

ASTES

Advances in Science, Technology & Engineering Systems Journal

Special Issue

Innovation in Computing, Engineering Science & Technology

2022

www.astesj.com
ISSN: 2415-6698

EDITORIAL BOARD
(Special Issue)

Editor-in-Chief

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

Guest Editors

Prof. Wang Xiu Ying
Chongqing University, China

Prof. Yu Xiao Yan
Chongqing Normal University, China

Prof. María Jesús Espinosa Trujillo
Universidad Tecnológica Metropolitana, Mexico

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

Prof. Mohamed Abdelaziz Hassan Eleiwa
University of Hail, KSA

Prof. Nicolae Tudoroiu
John Abbott College, Canada

Editorial

The Special Issue on Innovation in Computing, Engineering Science & Technology (2022) in the *Advances in Science, Technology and Engineering Systems Journal (ASTES Journal)* reflects the accelerating pace of technological convergence and the growing importance of integrative innovation across computational and engineering domains. In an era defined by rapid digital transformation and increasing system complexity, this issue presents a curated collection of research that highlights how novel computational techniques and engineering advancements are jointly shaping the future of science and technology.

A central theme of this issue is the expanding role of intelligent and data-centric methodologies in engineering applications. Several contributions explore the deployment of artificial intelligence, machine learning, and advanced data analytics to enhance system performance, enable predictive capabilities, and support real-time decision-making. These studies demonstrate how computing innovations are being seamlessly embedded into engineering processes, from smart manufacturing and automation to communication systems and intelligent control frameworks, thereby redefining traditional operational paradigms.

The issue also emphasizes innovation in core engineering sciences, particularly in the development of sustainable technologies, advanced materials, and high-efficiency systems. Authors investigate energy-aware designs, environmentally responsible engineering practices, and resilient infrastructures that address contemporary global challenges. The integration of computing tools with engineering innovation is especially evident in emerging areas such as the Internet of Things (IoT), cyber-physical systems, and next-generation network architectures, where interdisciplinary collaboration is essential for achieving scalable and adaptive solutions.

Methodological diversity is another strength of this special issue. The included papers employ a range of approaches, including theoretical modeling, computational simulations, experimental validation, and applied case studies. This blend of rigor and practicality ensures that the research contributes not only to academic advancement but also to industrial and societal applications. Many studies focus on scalability, robustness, and efficiency, offering solutions that are adaptable to a variety of technological environments and operational conditions.

Set against the backdrop of ongoing global transformation in 2022, the research presented in this issue reflects a heightened emphasis on digital resilience, system intelligence, and sustainable growth. The continued evolution of computing technologies, alongside advancements in engineering science, has reinforced the need for integrated approaches that can respond effectively to emerging challenges and opportunities.

The editorial team extends its sincere appreciation to the authors for their valuable contributions and to the reviewers for their diligent and constructive feedback. Their collective efforts have ensured the quality, relevance, and impact of this special issue.

This special issue provides a comprehensive overview of the innovations driving progress in computing and engineering science. By bringing together diverse perspectives and forward-looking research, it serves as a platform for advancing interdisciplinary knowledge and fostering continued innovation in technology and engineering systems.

Guest Editor

Prof. Wang Xiu Ying

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Special Issue **June 2022**

CONTENTS

Development of an Intelligent Road Anomaly Detection System for Autonomous Vehicles
by Paul Miracle Udah, Ayomide Ibrahim Suleiman, Jibril Abdullahi Bala, Ahmad Abubakar Sadiq, Taliha Abiodun Folorunso, Julia Eichie, Adeyinka Peace Adedigba and Abiodun Musa Aibinu

Lung Cancer Tumor Detection Method Using Improved CT Images on a One-stage Detector
by Young-Jin Park and Hui-Sup Cho

Meta-heuristic and Heuristic Algorithms for Forecasting Workload Placement and Energy Consumption in Cloud Data Centers
by Amine Bouaouda, Karim Afdel and Rachida Abounacer

The Perceptions of Students and Teachers When using ICTs for Educational Practices Matter: A Systematic Review
by Angela Pearce

Maintainability Improving Effects such as Insulation Deterioration Diagnosis in Solitary Wave Track Circuit
by Takayuki Terada, Hiroshi Mochizuki and Hideo Nakamura

Matching TCP Packets to Detect Stepping-Stone Intrusion using Packet Crossover
by Lixin Wang, Jianhua Yang, Austin Lee and Peng-Jun Wan

Multiple Criteria Decision-making: Risk Analyses for the Soft Target
by Dora Kotkova, Lukas Kralik, Lukas Kotek and Jan Valouch

Optimization of Query Processing on Multi-tiered Persistent Storage
by Nan Noon Noon, Janusz R. Getta and Tianbing Xia

An Improved Model to Analyze the Impact of Cyber-Attacks on Power Systems
by Muhammad Musleh Uddin, Kazi Rafiqul Islam and Md. Monirul Kabir

Assessment of Electromagnetic-Based Sensing Modalities for Red Palm Weevil Detection in Palm Trees
by Mohammed M. Bait-Suwailam, Nassr Al-Nassri and Fahd Al-Khanbashi

Regular Tessellation-Based Collective Movement for a Robot Swarm with Varying Densities, Scales, and Shapes
by Kohei Yamagishi and Tsuyoshi Suzuki

Assessing the Impact of Integrating Solar PV System using the Equal Area Criterion Method
by Abdul Ahad Jhumka, Robert Tat Fung Ah King and Chandana Ramasawmy

Advantages of 3D Technology in Stereometry Training
by Penio Lebamovski

Computer Vision Radar for Autonomous Driving using Histogram Method
by Hassan Facoiti, Ahmed Boumezzough and Said Safi

ARAIG and Minecraft: A Modified Simulation Tool
by Cassandra Frances Laffan, Robert Viktor Kozin, James Elliott Coleshill, Alexander Ferworn, Michael Stanfield and Brodie Stanfield

Profiling Attack on WiFi-based IoT Devices using an Eavesdropping of an Encrypted Data Frames
by Ibrahim Alwhbi Alharbi, Ali Jaber Almalki, Mnassar Alyami, Cliff Zou and Yan Solihin

Generalized Linear Model for Predicting the Credit Card Default Payment Risk
by Lu Xiong and Spendylove Duncan-Williams

A Proposal of Code Modification Problem for Self-study of Web Client Programming Using JavaScript
by Khaing Hsu Wai, Nobuo Funabiki, Khin Thet Mon, May Zin Htun, San Hay Mar Shwe, Htoo Htoo Sandi Kyaw and Wen-Chung Kao

Field Oriented Control and Commutation Based on Sensorless Methods for High-Speed Electrical Motors of Unmanned Multicopters
by Chen Zhao, Weisheng Kong, Federico Percacci and Patrik Gnos

μPMU Hardware and Software Design Consideration and Implementation for Distribution Grid Applications
by Ahmed Abdelaziz Elsayed, Mohamed Ahmed Abdellah, Mansour Ahmed Mohamed and Mohamed Abd Elazim Nayel

An Efficient Way of Hybridizing Edge Detectors Depending on Embedding Demand
by Habiba Sultana and A. H. M. Kamal

A Machine Learning Model Selection Considering Tradeoffs between Accuracy and Interpretability
by Zhumakhan Nazir, Temirlan Zarymkanov and Jurn-Guy Park

Frequency Oscillation Suppression of Interlinked Solar PV-Thermal Power System using HVDC Link
by Gulshan Sharma

The Security of Information Systems and Image Processing Supported by the Quantum Computer: A review
by Tarek Nouioua and Ahmed Hafid Belbachir
DEM models Calibration and Application to Simulate the Phosphate Ore Clogging
by Bouassale Nasr-Eddine, Sallaou Mohamed, Aittaleb Abdelmajid and Benaissa Elfahim

On the Prediction of One-Year Ahead Energy Demand in Turkey using Metaheuristic Algorithms

by Basharat Jamil, Lucía Serrano-Luján and José Manuel Colmenar

A Cloud Telemedicine Platform Based on Workflow Management System: A Review of an Italian Case Study
by Gianvito Mitrano, Antonio Caforio, Tobia Calogiuri, Chiara Colucci, Luca Mainetti, Roberto Paiano and Claudio Pascarelli

A Self-Adaptive Routing Algorithm for Real-Time Video Transmission in VANETs
by Marzouk Hassan, Abdelmajid Badri and Aicha Sahel

Metamaterial-Inspired Compact Single and Multiband Filters
by Ampavathina Sowjanya and Damera Vakula

Performance Adjustment Factor for Fixed Solar PV Module
by Kelebaone Tsamaase, Japhet Sakala, Kagiso Motshidisi, Edward Rakgati, Ishmael Zibani and Edwin Matlotse

Conception and Simulation of an Electronic Nose Prototype for Olfactory Acquisition
by Mostapha Harmouzi, Aziz Amari and Lhoussaine Masmoudi

A Comparison of Cyber Security Reports for 2020 of Central European Countries
by Kamil Halouzka, Ladislav Burita, Aneta Coufalikova, Pavel Kozak and Petr Františ

Secured Multi-Layer Blockchain Framework for IoT Aggregate Verification
by Ming Fong Sie, Jingze Wu, Seth Austin Harding, Chien-Lung Lin, San-Tai Wang and Shih-wei Liao

Automatic Counting Passenger System using Online Visual Appearance Multi-Object Tracking
by Javier Calle, Itziar Sagastiberri, Mikel Aramburu, Santiago Cerezo and Jorge García

Estimating a Minimum Embedding Dimension by False Nearest Neighbors Method without an Arbitrary Threshold
by Kohki Nakane, Akihiro Sugiura and Hiroki Takada

Regularity of Radon Transform on a Convex Shape
by Pat Vatiwutipong

Antenna System Design To Increase Power Transfer Efficiency with NFC Wireless Charging Technology
by Jérémy Quignon, Anthony Tornambe, Thibaut Deleruyelle and Philippe Pannier

Northern Leaf Blight and Gray Leaf Spot Detection using Optimized YOLOv3
by Brian Song and Jeongkyu Lee

Extended Buffer-referred Prefetching to Leverage Prefetch Coverage
by Jinhyun So and Mi Lu

Indoor Position and Movement Direction Estimation System Using DNN on BLE Beacon RSSI Fingerprints

by Kaito Echizenya and Kazuhiro Kondo

Long-term Bottom-up Modeling of Renewable Energy Development in Morocco
by Jabrane Slimani, Abdeslam Kadrani, Imad EL Harraki and El hadj Ezzahid

Automated Robotic System for Sample Preparation and Measurement of Heavy Metals in Indoor Dust Using Inductively Coupled Plasma Mass Spectrometry (ICP-MS)
by Heidi Fleischer, Sascha Statkevych, Janne Widmer, Regina Stoll, Thomas Roddelkopf and Kerstin Thurow

Scalability of Multi-Stage Nested Mach-Zehnder Interferometer Optical Switch with Phase Generating Couplers
by Masayuki Kawasako, Toshio Watanabe, Tsutomu Nagayama and Seiji Fukushima

Model Order Reduction and Distribution for Efficient State Estimation in Sensor and Actuator Networks
by Ferdinand Friedrich and Christoph Ament

Process Mining in Healthcare: A Systematic Literature Review and A Case Study
by Fabrizio Striani, Chiara Colucci, Angelo Corallo, Roberto Paiano and Claudio Pascarelli

Detection of Event-Related Potential Artifacts of Oddball Paradigm by Unsupervised Machine Learning Algorithm
by Rafia Akhter and Fred Beyette

Natural Tsunami Wave Amplitude Reduction by Straits – Seto Inland Sea
by Mikhail Lavrentiev, Andrey Marchuk, Konstantin Oblaukhov and Mikhail Shadrin

Blockchain Applications in Suning and PingAn
by Cong Qi, Yue Lei and Yuejun Cai

Designing the MIMO SDR-based Antenna Array for 5G Telecommunication
by Meriem Drissi, Nabil Benjelloun, Philippe Descamps and Ali Gharsallah

A Constrained Intelligent Nonlinear Control Method for Redundant Robotic Manipulators
by Dinh Manh Hung and Dang Xuan Ba

Association Rules for Knowledge Discovery From E-News Articles: A Review of Apriori and FP-Growth Algorithms
by Thilini Lakshika and Amitha Caldera

Estimating Subjective Appetite based on Cerebral Blood Flow
by Lai Kecheng, He Qikun, Hu Ning and Fujinami Tsutomu

Ensemble Extreme Learning Algorithms for Alzheimer's Disease Detection
by Vanamala H R, Samriddha Shukla and Vijaya krishna A

Emerging Trends in Green Best Practices and the Impact on Government Policy
by Emily Holt and Casey Corrado

Consideration of Ambiguity in the Analysis Phase of Data Warehouses
by Djamila Hammouche and Karim Atif

Interference-Aware Nodes Deployment of a LoRa-Based Architecture for Smart Agriculture in the Southern Region of Senegal
by El Hadji Malick Ndoye, Ousmane Diallo, Nadir Hakem and Emmanuel Nicolas Cabral

Optimizing Sensors Locations for Tsunami Warning System
by Mikhail Lavrentiev, Dmitry Kuzakov and Andrey Marchuk

Transfer and Ensemble Learning in Real-time Accurate Age and Age-group Estimation
by Anh-Thu Mai, Duc-Huy Nguyen and Thanh-Tin Dang

Mobility Intelligence: Machine Learning Methods for Received Signal Strength Indicator-based Passive Outdoor Localization
by Fanchen Bao, Stepan Mazokha and Jason O. Hallstrom

Technical Aspects and Social Science Expertise to Support Safe and Secure Handling of Autonomous Railway Systems
by Clemens Gnauer, Andrea Prochazka, Elke Szalai, Sebastian Chlup and Anton Fraunschiel

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 2, 1-13 (2023)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Development of an Intelligent Road Anomaly Detection System for Autonomous Vehicles

Paul Miracle Udah[*,1], Ayomide Ibrahim Suleiman[1], Jibril Abdullahi Bala[1], Ahmad Abubakar Sadiq[2], Taliha Abiodun Folorunso[1], Julia Eichie[3], Adeyinka Peace Adedigba[1], Abiodun Musa Aibinu[1]

[1]*Department of Mechatronics Engineering, Federal University of Technology, Minna,920101, Nigeria*

[2]*Department of Electrical and Electronics Engineering, Federal University of Technology, Minna,920101, Nigeria*

[3]*Department of Physics, Federal University of Technology, Minna,920101, Nigeria*

ARTICLE INFO

Article history:
Received: 24 August, 2022
Accepted: 08 January, 2023
Online: 11 March, 2023

Keywords:
Image Processing
Autonomous Vehicle
Road Anomaly
Fuzzy Logic
Speed Control
Maneuvering

ABSTRACT

Globally, road transportation has become one of the most reliable means of moving goods and services from one place to the other. It has contributed immensely to the standard of living and modern civilization. However, this means of transportation is characterised by some issues which are poised to be harmful to the human population if not properly addressed. One of such issues is the presence of potholes, bumps, and other road anomalies. Unfortunately, the late identification of road anomalies (Speedbumps and Potholes) and the inability of drivers to detect and slow down while approaching such road anomalies has also been a big challenge faced by many nations. Therefore, there is a need for an automatic and intelligent approaches to be built into vehicles to mitigate the number of road accidents caused by these anomalies. In this work, the development of an intelligent road anomaly identification and manoeuvring system for autonomous vehicle is presented. The developed system focuses on the detection of road anomalies specifically speedbumps and potholes; and the regulation of the vehicular speed when these anomalies are detected. A modified Histogram Oriented Gradient (HOG) and Fuzzy Logic Control (FLC) have been proposed in this work. Furthermore, promising results have been obtained and presented which depicts the proposed HOG algorithm outweigh other techniques in the identification and detection of speedbumps and potholes. In addition, the developed FLC was able to regulate the speed of the vehicle in the presence of speedbumps as well as navigate the vehicle accordingly in the presence of potholes.

1. Introduction

Road travel has facilitated the flow of goods and services in both developed and emerging countries, increased the quality of living, and has also led to global economic development. This goods and service distribution medium has now become one of the most important element of human society and may be one of the most dangerous elements for the human race if adequate steps are not enforced [1].

As since the advent of automobiles and other road vessels, traffic collisions have become the cumulative consequence of the adverse externalities of public transport [2]. Road accidents occur worldwide but are more frequent in dev eloping countries, triggered by over-speeding, traffic indifference, unfavorable weather, lack of vehicle maintenance and bad road conditions [2]. Some of such events are caused by human negligence or an unexpected natural phenomenon, which may be stopped with the correct precautionary measure. According to the World Health Organization (WHO), bad road conditions, especially in Nigeria, are the most unpredictable and the most prevalent cause of road accidents, and for every 100,000 vehicles accident, there are 1042 deaths worldwide in a year [3]. Another report found that in 2009 the annual road traffic fatality figure in Nigeria was 5,000. Poor roads condition and anomalies were the key contributors to these numbers, forcing drivers to swerve or loss control, which place them, and other road users at risk [3].

A study between the years 2000 and 2004 shows that about 36,000 Nigerians died and 125,000 were injured as a result of road crashes, increasing the death rate from traffic injuries in Nigeria as compared to other countries worldwide [4]. These accidents are

[*]Corresponding Author: Paul Miracle Udah, Federal University of TechnologyMinna, (+234)09097883956 , paulmiracle2014@gmail.com

https://dx.doi.org/10.25046/aj080201

usually caused by several factors such as inexperience, congestion, intoxication, intersection bottlenecks, distraction, drowsiness, and over speeding in the presence of road anomaly (Speedbumps and Potholes) [5] [6]. Speed bumps are structures that are built on the highways to limit traffic travel speeds in sensitive road areas, such as hospitals and schools [7], while a pothole is a bowl shaped deformity on a road surface that ranges from 150mm and above [8]. There are different types and sizes of potholes and they have harmful effects on pedestrians and cars regardless of their type or style. These road conditions account for damages to vehicles parts and causes drivers to lose control of their cars as they drive past them, which can result in a fetal road accident. Hence, the use of an advanced driver assistance systems (as found in autonomous vehicles) for the precise and timely identification of these anomalies would be idea to reduce and mitigate road accidents in developed and developing countries. Furthermore, this inclusion would also help provide driver with adequate information to aid navigation and speed control. [9].

Autonomous vehicles can be described as vehicles that that autonomously drives to the desired location without requiring any manual control [10]. Autonomous vehicles also have the capability to sense their surroundings and navigate around without the help of an external factor when subjected to different terrains [11]. The concept of autonomous vehicle (AV) involves computer vision (Perception), map navigation (Localization and Planning) and control integration. Each of these concepts must be present to ensure a vehicle is completely autonomous [12]. Regardless of the wide range of usage and nature of AV, human performance has proven to be reliable in driving ground vehicles, for instance, when overtaking another vehicle [13]. The implementation of computer vision and image recognition is very critical in facilitating proper object detection and self-navigation.

The application of computer vision is critically important in the development of AV as it is required for the acquistion of images of road and environment for ease of navigation and protection and safety of users [1]. Generally, computer vision provides machines with the functionality and ability to detect,analyse, interpret and comprehend videos and image information using techniques such as gradient, neutral network and edge detection [14]. Mostly, the technique involves viewing images as a two-dimensional signal, and then applying the traditional technique of image processing. When combined with an embedded device, highly efficient, low memory and robust solutions can be obtained [1]. In recent times, self-driving has been a topic of wide interest for Artificial Intelligence and the Automotive Industries. The irregularity detection on the road surface is a task with great challenges. In developing countries, it is very common to find un-marked speedbumps on road surfaces, which reduces the security and stability of self-driving cars [15]. In AVs, speedbump detection and other road anomalies can be achieved by using a number of techniques one of such is the use of mulitple cameras and microcontroller systems [16].

Some of the major limitations observed in the design and development of autonomous vehicle are blurriness and ineffectiveness in the detection of anomalies, as well as an absent or unsuitable control mechanism for vehicles to move over or around such anomalies. Therefore, the development of an intelligent road anomaly detection system, speed control and maneuvering system for autonomous vehicles is presented in this paper. The system analyses images of the surface of the road using image processing techniques (HOG and BRISK Features) for the detection of road anomalies (Speedbumps and potholes). The use of Fuzzy Logic Controller is used to regulate the speed and also the direction of the vehicle in the presence of r road anomalies (speedbumps and potholes).

In 2010, the number of vehicles used worldwide was estimated to be over 1 billion, and according to [17], AVs are widely anticipated to alleviate road congestion through higher throughput, and also reduce human errors to improve road safety and ease the stress of driving. Also, in [18], a review of previous work done on the effect of Artificial Intelligence on AVs was carried out. The use of intelligent transportation systems and artificial intelligence was employed in the improvement of the working of an autonomous vehicle, which was aimed at decreasing the possible error which can be encountered in AV development. The use of an artificial intelligence approach was used to show the possible setbacks in autonomous vehicles and an effective method using artificial intelligence in solving most of the problems in the autonomous vehicle was also proposed. This approach was proven to be possible and achievable with a low failure rate than the conventional approach used in other works reviewed. Although the algorithm employed in the development and improvement of an AV was effective, it was also time-consuming, difficult and ambiguous in development and implementation.

The remainder of this paper is organized as follows. Section 2 presents a review of related past works with respect to the identification of speedbumps and road anomalies specifically potholes. Also, presented in this section is the summary of the findings and gaps in the reviewed literatures. In section 3, the detailed methodology adopted for the research is presented with emphasis on the mathematical modelling of Avs, the Identification and detection of road anomalies using computer vision and AI. The technique for the speed control and navigation of the AV using fuzzy logic control system is also presented in this section. Section 4, presents and discusses the results obtained from the implementation of the proposed methodology as well as the performance evaluation of the methodology vis-à-vis standard metrics of evaluation. Section 4 presents the summary of findings, conclusion of the work and further research opportunities.

2. Literature Review

2.1. Related works in Pothole Detection

There exists in literature several works in the area of pothole identification and avoidance. In [19] the development of an image-based pothole detection system for intelligent transport system (ITS) service and road management system was proposed. In this work, , a pothole detection method based on two-dimensional (2D) images was proposed for improving the existing method and to be applied to ITS service and road management systems. The use of Modified Histogram Shaped-Based Thresholding (MHST) improved the accuracy of the system to 73% and the precision to 80%. However, the system exhibited a long image processing time

(46s average), which limits the practical application of the approach. Furthermore, the vibration of the vehicle also has effect on affects the detection unit of the proposed system

In addition, [20] developed a real-time pothole detection and road monitoring system. The work introduces a system to detect the potholes and informs the concerned authority about the pothole. The proposed system used an ultrasonic sensors and accelerometers to measure the depth of the pothole and jerking respectively. The system captures the geographical location of potholes using a Geograhic Position System (GPS) module. The acquired information is also stored in a cloud database. This serves as a valuable source of information to the vehicle drivers and Government authorities, through a web server with public access. The proposed system provides precautionary measures that can be taken to evade accidents. However, due to the placement of the Ultrasonic sensors, the system is unable to detect a pothole before the vehicle goes over it.

Furthermore, [21] implemented an android based pothole detection system using ultrasonic sensors. The system makes use of an integrated mobile app with the pothole detection system (Ultrasonic Sensor and Microcontroller) for monitoring the road path and anomalies. The system helps users monitor road paths and associated anomalies remotely with the aid of the mobile app. However, the system is limited in its capability to detected all types of road anomalies due to shortcomings of the ultrasonic sensors used. Another limitation of the proposed system is the quality of service of the internet/ communication network.

In a related development, [21] designed a real-time pothole detection technique using image processing and dimension estimation. The proposed system uses image processing and edge detection (Canny, Zero Cross) techniques identify and detect potholes in realtime. The proposed methodology ensures the estimation of the depth and width of potholes for the purpose of remediation such as filling and patching. The system applies a high level of image processing, filtering of images and application of scaling factors to accurately determine and calculate the area of the potholes.

Similarly, [22] designed an automatic pothole detection system using laser technology. The system captures the geographical location coordinates of the potholes and humps.. The developed system effectively detects humps and potholes without any huma intervention before the vehicle actually hits the potholes or the humps. With the use of laser technology, the system can calculate the distance between vehicles and potholes faster than an Ultrasonic Sensor. A major limitation of this system, however, is that it does not possess the ability to control the vehicles when maneuvering potholes.

In , [23] an intelligent system for pothole and hump identification on roads using Internet of Things (IoT) was proposed. The proposition makes use of a database to keep records of identified potholes and also sends the data to the appropriate authorities. The system is also capable of alerting the driver of a pothole ahead so that the vehicle maneuver or slowed down accordingly. However, the system does not assist the driver maneuvering potholes but only detects the potholes.

2.2. Related works in Speedbump Detection

There also exist a number of works in the literature as it relates to speedbumps . In 2010, the number of vehicles used worldwide was estimated to be over 1 billion, and according to [24], AVs are widely anticipated to alleviate road congestion through higher throughput, and also reduce human errors to improve road safety and ease the stress of driving.

In [25], a survey on possible production errors and security challenges faced by Autonomous Vehicles was carried out. The use of road segmented remedy, intersection-based solution, active & passive countermeasure was employed in carrying out the survey on possible production error and security challenges faced in autonomous vehicles.

In a related development a road surface condition monitoring and database system with the aid of a programmed accelerometer sensor was proposed by [26]. The system identifies different road conditions with the aid of vehicular vibrations due to their acceleration on road anomalies.. The proposed approach was reliable, durable and not time-consuming which makes it suitable for the development of road monitoring systems. The proposed approach offers a reliable and fast medium for the detection of road surface profiles. In the system acquired profiles are stored to form a roboust database for future references. Regardless of the benefit the system offers, the mechanism for the protection of the stored data on the database was not provided which pose a security risk to user data and this limits the application in AV systems as reagrds vehicle to vehicle communication [26].

Also, a smartphone road anomaly detection system was proposed and developed in [27]. The proposed system adopted a modified fuzzy logic guassian model for analysing and classification of road anomalies. The developed system employed a K-nearest neighbour (KNN) algorithm to classify the sensed anomalies into Potholes and Bumps. The proposed system demonstrated versatility, effectiveness with a 94% to 96% accuracy rate in the recognition of road anomaly and compatiblilty with different types of vehicles and its low cost-effectiveness. Regardless of the effectiveness, versatility and accuracy of the technique used in the recognition of road conditions, the technique is only applicable to only two types of road conditions namely potholes and bumps[27].

[13], proposed the use of a fuzzy logic controller for the development of a vision and control system for autonomous vehicles. The fuzzy logic controller comprises a steering controller and a speed controller. One of the added advantages of this method was its ability to add human knowledge and experience via the fuzzy decision tree language into the system to achieve higher accuracy and precision. Although the method is capable of performing automatic driving on straight roads, the scheme is challenging to implement in AV development [13].

The image processing was used in the development of a speed breaker and road marking detection system in [28]. The technique involved using image processing for speed breaker and road marking detection. The authors also integrated Optical character recognition and Hough transform alogrithms to recognize traffic signs and detect line markings respectively. An

accuracy rate of 79.97% for optical character recognition and 100% for the local binary patterns was obtained after analysis. One of the limitation of the system was the inability to detect speed breaker and road marking in the absence of light which made the system not fully functional at night.

2.3. Summary of findings and research gaps

In the light of the above, it is evident the application of image processing and computer vision have aided the identification and detection of road anomalies on motorways. Albeit, the blurriness of the road anomalies images as identified from the perspective of the moving vehicle and the inability to regulate the speed of vehicle when approaching anomalies are another issues. These and other issues identified in literature have resulted in the ineffectiveness of developed solution. Thus, to proffer solutions to these challenges, this work proposes the development of a improved Histogram of Oriented Gradient (HOG) image processing algorithm to detect road anomalies as well as estimate their locations. Furthermore, to control the speed and effectively navigate AVs, this work proposes the adoption of a modified fuzzy logic controller to regulate the speed of a vehicle when approaching a road anomaly. Table 1 shows some of the related work carried out in autonomous vehicle and pothole detection.

Table 1: Related works on Pothole and Autonomous vehicle

S/N	Author(s)	Work	Year	Description	Strength(s)	Weakness(es)
1	Seung-ki, R., Taehyeong, K. & Young-RO, K.	Image-Based Pothole Detection System for ITS Service and Road Management System.	2015	ITS (Intelligent Transportation System) service and road management system. A pothole detection method based on two dimensional (2D) images is proposed for improving the existing method and designing a pothole detection system to be applied to ITS service and road management system.	The use of a Modified Histogram Shaped-Based Thresholding (MHST) improved the Accuracy to 73% and Precision of 80%.	Long image processing time (46s average), accuracy can be improved with the use of more than one optical sensor, the vibration of vehicle affects the detection unit.
2	Ashwini, K., Kore, K., Akshada, K., & Ashwini, M.	Pothole Detection System Based on Android Device and Ultrasonic Sensor.	2018	The system makes use of an integrated mobile app with the pothole detection system (Ultrasonic Sensor and Microcontroller) for monitoring of the road path.	the mobile app helps the users monitor the road path remotely.	The use of just an ultrasonic sensor causes the system to be inefficient in detecting all forms of potholes. The system would be ineffective where there is poor or no network reception.
3	Rahul, H., Ashish, Z., & Nilesh, P.	Real time pothole detection technique using image processing and dimension estimation.	2019	The proposed system uses image processing and edge detection (Canny, Zero Cross) techniques to take real time data of potholes. It uses this data to estimate the depth and width of potholes for filling and patching purposes.	The system applies high level of image processing, filtering of images and application of scaling factors to accurately calculate the area of the pothole.	The processing time of the system will be relatively slow when applied to self driving cars for pothole detection due to the fact that self driving cars need only fast estimations of the pothole perimeters and not the depth.
4	Shivaleelavathi, B. G., Veeramma, Y., Yamini, V. S., & Spoorthi, T.	Design and Development of an Intelligent System for Pothole and Hump Identification on Roads.	2019	This project aims at identifying the potholes using image processing technology with the help of a camera. The system sends the location of these potholes to authorities using GPS. The humps are detected using the ultrasonic sensors and the electronics system alerts the driver about the humps and the driver can accordingly slow down the vehicle and hence avoid the accidents. The alert is provided in the form of voice through speakers and displayed on the LCD and using GSM, SMS is sent to the driver.	The system makes use of an IOT database to keep records of identified potholes and also sends the data to the appropriate authorities. The system is also capable of alerting the driver of a pothole ahead so that the vehicle can be manoeuvred or slowed down accordingly.	The system does not assist the driver manoeuvring potholes but only detect the potholes.
5	Youngtae, J., & Seungki, R.	Enhanced Pothole Detection Algorithm for Blurry Images.	2015	This system comprises of a pothole detection algorithm using various features in 2D images that improved the existing methods and accurately detected a pothole. The System implemented the use of	The major strength of this system is the ability of the system to be accurate enough to detect real potholes on the road with the aid of	There is a little the system can do if the image blur is caused by vibrations or shake in the camera. The restoration rate in

				image sharpening algorithms to prevent wrong detection of potholes.	a de-Blurring image sharpening algorithm.	situations like this is still very low.
6	Akshatha, P. , Chaitra, P. , Lakshmi, S. , & Richard, L.P.	Automatic Detection of Potholes	2019	Automatic detection of potholes is a human safety based project. This system provides cost effective solution for detection of potholes on the road by using ultrasonic sensors. The microcontroller fetches the location of the pothole using GPS and the GPS locations are sent with the help of GSM. The GSM at the vehicle part takes that location and sends the location to the concerned person and that person can see pothole on the map of that area.	This system provides cost effective solution for detection of potholes and humps on the road and indicate the road maintenance authority for maintenance. This system also provides the driver an indication that there is a pothole ahead and the desired action can be taken	The system does not possess the ability aid the vehicle in control. The alert system does not provide exact location of pothole on the road path to aid avoidance of the pothole.

3. Research Methodology

3.1. System Description

The design and development of the speedbump detection and speed control system for an autonomous vehicle involve numerous components. These components include Raspberry pi 3 microprocessor, speed sensor (LM 390), Pi camera, Model Car, Direct Current (DC) motors, servo motor, power supply.. The Raspberry pi microcontroller is the brain of the system. The inputs to the microcontroller are the images acquired from the camera as well as the speed of the vehicle at that instant. The outputs of the system are the speed of the DC motor and the angle of the servo motor. These outputsrl indicates the status of the vehicle in terms of speed and direction in the presence of road anomalies (Speedbumps and potholes). The raspberry pi microcontroller carries out all major processing and also houses the HOG and Fuzzy Logic Control algorithms.

When an anomaly is detected, the system will evaluate the distance of the vehicle from the anomaly, based on the data obtained from the camera, the fuzzy logic controller determines the appropriate speed and direction to be used when approaching an anomaly. When the vehicle is close to an anomaly, the speed will be reduced by the controller, and maneuver the vehicles in the appropriate direction. The vehicle will revert to its original speed and position after it has successfully passed the anomaly.

The block diagram representing system architecture is shown in Figure 1 and the flowchart of the system operation is shown in Figure 2.

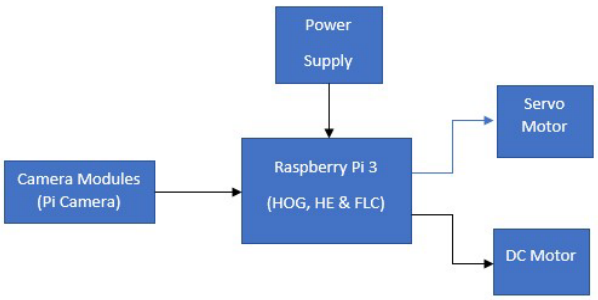


Figure 1: System Architecture

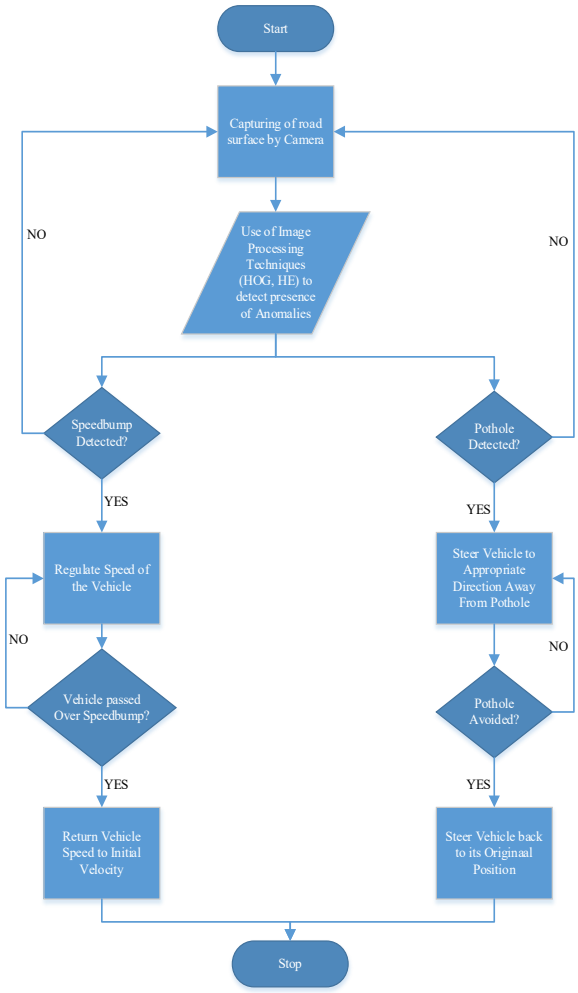


Figure 2: System Flowchart

Figure 2 shows the process of how information is being acquired, gathered, processed, and transmitted from one section of the system to the other. The image of the anomaly (speedbump or pothole) is captured by the camera by extracting the features of the road surface using the histogram of oriented gradient technique and Histogram Equalization respectively. When the road anomaly

has been detected, a command is sent to the microcontroller which regulates the speed of the vehicle, when the anomaly is a speedbump, and in the presence of a pothole, a command/ signal is sent to the system to change the direction of the vehicle by changing the angle of the servo motor. When the vehicle has successfully passed or steered passed the anomaly, the speed and the direction of the vehicle is returned to the original velocity and position respectively. This process continues the vehicle is in motion.

In this work, MATLAB software was used for the realization of the vehicle model, development and implementation of the image processing and fuzzy logic algorithms . MATLAB is an acronym for "Matrix Laboratory" which is a high-level programming language and an Integrated Development Environment (IDE) used for numeric computations, simulations and programming. MATLAB is used in different fields for data analysis, algorithm development and the development of system models. It has a vast variety of tools and commands essential to almost all engineering applications.

3.2. Mathematical Model

In this work, the autonomous vehicle was modelled in terms of the longitudinal dynamics of the vehicle. The external longitudinal forces acting on the vehicle include aerodynamic drag forces, gravitational forces, longitudinal tire forces and rolling resistance forces.

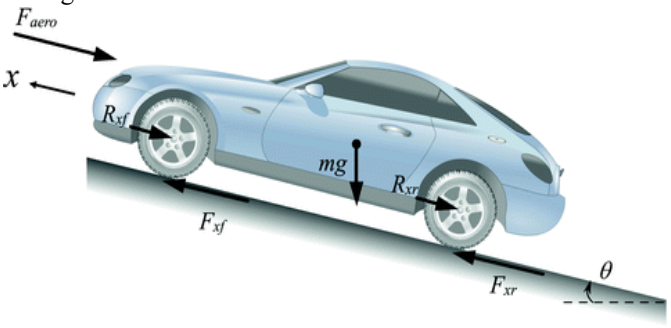


Figure 3: Longitudinal Vehicle Dynamics [29]

Figure 3 shows a diagram representing the longitudinal dynamics of the vehicle. The diagram shown in Figure 3 highlights the forces considered when modelling the longitudinal dynamics of a vehicle. F_{xr}, F_{aero}, F_{xf}, R_{xf}, mg, and R_{xr} represent the longitudinal tire forces at the rear, aerodynamic drag force, rolling resistive force at front tires, mass of the vehicle, and rolling resistive force at front tires, respectively [29]. The force balance along the vehicle longitudinal axis yield is showed in (1).

$$m\ddot{x} = F_{xf} + F_{xr} - F_{aero} + R_{xf} - R_{xr} - mgsin(\theta) \quad (1)$$

Where:

F_{xf} = longitudinal tire force at front tire
F_{xr} = longitudinal force at rear tire
F_{aero} =equivalent longitudinal aerodynamic drag force
R_{xf} = rolling resistance at front tire
R_{xr} =rolling resistance at rear tire
m = mass of the vehicle
g = acceleration due to gravity
θ = angle of inclination of the road the vehicle travels θ

The aerodynamic drag force on the vehicle can be represented as equation 5.

$$F_{aero} = \frac{1}{2}\rho C_d A_f (Vx + Vwind)^2 \quad (2)$$

Where:

ρ is the mass density of air.
C_d is the aerodynamic drag coefficient.
A_f is the frontal area of the vehicle which is the projected area of the vehicle in the direction of travel.
$V_x = \dot{x}$ is the longitudinal vehicle velocity, V_{wind} is the wind velocity (positive for a headwind and negative for a tailwind) [29].

3.3. Road Anomaly Detection

a) Image Acquisition

This process entails acquiring the image from the surface of the road, which will be used in determining the presence or absence of anomalies by applying image processing techniques. Image processing has become one of the main aspects of automation and safety-related applications in the electronic industries and these images can either be in a digital or analog form [15]. This process of acquiring the image will be carried out using a Pi camera with a resolution of 8 Megapixels, when the vehicle is moving with a speed of 5 to 10 m/s.

b) Image Preprocessing (Color space conversion, HE, HOG noise removal)

Images of road surfaces are captured using the Pi camera. The captured images are sent to the microcontroller to carry out image processing to detect and extract both speedbump and pothole features. The next process is done by the microcontroller to convert the image's colour space from RGB to grayscale, filter all distortion and blurriness using the median filter, crop out the Region of Interest (ROI) of the image, and HOG and HE respectively is applied on the image to enhance the image. The mathematical representation of the histogram equalization (HE) is in (3) and Histogram of Oriented Gradient (HOG) in (4) to (8).

Equation for HE:

$$h(v) = floor\left((CFD(v) - \frac{CFDmin}{(Mxn) - CFDmin}x\,(L-1)\right) \quad (3)$$

Where:

h = new grayscale value.

v = pixel number.

CDF(v)= Cumulative Density Function at v.

CDFmin = minimum Cumulative Density Function.

M x n = image rows & columns.

L = current grayscale level.

After histogram equalization, a feature extraction process is used to identify features on the image that are relevant to the scope of this project (potholes). This process is done by using the BRISK (Binary Robust Invariant Scalable Key points) points algorithm. The BRISK points represent the metric data on the region of interest with the most significant value. Pothole elements in images have more distinct metric values due to the disparity in pixel contrast compared to the road surface. This process is used to retrieve the location of the strongest metric data representing the pothole features recognized in the segmented image. The point's locations are then stored in an array.

Equation for HOG:

Using the 1st order differential coefficient, $f_x(x, y)$ and $f_y(x, y)$

$$fx(x,y) = f(x+1,y) - f(x-1,y) \quad (4)$$

$$fy(x,y) = f(x,y+1) - f(x,y-1) \quad (5)$$

where $f(x, y)$ in (1) and (2) is luminance at (x, y)

The magnitude and direction of the gradient (m & θ) can be calculated using (6) and (7)

$$(x, y) = \sqrt{fx(x, y)^2 + fy(x, y)^2} \quad \textit{(Magnitude)} \quad (6)$$

$$(x, y) = \arctan \frac{fx(x,y)}{fy(x,y)} \quad \textit{(Direction)} \quad (7)$$

After computation of (6) and (7), the HOG is generated by the following steps:

- determining the class which $\theta(x, y)$ belongs to.
- increasing the value of the class using the first step.
- repeating the above operation for all the gradients belonging to the cell.

Equation (8) can be used in finding the final HOG feature.

$$v = Vk \quad (8)$$

Where Vk = vector corresponding to a combined histogram for a block and V = normalize Vector.

c) Image Segmentation

Image segmentation entails the use of Sobel edge detection in identifying the edges present in an image, which is carried out based on the characteristics of the pixels in the image. The Sobel filter works by calculating the gradient of image intensity at each pixel within a given image. From the result gotten from the input of the Pi camera, the edges of the images were obtained by finding the approximate absolute gradient magnitude at each point in the image.

d) Feature Extraction (HOG and BRISK)

Feature extraction involves the identification and selection of the desired data from an image, and converting that data obtained into useful information The feature extracted provides the coordinate of the speedbump on the road surface when present with the use of the HOG Technique, while the feature and coordinate of the pothole of are provided using the Brisk points. These features are used in training the system in detecting the presence or absence of road anomalies (Speedbump and Potholes) respectively. During the feature extraction, the HOG points converge around or on the surface of the anomaly (i.e., Speedbump) and also during that of the BRISK points. Equation 9 was used in crrying out the HOG technique.

$$\boldsymbol{v} = \frac{\boldsymbol{Vk}}{\sqrt{||\boldsymbol{Vk}||2+1}} \quad (9)$$

e) Feature Recognition (ANN, DT, and SVM)

The use of Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree techniques were applied in training and testing of the designed system in recognizing the presence or absence of road anomaly (speedbump and potholes). The presence and absence of anomalies on the road surface were used as the target in training the system, where the presence of anomalies was represented by an output one (1), and the absence of anomalies was represented as an output zero (0).

3.4. Speed and Steering Control

a) Fuzzy Logic Design for Steering

Fuzzy Logic is a method from imprecise and ambiguous knowledge, which provides a definite solution. When using a series of statistical rules used to describe expertise dependent on membership degrees, this is achieved. Fuzzy Logic has widespread applications in control engineering, signal processing, and artificial intelligence.

The fundamental type of fuzzy logic control modelling, known as the fuzzy inference system (FIS), is a rule-based structure composed of three conceptual components:

- A rule base composed of a set of fuzzy if-then rules.
- A database describing the membership function (MF) used in these fuzzy rules.
- A reasoning process incorporating such rules into a mapping routine from inputs to the outputs of the system to drive a reasonable output.

There are essentially two types of base rule structures, namely Mamdani and Takagi-Sugeno-Kang (TSK). The fuzzy rule is expressed in language form in the Mamdani type. In this work, the fuzzy rule was represented using the Takagi-Sugeno-Kang rule structure as a logical function of the input variables which is more suitable for neuro-fuzzy systems. The fuzzy logic approach is made up of five steps:

- Define all input and output variables.
- Define the subset interval.
- Choose the membership function.
- Define the IF-THEN rules.
- Perform calculations and adjust the rules.

Due to its simplicity, ease of interpretation, acceptance of imprecise data and ability to integrate with other traditional control methods, this study will employ the use of a Fuzzy Logic Controller. Figure 4 presents a fuzzy logic control system.

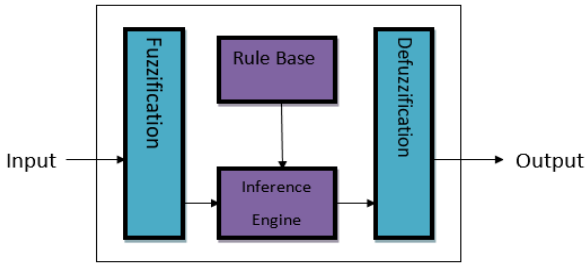


Figure 4: Fuzzy Logic System.

In this work , the steering of the vehicles in the presence of a pothole anomaly is determined by the rule base of the fuzzy system, the rule base is the representation of the IF-AND-THEN rules used in the implementation of the fuzzy controller. These rules are stated below, with the diagram representation shown in Figure 5:

i. If (POA is PRESENT) and (DTA is LEFT) then (STEERING is RIGHT)
ii. If (POA is PRESENT) and (DTA is RIGHT) then (STEERING is LEFT)
iii. If (POA is ABSENT) then (STEERING is MIDDLE).

where

POA : Presence of Abnormaly
DTA: Distance to Abnormaly
STEERING: Control of vehicle

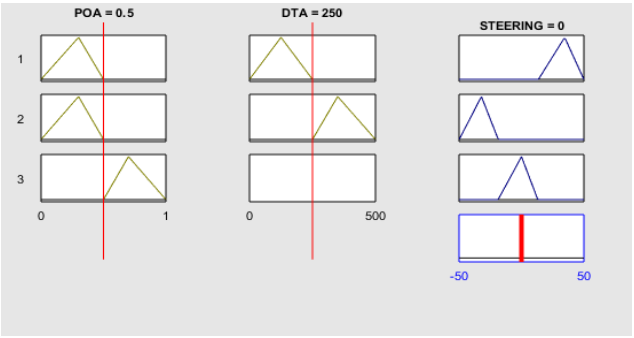


Figure 5: Rule Base for FIS.

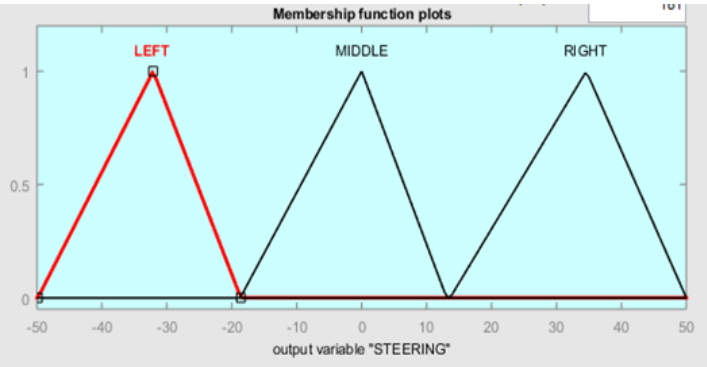


Figure 6: Output Membership Function for Steering.

The fuzzy logic controller uses the output membership function from the FIS in Figure 6 to determine the direction to control the autonomous vehicle in the presence or absence of road anomaly (pothole). The surface diagram describing the fuzzy logic controller's FIS is shown in Figure 7.

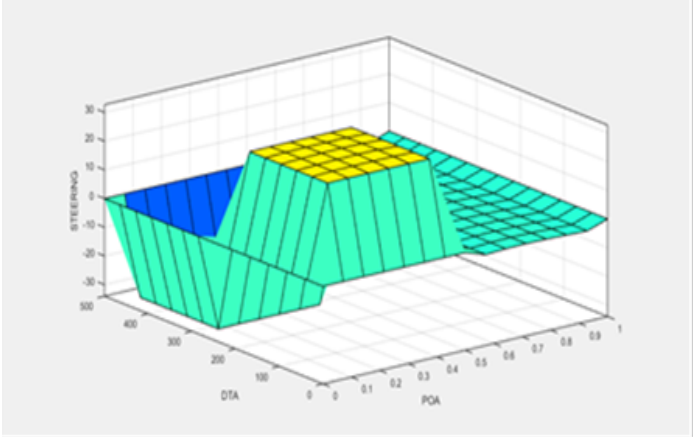


Figure 7: Surface Diagram of Fuzzy Logic Controller

b) Fuzzy Logic Design for Speed

The use of a surface diagram was employed to represent the relationship between the input and output parameters of the designed FLC system. Figure 8 shows the surface diagram of the design FLC system, by representing the relationship between the input membership parameter to the output membership parameters in a three-dimensional graphical representation, which was generated from the design of the FLC controller.

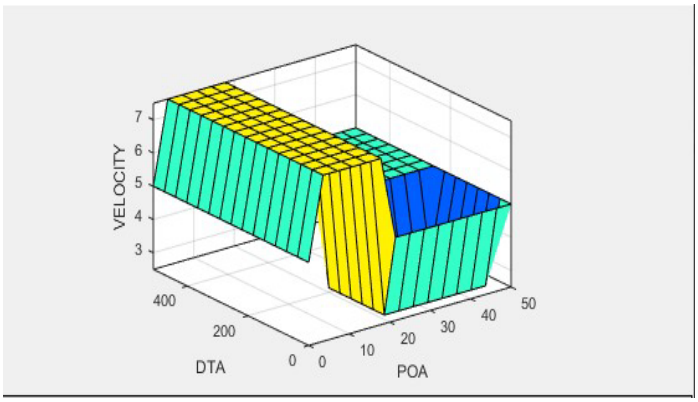


Figure 8: System Surface Diagram

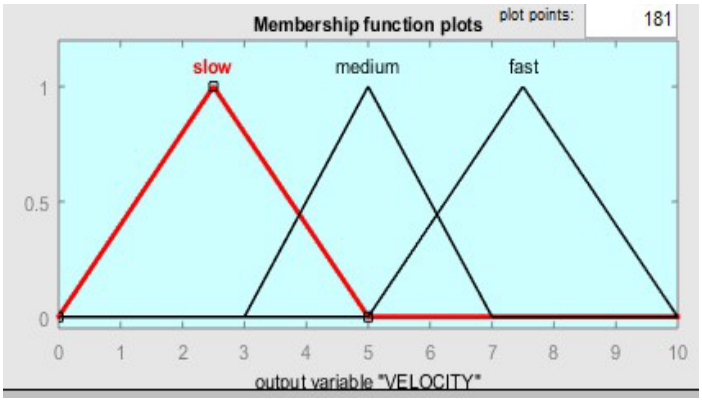


Figure 9: Output Membership Function for Speed Control.

The fuzzy logic controller uses the output membership function from the FIS in Figure 9 to determine the direction to control the autonomous vehicle in the presence or absence of road anomaly (speedbump).

In the design of the FLC system, rules were generated in order to provide the needed condition (rules) on which the FLC depend in controlling the speed of the vehicle in the presence or absence of speedbumps. Table 2 shows the rules generated by FLC in a graphical format. Where

POA : Presence of Anomaly

DOA: Distance of Anomaly

Table 2: FLC Rule Base

POA / DOA	Present	Absence
Close	Slow	Fast
Far	Med	Fast

3.5. Testing and Performance Evaluation

The performance evaluation of the system is carried out to determine the effectiveness, accuracy, precision and efficiency of the method and techniques used in the development of this project. The intelligent speedbump detection system is tested and evaluated based on precision, recall and accuracy in detecting the presence of speedbumps on the road. The metrics are represented mathematically as:

$$\text{Precision} = TP / (TP + FP) \quad (10)$$

$$\text{Accuracy} = TP / (TP + FN) \quad (11)$$

$$\text{Sensitivity} = FP / (TN + FP) \quad (12)$$

where:
TP = true positive (number of abnormally classified correctly).
FP = false positive (number of negative cases of abnormally classified correctly).
TN = true negative (number of correct cases of abnormally classified wrongly). FN = false negative (number of incorrect cases of abnormally classified wrongly).

The fuzzy logic model was also evaluated based on the response of the system to the rules generated by the fuzzy inference system in regulating the speed and direction of the vehicle when there is the presence of road anomaly. In carrying out the performance evaluation of the system, the accuracy rate which indicates how well the system was able to correctly detect the positive cases of a speedbump.

4. Results

This section presents the result gotten from the detection of road anomaly, i.e. speedbumps detection system in section 4.1 and that of the pothole detection system in section 4.2 respectively.

4.1. Road Anomaly Detection

a) Speedbump

The results obtained from the development of an intelligent speedbump and speed control system for autonomous vehicle, with the use of the HOG technique is shown in Figure 10. The result was obtained with the implementation of the HOG feature extraction technique in order to extract the needed data from the image acquired by the camera (i.e., speedbump). The feature extracted provides the coordinate of the speedbump on the road surface when present, and it's also used in training the system in detecting the presence or absence of road anomaly. During the feature extraction, the HOG points converge around or on the surface of the anomaly (i.e., Speedbump).

The use of Artificial Neural Networks (ANN) and Decision Tree techniques was applied in training and testing of the designed system in recognizing the presence or absence of road anomaly (speedbump). The results obtained from the training the system with the use of ANN and Decision Tree are presented in Figure 11 to Figure 14.

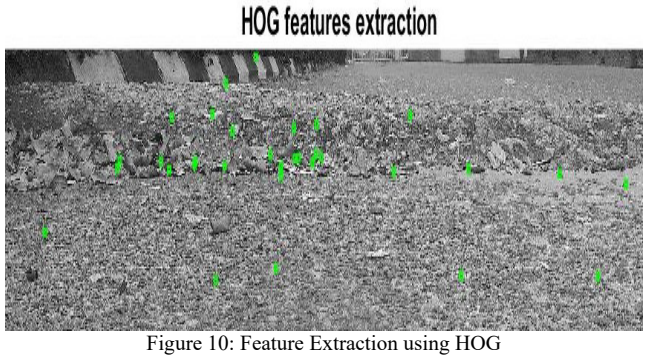


Figure 10: Feature Extraction using HOG

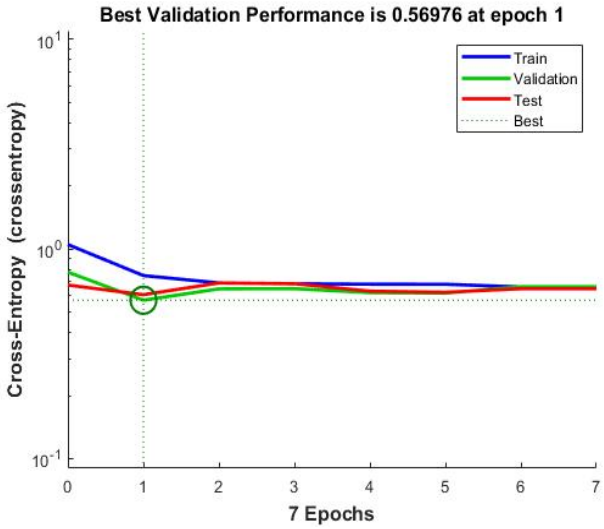


Figure 11: Performance Validation of System (ANN)

From Figure 11, it is shown that the best performance of the trained model occurred at a value of 0.56976 before the occurrence of overfitting at 7 Epochs. The confusion matrix of the system in Figure 12 presents the result in the detection of an anomaly. The confusion matrix shows that the system had a testing performance of 58.4%, a training performance of 71.4%, a validation performance of 71.4% and total performance of 62.3% for the detecting of speedbumps on road surfaces.



Figure 12: Confusion Matrix of Neural Network

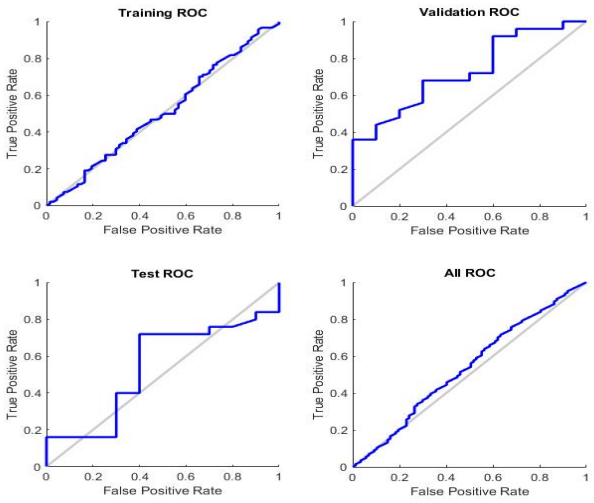


Figure 13: Receiver Operation Characteristics of Neural Network

The evaluation of the result between the output of the model to the true positive rate and false positive rate in the classification of the data. From the ROC shown in Figure 13, the system had a higher True Positive Ratio compared to the False Negative Ratio in the validation, training and testing ratio graph in the detection speedbump.

Figure 14 presents a graphical representation of the output gotten from the application of training the system with the use of a classification learner (Decision Tree). The Classifier was also provided with a dataset containing 231 road images (144 images with speedbump and 87 images without speedbump). The Decision tree classifier was also trained with 144 features extracted from the images and had an accuracy of 75.7%.

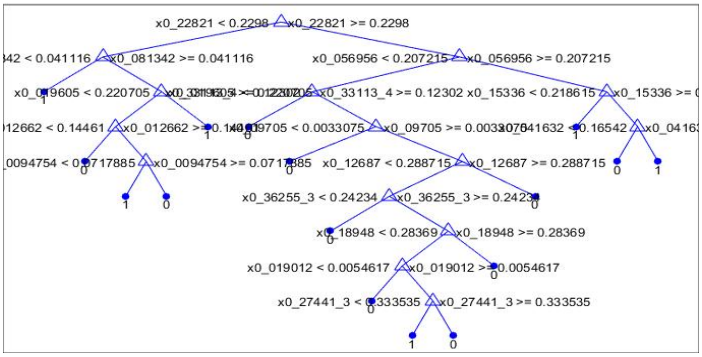


Figure 14: Classification Tree Model for Decision Tree for Speedbump

b) Pothole

The results obtained from the image processing algorithm, image acquisition, image preprocessing using Histogram Equalization, and feature extraction are shown in Figure 15.

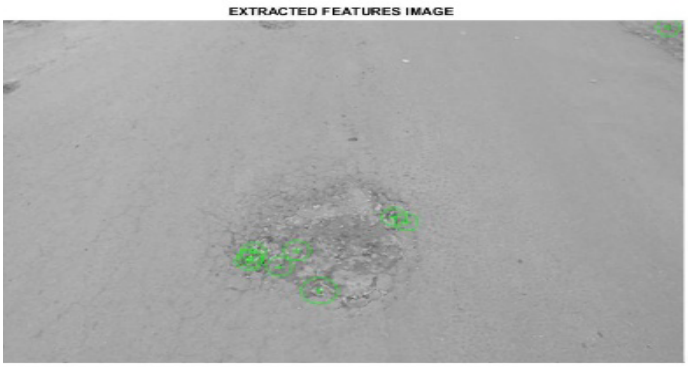


Figure 15: BRISK Points Extracted from the Image.

4.2. Vehicle Maneuvering

This section presents the result gotten from the design of the fuzzy logic system in the steering and speed control of the vehicle. The result of the steering fuzzy system is shown in section 4a and the speed control is in section 4b.

a) Steering

Table 3 presents the extracted points from the image. these points will be fed into the feature recognition (ANN) model to predict if a pothole is present in the image.

Table 3: Brisk Points Location

BRISK POINT	X AXIS	Y AXIS
1	404.0000	269.0000
2	287.3419	337.0000
3	258.8189	324.8189
4	305.3646	314.9544
5	416.9694	275.9694
6	328.2500	370.0000
7	256.0000	329.0000
8	261.0000	317.2109
9	256.4589	274.0000

Table 4: Fuzzy Logic Controller Performance results

Presence of anomaly (POA) (Meter)	Distance to Anomaly (DTA) (Meter)	Output (Velocity) (M/s)	Accuracy	Remark
0.1	0	1	Accurate	Slow Velocity
0.2	20	2	Inaccurate	Fast Velocity
0.3	40	3	Accurate	Slow Velocity
0.4	60	4	Accurate	Medium Velocity
0.5	80	5	Accurate	Medium Velocity
0.6	100	6	Inaccurate	Fast Velocity
0.7	120	7	Accurate	Slow Velocity
0.8	140	8	Accurate	Fast Velocity
0.9	160	9	Inaccurate	Fast Velocity
1	180	10	Accurate	Fast Velocity
2	200	10	Accurate	Fast Velocity
3	220	6	Inaccurate	Slow Velocity
4	240	4	Inaccurate	Slow Velocity
5	260	7	Accurate	Fast Velocity
6	280	10	Accurate	Fast Velocity
7	300	10	Accurate	Fast Velocity
8	320	8	Inaccurate	Slow Velocity
9	340	10	Accurate	Fast Velocity
10	360	9	Inaccurate	Medium Velocity
11	380	10	Accurate	Fast Velocity

b) Speed

The Fuzzy Logic Controller was used in the control of the vehicle speed in the presence of a speedbump, this was also carried out using a maximum velocity of 10M/s in controlling the speed of the vehicle. The presence the distance of the speedbump was used in the determination of the velocity at the vehicle moves at. Table 4 present the response of the FLC in regulating the speed of the vehicle in relation to the presence and distance of rod anomaly (speedbump) to the vehicle.



Figure 16: Developed MATLAB Application for Speedbump Analysis on Images.

4.3 Testing and Performance Analysis

a) Developed Application Interface

The feature recognition system was evaluated by developing a GUI application in testing the detection accuracy of the ANN model to that of the Decision Tree model and the SVM for both speedbump and pothole anomaly as shown in Figure 16 and Figure 17 respectively. Also, the results obtained from the ANN, SVM and Decision Tree model displayed on the MATLAB application used for the image analysis are shown in Figure 16.

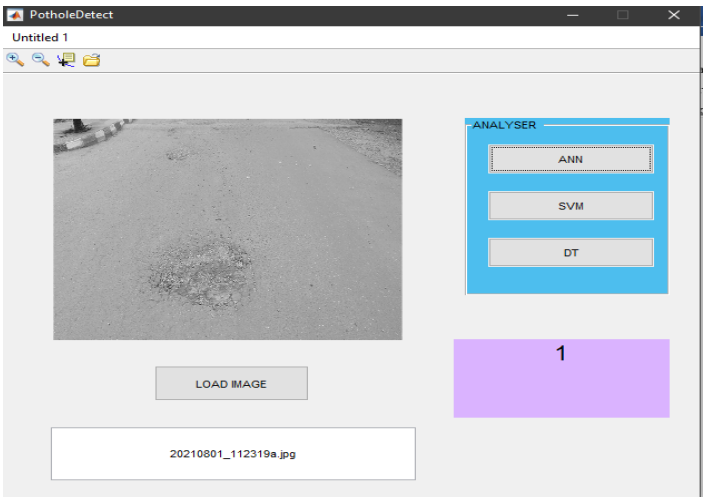


Figure 17: Developed MATLAB Application for Pothole Analysis on Images.

From the results presented in the application in Figure 17, an output of 1 was achieved signifying that an anomaly (pothole) is present in the image.

b) Comparative Analysis

The ANN, SVM and DT models' performance was evaluated based on accuracy, precision and sensitivity using Equations (9), (10), and (11). The evaluated performances are shown in Table 4 and Table 5 for both speedbump and pothole respectively.

After the two models have been designed, trained and tested, a comparison between both techniques was carried out, based on the accuracy, recall and precision of their results. Table 2 shows the comparative analysis between the ANN and Decision Tree models. The Precision, Accuracy and Recall of both trained models were calculated using (9), (10) and (11). The aim of the result shown in Table 5, is to determine the most optimal and suitable prediction technique for detecting speedbumps on the road surface. From Table 5, it was seen that the Decision Tree model produces a higher accuracy, precision and recall, with a percentage of 75%, 82% and 77% respectively.

Table 5: Comparison Table of ANN and Decision Tree

Parameters/ Technique	Artificial Neural Network (ANN)	Decision Tree
Accuracy	62.3%	75.7%
Precision	70%	82%
Recall	61%	77%

From the comparison in Table 6, the SVM model provided the best performance in terms of accuracy and precision. While the ANN and Decision Tree models produced similar but optimal performances in accuracy and precision which were over 75%.

Table 6: Performance Evaluation for ANN, SVM AND DT models.

	ACCURACY (%)	PRECISION (%)	SENSITIVITY (%)
ANN	78.6	78.6	78.6
SVM	86.0	86.7	79.4
DECISION TREE	78.5	76.4	71.75

5. Conclusion

The result of the HOG ANN training model used for recognizing the presence or absence of speedbump had an accuracy of 62.3%, a precision of 0.70 and a recall of 0.61, while that of the decision tree had an accuracy of 75.7%, a precision of 0.82 and a recall value of 0.77 as shown in Table 3. The results of the HE image processing model and ANN training model showed good accuracy of 83.3% for the sample images used for testing as shown in table 5, but the model was unable to detect the presence of anomalies in low-resolution images as the algorithm could not plot 9 BRISK points on images that had low pixel quality. The trained ANN model provided a total accuracy of 78.6%.

The fuzzy logic controller was able to regulate the speed of the vehicle in the presence of a speedbump and also change the direction of the vehicle in the presence of potholes. It was also observed that the use of the decision classification tree approach in detecting speedbump detection provided more accurate results when compared to that of Artificial Neural Network (ANN) with an accuracy of 65% in testing and training of the system, while the SVM models in the detection of potholes provided the best performance in pothole detection, but the ANN model provided an optimal performance for pothole detection with an accuracy of 78.6%.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

The authors wish to acknowledge the Tertiary Education Trust Fund (TETFUND), Nigeria for funding this work under the National Research Fund (NRF) grant titled “Novel Road Accident Monitoring and Prevention System for Nigerian Roads and Highways” with Grant Number: TETF/ES/DR&D-CE/NRF2020/SETI/95/VOL1.

References

[1] P. U. Miracle, J. A. Bala, A. M. Aibinu, T.A. Folorunso, "Intelligent Speed Bump Identification and Speed Control System for Autonomous Vehicles: A Conceptual Design," *1st International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS)*, 2021, 1-6, doi: 10.1109/ICMEAS52683.2021.9692403.

[2] J.A. Afolabi, K.T. Gbadamosi, (2017). “Road Traffic Crashes in Nigeria: Causes and Consequences.” Transport & Logistics: The International Journal, **17**(42), 41.

[3] N.C. Onyemachi, U.R. Ofoma, “The Public Health Threat of Road Traffic Accidents in Nigeria: A Call to Action.” Annals of Medical and Health Science Research, **6**(4), 199-204, 2016, doi: 10.4103/amhsr.amhsr_452_15.

[4] G.F. Pepple, C.S. Ejimadu, “Relationship between Road Traffic Accidents and Visual Status of Commercial Motor Vehicle Drivers in an Urban Area of Niger Delta Region of Nigeria.” Ophthalmology Research: An International Journal, 1-7, 2019.

[5] M. Bashiri, C.H. Fleming, “A platoon-based intersection management system for autonomous vehicles.” In IEEE Intelligent Vehicles Symposium, Proceedings **4**, 667–672, Redondo Beach, CA, USA, 2017. https://doi.org/10.1109/IVS.2017.7995794

[6] W. Zhang, E. Jenelius, H. Badia, “Efficiency of semi-autonomous and fully autonomous bus services in trunk-and-branches networks.” Journal of Advanced. Transportation, 2019. https://doi.org/10.1155/2019/7648735

[7] S. Adarsh, Real-Time Speed Bump Detection Using Image Segmentation for Autonomous Vehicles. Intelligent Computing, Information and Control System (ICICCS). 1039, 308, 2019.

[8] J. Youngtae, R. Seungki, “Enhanced Pothole Detection Algorithm for Blurry Images.” Journal of Emerging Trends in Computing and Information Sciences, **6** (5), 245, 2015.

[9] H.S. Yun, T.H. Kim, T.H. Park, “SpeedBump Detection for Autonomous Vehicles by Lidar and Camera.” Journal of Electrical Engineering & Technology, **14**(5), 2155-2162, 2019. https://doi.org/10.1007/s42835-019-00225-7.

[10] H.S. Yun, T.H.. Kim, T.H. Park, “Speed-Bump Detection for Autonomous Vehicles by Lidar and Camera." Journal of Electrical Engineering & Technology, **14**(5), 2155- 2162, 2019, https://doi.org/10.1007/s42835-019-00225-7.

[11] H. Bello-Salau, A.M. Aibinu, A.J. Onumanyi, S. Ahunsi, E.N. Onwuka, J.J. Dukiya, “Development of a road surface condition monitoring and database system.” In 2nd International Conference on Information and Communication Technology and Its Applications (ICTA 2018), Federal University of Technology, Minna, Nigeria, 6, 2018.

[12] A.B. Saeed, Madjid, T, “Autonomous Vehicles: Challenges, Opportunities and Implications.” Journal on Modern Transportation, **24**(8), 287, 2016, https://doi.org/10.1007/s40534-016-0117-3.

[13] U. Anand, “Fuzzy logic vision and control of autonomous vehicles.” *IPASJ Int. J. Comput. Sci*, **4**(1), 1-7, 2016.

[14] J.A. Bala, S.A. Adeshina, A.M. Aibinu, "Advances in Road Feature Detection and Vehicle Control Schemes: a Review," *2021 1st International Conference on Multidisciplinary Engineering and Applied Science (ICMEAS)*, 2021, pp. 1-6, doi: 10.1109/ICMEAS52683.2021.9692414.

[15] L.B. Ana, O. Ivan, O.J. Arturo, “Speedbump Detection on Roads using Artificial Vision.” Research in Computer Science, **148**(9),71-82, 2019.

[16] Y. Hyeong-Seok, K. Tae-Hyeong, P. Tae-Hyoung, ”Autonomous Vehicle by Lidar and Camera.” Journal of Electrical Engineering & Technology, **14**(5), 2155-2162, 2019, https://doi.org/10.1007/s42835-019-00225-7.

[17] H. Ziaul, "Basic Concept of GPS and Its Applications." IOSR Journal of Humanities and Social Science, **21**(3), 31, 2016.
[18] R. Seung-ki, K. Taehyeong, K. Young-RO, "Image-Based Pothole Detection System for ITS Service and Road Management System." Journal of Mathematical Problems in Engineering, **20**(15), 1-11, 2015, https://doi.org/10.1155/2015/968361.
[19] T. Amitha, J. Shiji, V. Sinu, P. Sooraj, N. Indrasena, "Real Time Pothole Detection and Road Monitoring System." International Journal of Innovative Research in Science, Engineering and Technology, **8**(6), 6740-6744, 2019.
[20] K. Ashwini, K. Kore, K. Akshada, M. Ashwini, "Pothole Detection System Based on Android Device and Ultrasonic Sensor." International Journal of Innovative Research in Computer and Communication Engineering, **6**(9), 7688-7692, 2018.
[21] H. Rahul, Z. Ashish, P. Nilesh, "Real time pothole detection technique using image processing and dimension estimation." International Journal of Engineering Development and Research, **7**(2), 108-111, 2019.
[22] S. Yamuna, S. Raja, M. Pavithra, M. Ranjitha, R. Shubashni, S. Sneha, "Automatic Pothole Detection System using Laser Technology." International Journal for Research in Applied Science & Engineering Technology (IJRASET), **5**(3), 1226-1231, 2017.
[23] Y. Wiseman, *Autonomous Vehicles*. Encyclopaedia of organizational knowledge, Administration, and Technologies, First Edition 2019.
[24] S.D. Pendleton, H. Andersen, X. Du, X. Shen, M. Meghjani, Y.H. Eng., M.H. Ang, "Perception, planning, control, and coordination for autonomous vehicles." *Machines*, **5**(1), 6, 2017, **https://doi.org/10.3390/machines5010006** .
[25] J. Cui, L. S. Liew, G. Sabaliauskaite, F. Zhou, "A review on safety failures, security attacks, and available countermeasures for autonomous vehicles." *Ad Hoc Networks*, *90*, 101823, 2019, https://doi.org/10.1016/j.adhoc.2018.12.006.
[26] A.M. Nascimento et al., "A Systematic Literature Review About the Impact of Artificial Intelligence on Autonomous Vehicle Safety," in IEEE Transactions on Intelligent Transportation Systems, **21**, (12), 4928-4946, Dec. 2020, doi: 10.1109/TITS.2019.2949915.
[27] R. Du, G. Qiu, K. Gao, L. Hu, L. Liu, "Abnormal Road Surface Recognition Based on Smartphone Acceleration Sensor." *Sensors*, **20**(2), 451, 2020, **https://doi.org/10.3390/s20020451**.
[28] O. Adeyemi, M. Irhebhude, A. Kolawole, "Speed Breakers, Road Marking Detection and Recognition Using Image Processing Techniques." *Advances in Image and Video Processing*, **7**(5), 30-42, 2019.
[29] R. Rajamani, Vehicle dynamics and control. Springer Science & Business Media, 2011.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 4, 01-08 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Lung Cancer Tumor Detection Method Using Improved CT Images on a One-stage Detector

Young-Jin Park, Hui-Sup Cho*

Division of Electronics and Information System, DGIST, Daegu, 42988, Republic of Korea

ARTICLE INFO

Article history:
Received: 26 May, 2022
Accepted: 23 June, 2022
Online: 18 July, 2022

Keywords:
Lung Cancer Tumor Detection
DICOM
CAD system

ABSTRACT

Owing to the recent development of AI technology, various studies on computer-aided diagnosis systems for CT image interpretation are being conducted. In particular, studies on the detection of lung cancer which is leading the death rate are being conducted in image processing and artificial intelligence fields. In this study, to improve the anatomical interpretation ability of CT images, the lung, soft tissue, and bone were set as regions of interest and configured in each channel. The purpose of this study is to select a detector with optimal performance by improving the quality of CT images to detect lung cancer tumors. Considering the dataset construction phase, pixel arrays with Hounsfield units applied to the regions of interest (lung, soft tissue, and bone region) were configured as three-channeled, and a histogram processing the technique was applied to create a dataset with an enhanced contrast. Regarding the deep learning phase, the one-stage detector (RetinaNet) performs deep learning on the dataset created in the previous phase, and the detector with the best performance is used in the CAD system. In the evaluation stage, the original dataset without any processing was used as the reference dataset, and a two-stage detector (Faster R-CNN) was used as the reference detector. Because of the performance evaluation of the developed detector, a sensitivity, precision, and F1-score rates of 94.90%, 96.70%, and 95.56%, respectively, were achieved. The experiment reveals that an image with improved anatomical interpretation ability improves the detection performance of deep learning and human vision.

1. Introduction

Lung cancer is the leading cause of cancer-related deaths (18.0% of the total cancer deaths), followed by colorectal (9.4%), liver (8.3%), stomach (7.7%), and female breast (6.9%) cancers [1]. Early diagnosis and treatment may save lives. Although computerized tomography (CT) scan imaging is the best imaging technique in the medical field, it is difficult for doctors to interpret and identify cancer using CT scan images [2]. In addition, because lung cancer detection can increase the detection time and error rate depending on the skill of the doctor, computer-aided diagnosis (CAD) studies to passively assist the detection are on image segmentation, denoising, and 3D image processing using image processing [3] and neural network optimization [4–6].

To improve the anatomical interpretation ability of CT images, this study is designed to improve the cognitive ability of detectors by setting lung, soft tissue, and bone as the regions of interest and utilizing a dataset which is visually easy to distinguish between each region's features in deep learning.

In the dataset construction phase, the dataset was constructed by preprocessing the Digital Imaging and Communications in Medicine (DICOM) files provided by the Lung-PET-CT-Dx dataset [7,8]. The region of interest to which Hounsfield Unit (HU) windowing is applied is composed of a fundamental three-channel dataset (3ch-ORI) to generate an image, and the characteristics of each region can be recognized in one image. Because this process can improve the cognitive ability to visually classify the regions of interest, it is relevant to feature extraction through anatomical analysis in the training process of a neural network, mimicking the human brain and lung cancer detection process. In addition, to enhance the contrast of the 3ch-ORI, a detector with optimal performance was selected by comparing the results of deep learning on a dataset (3ch-CLAHE) to that which the contrast-limited adaptive histogram equalization (CLAHE) was applied. Considering the deep learning phase, deep learning was performed using reference datasets and a reference detector to determine the optimal train customization settings.

The Lung-PET-CT-Dx dataset used in this study provides version 1 (release date: June 1, 2020) to version 5 datasets (release

*Corresponding Author: Hui-Sup Cho, mozart73@dgist.ac.kr

https://dx.doi.org/10.25046/aj070401

date: December 22, 2020); nonetheless, it has been released recently and the related studies [9,10] are insufficient. Therefore, in the evaluation stage, the raw original dataset (1ch-ORI) was used as a reference for comparison. Regarding performance evaluation, the intersection of union (IoU) was calculated to achieve high-level results with a sensitivity, precision, and F1-score rates of 94.90%, 96.70%, and 95.56%, respectively.

2. Related studies

This chapter describes the existing studies that have used methods such as structural separation, noise removal, and three-dimensional (3D) technology for visualization of CT images using the Lung Image Database Consortium and Image Database Resource Initiative dataset (LIDC-IDRI). A novel objective evaluation framework for nodule detection algorithms using the largest publicly available LIDC-IDRI dataset or subset lung nodule analysis 2016 (LUNA16) is a challenge [11]. This set of additional nodules for further development of the IDRI-IDRI dataset that was initiated by the National Cancer Institute (NCI) [12,13] have been released.

In [14], they proposed a novel pulmonary nodule detection CAD system and developed to detect nodule candidates using improved Faster R-CNN. They have archived sensitivity of 94.6%. In [15], the noise present in the CT image was removed by applying the weighted mean histogram equalization (WMHE) method, and the quality of the image was improved using the improved profit clustering technique. Consequently, minimum classification errors of 0.038% and 98.42% accuracies were obtained. The method [16] using the modified gravity search algorithm (MGSA) for the classification and identification of lung cancer in CT images achieved a sensitivity, specificity, and accuracy of 96.2%, 94.2%, and 94.56%, respectively, owing to the application of the optimal deep neural network (ODNN). Furthermore, a threshold-based technique for separating the nodules of lung CT images from other structures (e.g., bronchioles and blood vessels) was proposed [17], and from the evaluation, a sensitivity of 93.75% was achieved. The 3D region segmentation of the nodule in each lung CT image achieved 83.98% [18] because of image reconstruction using the sparse field method. In many other studies, many methods for detection and classification using image processing and deep learning have been proposed, and their performance is quite high. These studies aimed at assisting medical staff with visualization based on image processing. Therefore, various methods need to be continuously studied for CAD systems, where even a 0.01% performance improvement is significant. In this study, the improved CT image is used for deep learning to improve the anatomical analysis ability of each region of interest in the CT image. The dataset aimed at enhancing the quality of the CT image in the pre-processing of the DICOM file without using a complicated image processing method to achieve a high level of result. If the improved CT image is applied to the method proposed in previous studies, a better performance is expected.

3. Materials and Methods

3.1. Dataset construction phase

The preprocessing step of this study includes obtaining a purified CT image through structural analysis of the Lung-PET-CT-Dx dataset, pixel range normalization of the DICOM file, and HU windowing for each region of interest. The Lung-PET-CT-Dx dataset consists of CT and PET-CT DICOM images of lung cancer subjects with XML annotation files that indicate tumor location with bounding boxes. The subjects were grouped according to tissue histopathological diagnosis. Patients with names/IDs containing letter 'A' were diagnosed with Adenocarcinoma, 'B' corresponded to Small Cell Carcinoma, 'E' indicated Large Cell Carcinoma, and 'G' corresponded to Squamous Cell Carcinoma [19].

Object detection performs classification and localization to obtain detection and classification results for each class. However, because this study focuses on the performance of detecting lung cancer, one class was evaluated using only the adenocarcinoma class, without using a dataset with a different number for each class. Therefore, the results of the classification are meaningless, and only the results of localization are used to evaluate the performance. The adenocarcinoma class consists of sub-directories divided for each slice in 265 main directories, and 21 main directories that do not have annotation files or do not match the annotation xml information are excluded from the dataset configuration. In addition, the DICOM files existing in each directory were merged into one directory for easy management and quick data access. The annotation files matching 1:1 with the DICOM file were stored in one common csv file, and after xml parsing, they were stored in the DICOM file. Unlike greyscale images which are in a range of 0 to 255, DICOM files are converted to 12-bit pixel arrays, and DICOM files are composed of HU [20] (a unit that expresses the degree of attenuation of X-rays when penetrating the body). Therefore, as shown in Figure 1, the DICOM file can be viewed more clearly by normalization and HU windowing.

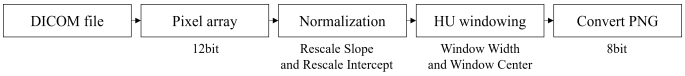


Figure 1: Process steps of the DICOM file

First, the 12-bit (4096 level) pixel array extracted from the DICOM image was normalized according to the unit defined in the HU. Depending on the CT equipment, the pixel range was stored as 0 to 4095 or -2048 to 2047. Using Equation (1), linear transformation was applied to the 'Rescale slope' and 'Rescale intercept' fields to remap the image pixel:

$$Out\ pixel = rescale\ slope * input\ pixel + rescale\ intercept \quad (1)$$

For example, as shown in the figure, in the 0-to-4095-pixel range, the rescale intercept has a value of − -2048, and the rescale slope has a value of 1; therefore Equation (1) is used to convert it to a value in the range of − -2048 to 2047. In contrast, the rescale intercept of the DICOM file stored in the pixel range of -2048 to 2047 is 0; hence, there is no change even if the above formula is used. Therefore, normalization is applied to the range shown in Figure 2, and all DICOM files are placed within the same pixel range. Considering the reference, dataset, the rescale slope and rescale intercept attributes do not exist in the properties of the DICOM file, they are excluded from the dataset configuration.

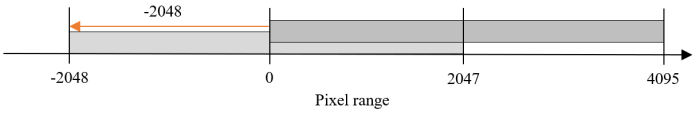


Figure 2: Brightness settings for DICOM image

Subsequently, the normalized pixel array performs the HU windowing process by applying the window width and center properties to each region of interest as shown in Table 1.

Table 1: Window setting using the Hounsfield Unit

Window	Lung	Soft tissue	Bone
Window Center	-700	40	500
Window Width	1400	350	2000

The overall flow of the data construction phase that processes the DICOM file and composes each dataset is shown in Figure 3.

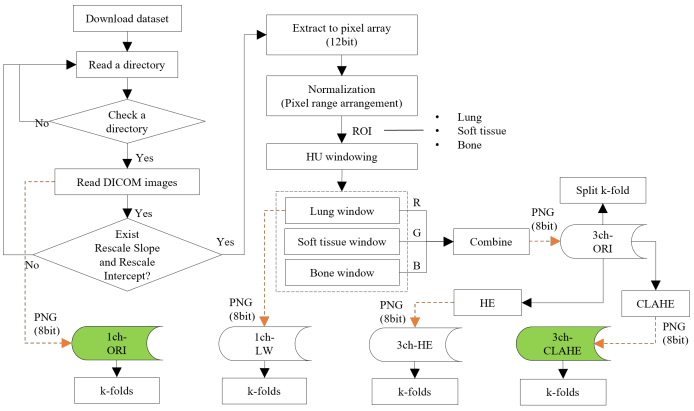


Figure 3: Flow of the dataset construction phase

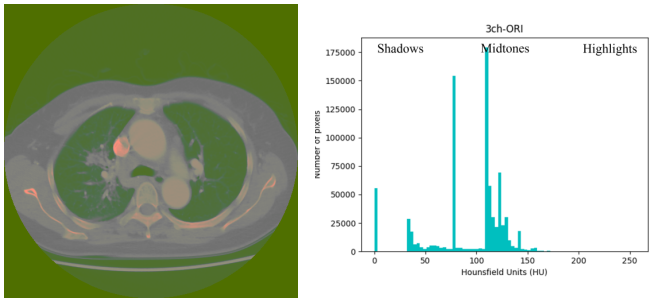
In dataset composition step, a method of composing an image in three-channel and generating images with enhanced contrast by applying CLAHE is described. The CLAHE is an algorithm that uniformly divides an image and distributes pixels of a specific height to each area. After setting the clip limit (the threshold), the height of the histogram was limited.

Table 2 lists the composition and use of the dataset employed. The 1ch-ORI is a dataset consisting of images converted directly into a PNG format from the original DICOM image without any processing and is used as a reference in the experiment. On the contrary, the 3ch-HE, which applies histogram equalization (HE) equally to all pixels, is used as a reference dataset for comparison with the 3ch-CLAHE. In addition, the 3ch-ORI is a fundamental three-channel dataset before the application of equalization.

Table 2: Datasets used for the experiment

Name	Configuration	Usage
1ch-ORI	Raw dataset	Reference dataset
3ch-ORI	3-channel dataset before equalization	Fundamental dataset of 3-channel
3ch-CLAHE	3-channel dataset after CALHE	Proposed dataset
3ch-HE	3-channel dataset after HE	Reference dataset for equalization

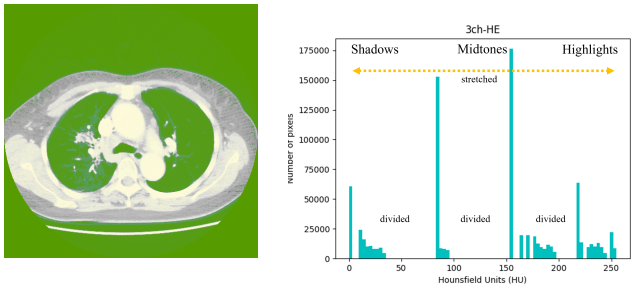
Contrast enhancement using image processing can acquire more detailed information by improving visual recognition ability; thereby, increasing the analysis ability of CT images in the process of human visual and feature recognitions in deep learning. Because methods such as linear combination [21] used for contrast enhancement use multiple images for one-channel, it may affect the contrast range when observing the HU-applied window. Contrast enhancement is a specific characteristic enhancement of image enhancement processing. Histogram equalization is a popular method for image contrast enhancement [22]. Therefore, in this study, the histogram processing technique is applied to the 3ch-ORI, in which the window of interest is set as the data for each channel. However, histogram equalization is not the best method for contrast enhancement because the mean brightness of the output image is significantly different from that of the input image [23]. Brightness is used as an important feature along with shape information when configuring channels, and it is difficult to distinguish between regions because CT images that are not pre-processed may be dark and noise may exist as shown in Figure 4-(A). The histogram can be divided into left, right, and midtones. In Figure 4-(B), where equalization is not applied, the highlighted area is empty; hence, it is difficult to distinguish each area as dark as shown in Figure 4-(A).



(a) 3ch-ORI image b) Histogram

Figure 4: Exemplary image and histogram of 3ch-ORI

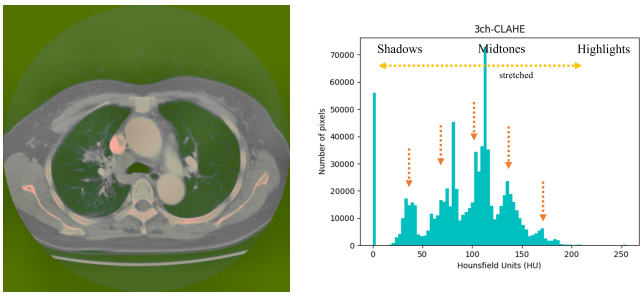
If HE is applied to all pixels at once, equalization is performed indiscriminately. This may cause noise in extremely dark or bright areas or loss of necessary information. Considering Figure 5-(B), where histogram equalization is applied, the number of pixels in the highlighted area is increased (yellow dotted arrows), and it can be seen that Figure 5-(A) affects the increase in pixel intensity and brightness. However, the midtone area decreased, resulting in the spread of shadow and highlight directions. Because it is more difficult to classify each area owing to the addition or loss of information in a specific area, the CLAHE method is used in this study to prevent noise overamplification.



(a) 3ch-HE image (b) Histogram

Figure 5: Example image and histogram of 3ch-HE

In Figure 6, to which CLAHE is applied, it can be observed that the level of pixel areas is spread around the midtone area (yellow dotted arrows), the number of pixels is evenly distributed, and the average value is decreased (orange dotted arrows), enhancing the contrast.



(a) 3ch-CLAHE image (b) Histogram

Figure 6: Exemplary image and histogram of the 3ch-CLAHE

In Figure 4, the contrast is too low to distinguish each region using human eyes; therefore, Figures 5 and 6 with enhanced contrast are used for the experiment. Considering a human point of view, the characteristics of each area in Figure 6 can be distinguished better than in Figure 5; nevertheless, an accurate judgment is made by comparing the deep learning results. Equalization of the histogram using brightness rather than the color of the image was applied after configuration as three three-channels because if three three-channels were configured by applying them to each gray image, different contrasts could be applied to each channel. Therefore, because it is different from the intended image when combined with the color model, three-channel, unintended results of anatomical organs, structures, or artifacts in the human body can have a significant impact on CT image analysis.

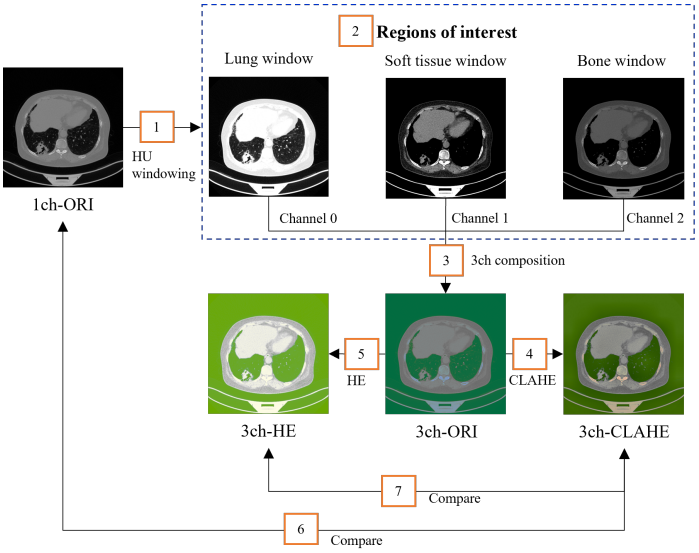


Figure 7: Dataset construction flow

The configuration flow of the dataset is shown in Figure7. The three types of areas of interest (number 2 in the orange box), lung, soft tissue, and bone window, appear clearly after HU windowing. However, it is difficult to anatomically distinguish each area owing to the addition or loss of the specific areas.

After applying each method in the 12-bit pixel array, each dataset was converted to an 8-bit PNG format and stored on a disk. A total of 13,233 images were divided into train- and test-sets in a ratio of 8:2 (10,586:2,647). In addition, because data bias in each dataset can affect the evaluation results of the deep learning model, five-fold cross-validation was performed as shown in Figure 8 to select the optimal fold to be used in the experiment.

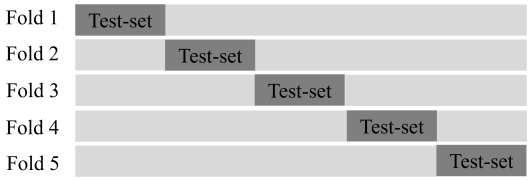


Figure 8: Dataset composition of the cross-validation

3.2. Deep Learning phase

In general, object detection is categorized into one- and two-stage detectors as shown in Table 3. One-stage detectors perform classification and localization concurrently. Therefore, they are fast; however, they are low in accuracy. The two-stage detectors use the Legion Proposal Network (RPN) to select candidate areas where objects are expected to be detected, making them slow; nonetheless, they are high in accuracy. The most recent one-stage detectors exceed the accuracy of two-stage detectors; hence, classification according to accuracy is less meaningful.

In this study, we compared RetinaNet [24] using two-dimensional (2D) image-based anchor-based detectors and Faster R-CNN [25–27] as a reference, modified and utilized Faster R-CNN [28,29] and RetinaNet [30] cloned from the GitHub repository to create a model.

Table 3: Comparison of one- and two-stage detectors

Detector	Anchor based	Detector (Deep learning algorithm)
one-stage	O	YOLO-v1, v2, v3 (2016), SSD (2016) – RetinaNet (2017)
	X	CornerNet (2018) – ExtremeNet (2019) – CenterNet (2019)
two-stage	O	R-CNN (2013) – Fast R-CNN (2015) – Faster R-CNN (2015) – Mask RCNN (2017)

In deep learning, train customization is an important experimental step for selecting the optimal dataset configuration and detector. In this study, ResNet-{50, 101, 152} pretrained with ImageNet [31] was used for transfer learning, and random flip and data shuffle were applied for data augmentation. Moreover, for stable optimization, Adam [32] was used as the optimizer, and the learning rate was set to 1e-5. Considering the reference, Faster R-CNN, which is divided into two stages of RPN and classifier, sets the customization of both elements similarly. Furthermore, the epoch size was set to 500, and the batch size was set to four or eight. Experiments were performed in the environment of Python 3.7, Cuda-10, and a GPU on a 64bit Ubuntu18.04LTS operating system.

After generating a model using the training set and evaluating it using the test set, the overall flow of the deep learning phase (the process of determining the final performance) is shown in Figure 9.

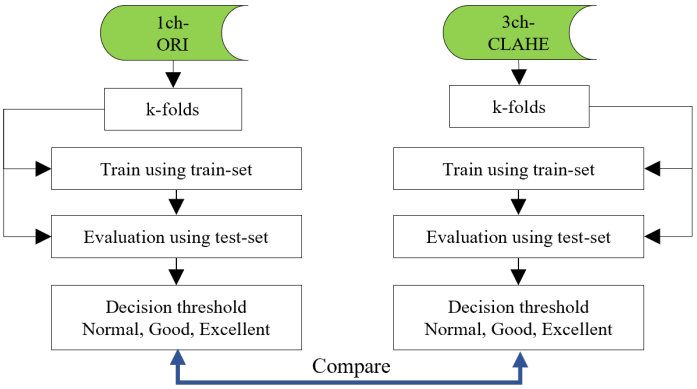


Figure 9: Flows of the deep learning phase

Object detection performs both classification, which classifies objects in the bounding box, and localization, which is a regression process for finding the bounding box. However, because the classification performance of this study is recognized as only one class using a single class, it is not reflected in the evaluation, and the IoU of the bounding box detected by localization and the ground truth (GT) box, which is the annotation information, is calculated. Considering the reference dataset, when multiple bounding boxes are detected in the test-set image, the bounding box with the highest IoU is selected as the IoU of the image. In many cases, natural scene images can be judged by predicting low-level detection results of objects such as people or automobiles with the human eye. However, lung cancer tumors have a non-standard shape; thereby, requiring a higher performance. In this study, to increase the reliability of the detection performance as shown in Table 4, the decision thresholds for each image are set to be narrower than those of the natural scene image for final judgment. However, the narrower the threshold setting range is, the higher the reliability and the lower the statistical evaluation results. Because the achievement result can be relatively decreased, it must be carefully set according to the field of use

Table 4: Setting of the decision threshold

Decision	Natural scene	Proposed
Normal	>=0.50	>=0.60
Good	>=0.70	>=0.75
Excellent	>=0.90	>=0.90

Table 5: Confusion matrix

Name	Threshold	Description
TP	>= 0.6	Lung cancer exists, detected correctly
TN	No use	No lung cancer exists, identified correctly
FP	< 0.6	No lung cancer exists, detected incorrectly
FN	-	Lung cancer exists, missed

The evaluation of models using the test-set images uses the outcomes of four kinds of statistical confusion matrices as shown in Table 5. True Positive (TP) is determined to correctly detect lung cancer tumors when the IoU is 0.6 or higher, and false positive (FP) is determined to be incorrectly detected when the IoU value of the detected bounding box is 0.6 or less. Considering the false negatives (FN), because there is no detected bounding box, the IoU for the GT box cannot be calculated; therefore, a value excluding TP from the total dataset is used. Regarding the reference dataset, negative (TN) is used when lung cancer tumors do not exist and is not used in the field of object detection for a dataset consisting of one class in which the GT box exists in all the test-set images. The total test-set image length and number of GT boxes were the same.

The experimental results were evaluated using statistical performance measurement methods such as sensitivity, precision, and F1-score. Sensitivity represents the predicted positive among all positives and is calculated using Equation 2:

$$Sensitivity\ (or\ Recall) = TP / (TP + FN) \quad (2)$$

Precision is the proportion of true positives among the predicted positives, calculated using Equation 3:

$$Precision = TP / (TP + FP) \quad (3)$$

Because the indicators of sensitivity and precision are inversely proportional, it cannot be concluded that a high value of one of them has good performance. Finally, the F1-score is a harmonic mean that considers both sensitivity and precision. The optimal value is one, and the higher the value is, the better the performance, and it is calculated using Equation 4:

$$F1\text{-}score = 2 * (Precision * Recall) / (Precision + Recall) \quad (4)$$

3.3. Results and Discussion

In this chapter, experiments are conducted using each evaluation element, as shown in Figure 10, and the results are discussed.

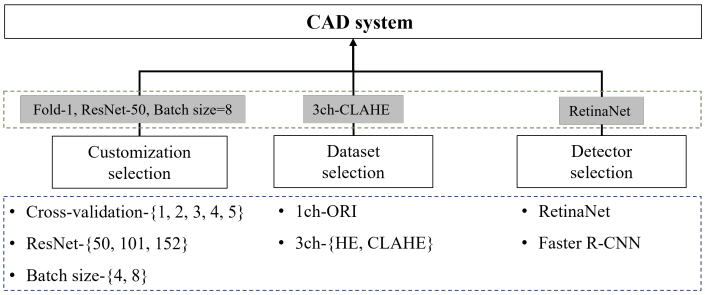


Figure 10: Evaluation factors and flows.

Train customization selection (blue dotted box) through cross-validation experiments, ResNet-depth, and batch size. The data selection and detector selection steps describe the process of selecting a detector with the best performance (green dotted box) through evaluation and using it as a CAD system. First, the experimental results of train customization according to the conditions of cross validation, ResNet-depth, and batch size for the selection of a detector to be used in the CAD system are shown in Table 6. When the Fold-Num was set to one and the batch size was set to eight, the highest result was a sensitivity of 94.9% and precision of 96.7% in ResNet-50, and the F1-score result was the highest in ResNet-101 with 95.8%. Considering the reference dataset, the result of the F1-score in ResNet-50 was 95.6%, being the second highest result.

Table 6: Comparison of ResNet-50 that changed the conditions of customization and ResNet-{101, 152}

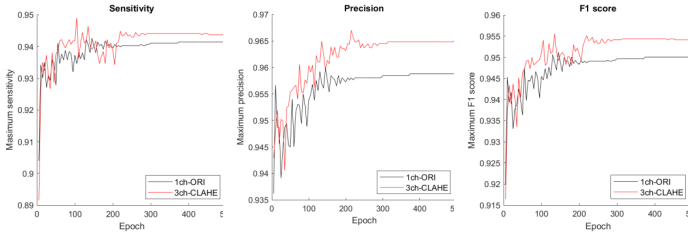
Fold	Depth	Batch Size	Sensitivity (%)			Precision (%)			F1-score (%)		
			Normal	Good	Excellent	Normal	Good	Excellent	Normal	Good	Excellent
1	50	8	94.9	72.3	14.5	96.7	73.9	14.8	95.6	73.1	14.6
2	50	8	93.7	72.9	13.8	96.3	75.6	14.3	94.7	74.2	14.1
3	50	8	93.8	70.3	13.4	96.4	73.1	13.8	94.8	71.5	13.6
4	50	8	93.1	72.0	13.9	95.3	73.7	14.3	94.2	72.9	14.1
5	50	8	93.0	72.4	13.9	96.0	74.8	14.4	94.4	73.6	14.1
1	101	8	94.8	72.5	13.7	96.7	74.0	14.0	95.8	73.2	13.8
1	152	8	94.6	72.8	14.3	96.5	74.6	14.7	95.4	73.7	14.5
1	50	4	94.4	72.5	14.8	96.6	74.7	15.3	95.6	73.6	15.0
1	101	4	94.5	72.6	14.9	96.6	74.5	15.3	95.2	73.5	15.1
1	152	4	94.6	72.8	14.3	96.5	74.6	14.7	95.4	73.7	14.5

Considering the experimental results, Fold-1, ResNet-50, and batch size: eight (which show the best overall performance), were selected the customization setting values of the detector. Since the performance was the best when using the Fold-1 dataset trained using the ResNet-50 neural network, the experiments using the ResNet-101 and ResNet-152 neural networks were compared with the ResNet-50 using only the Fold-1 dataset.

Table 7 shows a comparison between the 3ch-CLAHE and reference datasets using customization setting values in ResNet-50. First, comparing the results with 1ch-ORI (an unprocessed CT image) showed a performance improvement in the sensitivity (+0.74%), precision (+0.70%), and F1-score (+0.44%). Additionally, comparing the result with 3ch-HE, which applied HE to all pixels, showed a performance improvement in the sensitivity (+0.94%), precision (+0.33%), and F1-score (+0.60%). Therefore, it was found that the dataset in which each ROI was composed of three three-channels and CLAHE applied for contrast enhancement had a significant effect on the performance improvement of deep learning.

Table 7: Comparison of proposed dataset with reference datasets

Dataset	Sensitivity (%)	Precision (%)	F1-score (%)
3ch-CLAHE	94.90	96.70	95.56
1ch-ORI	94.26	96.00	95.12
3ch-HE	93.96	96.37	94.96

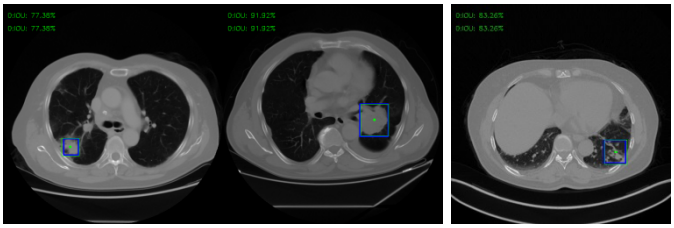


(a) Sensitivity (b) Precision (c) F1-score

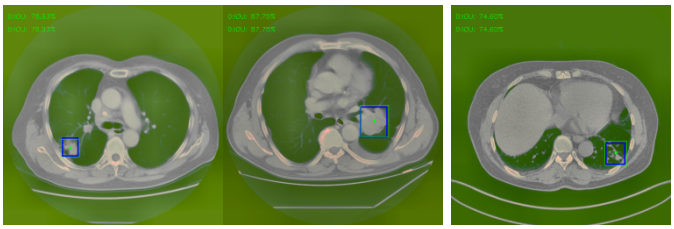
Figure 11: Comparison of the proposed dataset (3ch-CLAHE) with the reference dataset (1ch-ORI)

The dataset with the best performance can be obtained from the experimental results in the table; however, visual performance analysis using a graph as shown in Figure 11 can be used as a tool for selecting an appropriate model and determining when to stop learning at the highest performance. The maximum measurement value (y-axis) of each epoch (x-axis) for the comparison datasets 1ch-ORI and 3ch-CLAHE appeared before approximately 200 epochs; nonetheless, stable learning results appeared after approximately 250 epochs. Two hundred and fifty epochs indicate that the learning efficiency is the best.

Figure 12 shows only a few cases among the actual detection results using the test set of 1ch-ORI and 3ch-CLAHE. Using the coordinate values of the GT box (blue box) and bounding box (green box) shown in each image, IoU was calculated and used for the evaluation.

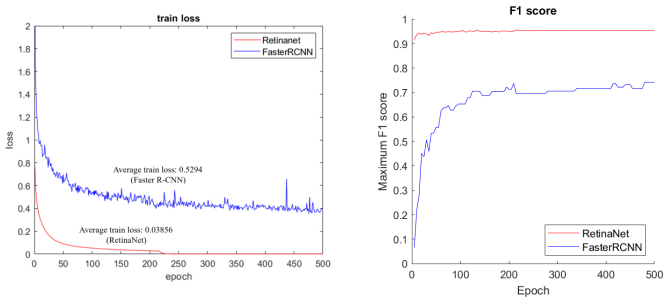


(a) Lung cancer detected sample images of 1ch-ORI



(b) Lung cancer detected sample images of 3ch-CLAHE

Figure 12: Examples of detected result images



(a) Train loss (b) F1 score.

Figure 13: Comparison of RetinaNet with the Faster R-CNN

Finally, the performance was compared to Faster R-CNN, which was used as a reference detector to select the final detector. As shown in Figure 13-(A), which compares the average train loss, RetinaNet shows a result of 0.03856. This a big difference from the Faster R-CNN, which shows a result of 0.5294, and the difference in performance and stability during the training process is reflected in the evaluation result. As shown in Figure 13-(B), RetinaNet (red line) shows a high F1-score of 0.9 or higher. However, the train loss of Faster R-CNN (blue line) is unstable, and the F1-score shows a result between 0.7 and 0.8 in Figure 13-(B).

4. Conclusion

In this study, to detect lung cancer quickly and accurately, we attempted to improve the detection performance by improving the image quality. Novel lung cancer detection methods using image processing provide a high level of accuracy by applying noise removal from CT images, segmentation techniques, and methods using 3D images for deep learning. The segmentation technique,

which is mainly used to find a small nodule, has the advantage of concentrating on the area. Nonetheless, it also consumes a lot of application time and resources and has the disadvantage that the shape and boundary line may appear irregularly for each image. Because the 3D visualization method of CT images can represent the lungs more realistically, it consumes a lot of resources compared to the other methods although it is used to detect the shape of the lesion with more details.

In this study, we propose a CLAHE-based three-channel dataset construction method that automatically detects lung cancer tumors. Although this method processes CT images in a relatively simple way compared to the novel lung cancer detection methods, high performance has been confirmed through several comparative experiments, and a better performance is expected when applied to the methods of other studies. In addition, the customization of the deep learning process is as important as the configuration of the dataset, and the experimental results reveal that the CT image with improved human visual perception is important for the neural network that mimics the human brain. However, owing to the lack of reference studies, the study was conducted with the goal of improving the performance of the original dataset, and achieved a sensitivity, precision and F1-score rates of 94.90%, 96.70%, and 95.56%. In the results of this study, the one-stage detector showed better performance in train stability and object detection rate than the two-stage detector. Since the images used in this study are medium or small size objects, different results may appear when big size objects are detected using a natural scene dataset, etc.

In addition, although this study cannot be directly compared to studies using the popular public dataset, it serves as a prior study using a dataset that has insufficient comparative studies.

5. Data Availability

The CT scan images used to support the findings of this study have been collected from the Cancer Imaging Archive (TCIA) (link: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70224216)

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This study was supported by the Technology Innovation Program (20011875, Development of AI-Based Diagnostic Technology for Medical Imaging Devices) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

References

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA: A Cancer Journal for Clinicians, **71**(3), 209–249, 2021, doi:10.3322/caac.21660.

[2] S. Makaju, P.W.C. Prasad, A. Alsadoon, A.K. Singh, A. Elchouemi, "Lung Cancer Detection using CT Scan Images," Procedia Computer Science, **125**, 107–114, 2018, doi:10.1016/J.PROCS.2017.12.016.

[3] D. Sharma, G. Jindal, "Identifying lung cancer using image processing techniques," in International Conference on Computational Techniques and Artificial Intelligence (ICCTAI), Citeseer: 872–880, 2011.

[4] W. Sun, B. Zheng, W. Qian, "Computer aided lung cancer diagnosis with deep learning algorithms," in Medical imaging 2016: computer-aided diagnosis, SPIE: 241–248, 2016.

[5] A. El-Baz, G.M. Beache, G. Gimel'farb, K. Suzuki, K. Okada, A. Elnakib, A. Soliman, B. Abdollahi, "Computer-aided diagnosis systems for lung cancer: challenges and methodologies," International Journal of Biomedical Imaging, 2013, 2013.

[6] Y. Abe, K. Hanai, M. Nakano, Y. Ohkubo, T. Hasizume, T. Kakizaki, M. Nakamura, N. Niki, K. Eguchi, T. Fujino, N. Moriyama, "A computer-aided diagnosis (CAD) system in lung cancer screening with computed tomography," Anticancer Research, **25**(1 B), 2005.

[7] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," Journal of Digital Imaging, **26**(6), 2013, doi:10.1007/s10278-013-9622-7.

[8] P.. W.S.. L.T.. L.J.. H.Y.. & W.D. Li, A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis, doi:https://doi.org/10.7937/TCIA.2020.NNC2-0461.

[9] S. Mazza, D. Patel, I. Viola, "Homomorphic-encrypted volume rendering," IEEE Transactions on Visualization and Computer Graphics, **27**(2), 2021, doi:10.1109/TVCG.2020.3030436.

[10] D. Gu, G. Liu, Z. Xue, "On the performance of lung nodule detection, segmentation and classification," Computerized Medical Imaging and Graphics, 89, 2021, doi:10.1016/j.compmedimag.2021.101886.

[11] A.A.A. Setio, A. Traverso, T. de Bel, M.S.N. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M.E. Fantacci, B. Geurts, R. van der Gugten, P.A. Heng, B. Jansen, M.M.J. de Kaste, V. Kotov, J.Y.H. Lin, J.T.M.C. Manders, A. Sóñora-Mengana, J.C. García-Naranjo, E. Papavasileiou, M. Prokop, M. Saletta, C.M. Schaefer-Prokop, E.T. Scholten, L. Scholten, M.M. Snoeren, E.L. Torres, J. Vandemeulebroucke, N. Walasek, et al., "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," Medical Image Analysis, **42**, 2017, doi:10.1016/j.media.2017.06.015.

[12] LIDC, The Lung Image Database Consortium image collection, Https://Wiki.Cancerimagingarchive.Net/Display/Public/LIDC-IDRI,.

[13] I.S.G. Armato, H. MacMahon, R.M. Engelmann, R.Y. Roberts, A. Starkey, P. Caligiuri, G. McLennan, L. Bidaut, D.P.Y. Qing, M.F. McNitt-Gray, D.R. Aberle, M.S. Brown, R.C. Pais, P. Batra, C.M. Jude, I. Petkovska, C.R. Meyer, A.P. Reeves, A.M. Biancardi, B. Zhao, C.I. Henschke, D. Yankelevitz, D. Max, A. Farooqi, E.A. Hoffman, E.J.R. Van Beek, A.R. Smith, E.A. Kazerooni, G.W. Gladish, et al., "The Lung Image Database Consortium ({LIDC}) and Image Database Resource Initiative ({IDRI}): A completed reference database of lung nodules on {CT} scans," Medical Physics, **38**(2), 2011.

[14] J. Ding, A. Li, Z. Hu, L. Wang, "Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2017, doi:10.1007/978-3-319-66179-7_64.

[15] P.M. Shakeel, M.A. Burhanuddin, M.I. Desa, "Lung cancer detection from CT image using improved profuse clustering and deep learning instantaneously trained neural networks," Measurement: Journal of the International Measurement Confederation, **145**, 2019, doi:10.1016/j.measurement.2019.05.027.

[16] S.K. Lakshmanaprabu, S.N. Mohanty, K. Shankar, N. Arunkumar, G. Ramirez, "Optimal deep learning model for classification of lung cancer on CT images," Future Generation Computer Systems, **92**, 2019, doi:10.1016/j.future.2018.10.009.

[17] N. Khehrah, M.S. Farid, S. Bilal, M.H. Khan, "Lung nodule detection in CT images using statistical and shape-based features," Journal of Imaging, **6**(2), 2020, doi:10.3390/jimaging6020006.

[18] S. Saien, H.A. Moghaddam, M. Fathian, "A unified methodology based on sparse field level sets and boosting algorithms for false positives reduction in lung nodules detection," International Journal of Computer Assisted Radiology and Surgery, **13**(3), 2018, doi:10.1007/s11548-017-1656-8.

[19] TCIA, A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis (Lung-PET-CT-Dx) , Https://Wiki.Cancerimagingarchive.Net/Pages/Viewpage.Action?PageId=70224216,.

[20] S.J. DenOtter TD, Hounsfield Unit, StatPearls Publishing, 2020, doi:10.32388/aavabi.

[21] S. Ullman, R. Basri, "Recognition by Linear Combinations of Models," IEEE Transactions on Pattern Analysis and Machine Intelligence, **13**(10), 1991, doi:10.1109/34.99234.

[22] Scott E Umbaugh, Computer Vision and Image Processing, Prentice Hall:

New Jersey 1998, 1988.
[23] O. Patel, Y. P. S. Maravi, S. Sharma, "A Comparative Study of Histogram Equalization Based Image Enhancement Techniques for Brightness Preservation and Contrast Enhancement," Signal & Image Processing : An International Journal, **4**(5), 2013, doi:10.5121/sipij.2013.4502.
[24] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, "Focal Loss for Dense Object Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 2020, doi:10.1109/TPAMI.2018.2858826.
[25] R. Girshick, "Fast R-CNN," Proceedings of the IEEE International Conference on Computer Vision, 2015 Inter, 2015.
[26] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014, doi:10.1109/CVPR.2014.81.
[27] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, **39**(6), 2017, doi:10.1109/TPAMI.2016.2577031.
[28] Kentaro Yoshioka, FRCNN, Https://Github.Com/Kentaroy47/Frcnn-from-Scratch-with-Keras,.
[29] Young-Jin Kim, FRCNN, Https://Github.Com/You359/Keras-FasterRCNN,.
[30] Yann Henon, pytorch-retinanet, Https://Github.Com/Yhenon/Pytorch-Retinanet,.
[31] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2010, doi:10.1109/cvpr.2009.5206848.
[32] D.P. Kingma, J.L. Ba, "Adam: A method for stochastic optimization," in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 1, 01-11 (2023)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Meta-heuristic and Heuristic Algorithms for Forecasting Workload Placement and Energy Consumption in Cloud Data Centers

Amine Bouaouda*, Karim Afdel, Rachida Abounacer

[1]*Department of Computer Science, LabSIV, Ibn Zohr University, Agadir, 80000, Morocco*

ARTICLE INFO

Article history:
Received: 30 August, 2022
Accepted: 03 December, 2022
Online: 24 January, 2023

Keywords:
Data center Energy
Forecasting Energy
Container Placement
Genetic Algorithm
First Fit Decreasing

ABSTRACT

The increase of servers in data centers has become a significant problem in recent years that leads to a rise in energy consumption. The problem of high energy consumed by data centers is always related to the active hardware especially the servers that use virtualization to create a cloud workspace for the users. For this reason, workload placement such as virtual machines or containers in servers is an essential operation that requires the adoption of techniques that offer practical and best solutions for the workload placement that guarantees an optimization in the use of material resources and energy consumption in the cloud. In this article, we propose an approach that uses heuristics and meta-heuristics to predict cloud container placement and power consumption in data centers using a Genetic Algorithm (GA) and First Fit Decreasing (FFD). Our algorithms have been tested on CloudSim and the results showed that our methods gave better and more efficient solutions, especially the Genetic Algorithm after comparing them with Ant Colony Optimization (ACO) and Simulated Annealing (SA).

1 Introduction

This paper is an extension of work originally presented in the Fifth Conference on Cloud and Internet of Things (CIoT) [1]. By adding other methods such as the Genetic Algorithm, First-Fit, Random-Fit, and Simulated Annealing, our approach will predict the energy consumed by the data centers and the workload placement (containers) in the servers, offering best and optimal solutions to reduce energy consumption, waste of cloud resources, and have an energy-efficient container placement policy.

Generally, Cloud Computing is based on a large data center (server farm), in which many servers are connected to achieve high performance. The data centers represent an infrastructure of several instances (hosts, virtual machines...) [2]–[4]. Each of these instances requires an allocation at the data center because of the growing demand for hosting services. For these reasons, they are considered to be heavy consumers of resources and energy [3]. In data center, the energy consumed by active servers represents a large proportion of the total energy [4], [5]. More clearly, the energy consumed by the hosts or hosting servers plus network and storage equipment represents about 40% of the total energy [6]–[8]. Cooling equipment uses between 45% and 50% of the total energy, and the rest is shared among other systems such as lighting [6]–[11]. With the energy of the cooling systems, the costs in the data centers are experiencing a big explosion, which require a reduction in their expenses [12]. According to [5][9][13], the main challenges in data centers are to minimize the heat and energy consumed by cloud infrastructures and to secure them against threats.

So, to optimize the use of energy in data centers, it is necessary to define the servers that must be active according to the current workload and to avoid traditional techniques that negatively influence the quality of services (QoS) such as stopping components or reducing their performance [6], [9], [10]. In most data centers, the consumption of hardware and software resources of each active physical machine is between 11% and 50% with power consumption between 50% and 70% compared to a server whose resources are used entirely by the hosting of the instances and applications executed on these instances [2], [9], [14]. For this, the efficient placement of virtual instances is very important to control the use of material resources and prohibit any kind of their waste that can lead to an increase in energy consumption.

Our approach will have two objectives. The first will be the prediction of workload placement using the Genetic Algorithm and the First Fit Decreasing to define a better placement of a new type of virtual instance called containers. This placement will be constrained to define the best lower number of servers to host container

*Corresponding Author: Amine Bouaouda, +212604129135, amine.bouaouda@edu.uiz.ac.ma

instances. The second objective is the main one, which will be the prediction of energy consumed by a data center, by applying our container placement algorithms to have the best solutions in terms of reducing the energy consumed by a data center and optimizing the use of hardware resources such as RAM, Storage, Bandwidth, etc.

The remainder of this paper is coordinated as follows. We present the related works in the second section. In the third section, we propose our methodology. We perform our algorithms in the fourth section, and the paper is concluded in section five.

2 Related Work

The estimation or prediction of energy consumption in data centers has become a necessary operation in recent years. Due to the huge growth in user demands for cloud services, the number of servers has started to increase to provide the necessary hardware resources. This implied an increase in energy consumption, which forced large cloud companies to do studies on the energy consumed by their data centers, to optimize it in the future, or replace their power source with a cheaper and guaranteed one.

In this context, several approaches estimate energy consumption in data centers. Most of these methods focus on servers to estimate energy. Each server or host is a set of resources such as RAM, HDD, and CPU. In this case, the energy of the host is relative to the sum of the power of all its resources, or according to some [15][16]. Mathematically, the energy consumed by a host is represented in the literature by linear functions that depend on a resource like CPU [15][17], or non-linear, whose functions are quadratic of the CPU resource use [17][18]. Heuristics and metaheuristics are methods that are widely used to estimate the energy in data centers based on the placement of virtual instances in servers especially metaheuristics as they are adaptive for complex problems that require considerable calculation.

In [19], they proposed the placement of virtual machines based on energy consumption by the resources of the data center. One of the objectives of this approach was to reduce energy consumption using a simulated annealing algorithm. This technique generates an initial solution called initial configuration which contains the placement of the VMs in servers, on which they applied at each iteration one of the three simulated annealing techniques: inversion, translation, and switching, to get the next configuration. To define which solution is better, they calculate the energy consumed by the data center in both cases and choose the one that gives the small energy value that will be the best solution. The disadvantages of this approach are the execution time which is very long for large instances, and its process ends if it finds that a new configuration is better than the previous one even if they remain several iterations. This decision does not necessarily indicate that the new configuration is the best among all other solutions.

The same thing in [20], they used the simulated annealing algorithm to propose an economic placement in terms of energy for virtual machines. The proposed algorithm goes through the four stages of the simulated annealing (generation of the initial configuration, obtaining the next generation, definition of the objective function, and timing of temperatures and evolution time). The simulated annealing proposed was compared by the First Fit Decreasing multi-start random searching approach. The results obtained show the effectiveness of the proposed method, but in the case of large instances, it takes a lot of time to find the best solution. For this, they used a time limiter to stop the calculation process in a solution very close to the best.

Other researchers have proposed multi-objective metaheuristics for the placement of workloads in servers by calculating the energy consumed by data centers, to select the best placement in terms of energy minimization. In [21], the authors used the multi-objective genetic algorithm to provide virtual machine placement by minimizing energy consumption and improving the quality of services and the use of resources. The contradictory nature of the objectives defined in this approach has necessarily influenced the distribution of the load between the resources in a data center.

The same thing in [22] and [23], which proposed an approach based on the ant colony algorithm to calculate and minimize the energy consumed by a cloud system based on the placement of virtual machines. In [23], they built a mechanism for measuring indirect energy consumption for virtual machines based on a model for calculating the energy consumed by these machines deployed in [24], because it is difficult to deduce the energy consumption directly from the material because of the existence of the virtualization technology.

Our approach will be different from the old works because we will use cloud containers instead of virtual machines. More of this, we will propose a genetic algorithm and a First-Fit-Decreasing algorithm to provide best container placement without wasting the resources of active servers, with the prediction of energy consumed by any data center to define which algorithm offers best and optimal solutions for the minimization of energy consumed and the optimization of material resources.

3 Methodology

Our main objective is to predict the energy consumed in a data center for a given workload using the genetic algorithm and First-Fit-Decreasing to predict the best or optimal placement of this load (cloud containers). The best algorithm will be one that offers optimal solutions in terms of minimizing energy consumption and optimizing the use of the hardware resources of a data center operating in a specific context.

3.1 Workload Placement

The placement of workloads is an operation that has a great impact on several problems in cloud computing such as minimizing energy consumption. The concept of this placement is to define the servers that will be active to host several virtual instances. To have an optimal placement, it is necessary to choose an optimal number of servers without wasting the material resources of the data center. The workloads in our approach will be the containers.

Containers are small virtual instances in terms of hardware and software resources [25]–[27]. They provide a virtual platform such as Docker, with which multiple users can drive and run their applications or images of operating systems directly on the physical

machine [27], [26].

The containers are efficient compared to the virtual machines because they are lightweight and their installation takes a few seconds and they run directly on the operating system and the hardware of the physical machine. On the other hand, virtual machines take a long time to install and require hypervisors to run [26][25]. In [26], the author represents a containerization technology for creating container instances. It allows users to deploy and run their applications in process containers.

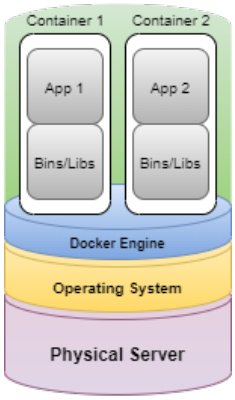


Figure 1: Operation of a container

Our approach will propose two algorithms to predict the best and optimal placement for containers. The main objective will be the choice of an optimal number of servers sufficient to host the containers according to their material resources. This objective is defined by the following equation:

$$F(X) = \sum_{Host=1}^{n} X_{Host} \tag{1}$$

$$X_{Host} = \begin{cases} 1, & \text{if the host is active} \\ 0, & \text{otherwise} \end{cases}$$

To achieve this goal, we have identified three necessary constraints. First, each container must have a placement at the end. The second constraint insists that each container must be placed in a single host. The latter checks that all containers must not exceed the host regarding material resources. For example, for RAM, we have decided to reserve 80% of each active host for container placement and keep the rest for user processing, ensuring proper load balancing between servers.

3.1.1 Genetic Algorithm

The genetic algorithm [28] is an optimization method (metaheuristic) [29] first presented by John Holland in the 1970s. It is based on techniques derived from genetics and the evolutionary mechanisms of nature: crossover, mutation, and selection.

It represents a method for solving complex optimization problems, with or without constraints, based on a natural selection process (a process similar to that of biological evolution). More precisely, it provides solutions to problems that don't have calculable solutions in a reasonable time analytically[30][28].

The process of the genetic algorithm begins with the random creation of thousands of more or less good solutions, then they will be subjected to an evaluation to select the most suitable ones according to constraints. The population continues to evolve through the creation of other generations, by crossing the best solutions between them and having them mutate, then they are brought together with the best already chosen in the selection. This process will be restarted in a certain number of iterations to arrive at the best solutions.

Generation of the initial population: The genetic algorithm in its nature is a population-based method, i.e., it begins with a set of initial solutions named the initial population, and in each iteration, it produces a new generation of solutions of the same size as the initial one. To have an initial population, it must be generated from a set of solutions, which are called individuals (chromosomes). The total number of individuals generated represents the size of the population.

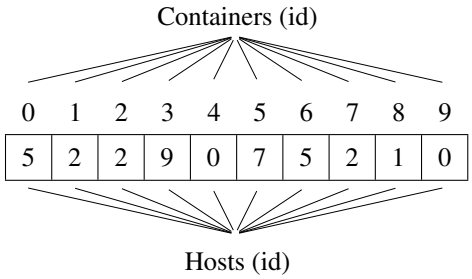


Figure 2: Representation of a chromosome (individual)

So to generate the initial population, you have to create a set of chromosomes as in the figure 3. This generation will be random within the constraints of the approach to ensure that each respects all the constraints of the problem to have a correct initial population.

	0	1	2	3	4	5	6	7	8	9
Chromosome 0	5	2	2	9	0	7	5	2	1	0
Chromosome 1	9	5	0	1	6	2	5	2	1	8
Chromosome 2	2	2	4	7	3	7	5	2	1	0
Chromosome 3	8	0	5	4	4	3	0	2	1	0

Figure 3: Representation of the structure of an initial population

Selection: For our approach, we decided to set the number of chromosomes in four for the initial population and the one to be built in each iteration and select the two best individuals (parents) among each created population, on which the crossing operator will apply to have two new individuals (sons) to build a new generation with new features. The selection of the best individuals is based on a criterion called the Fitness function. It is an equation defined

according to the problem studied, to determine the best solutions that satisfy this function.

To determine the best individuals, we have based on the principle of our objective which serves to minimize the number of active hosts. More clearly, if a chromosome of a population respects all the constraints of the problem and contains the minimum possible hosts, then it represents a good solution among those in the studied population. This choice makes it possible to calculate the number of active hosts in an individual and choose two individuals which contain the minimum number of hosts, or the reciprocal of the number of active hosts, and then choose the two large values obtained. We represent below the equation of the Fitness function to select the best individuals.

$$f(I) = 1/N_{AH} \tag{2}$$

- I : The individual being studied.
- $f(I)$: The Fitness function of individual I.
- N_{AH} : Number of active hosts in the individual I.

The Fitness function will be applied at the beginning of the initial population to determine the individual parents, who will be the inputs of the next phase (the crossover). This function will be applied to each new generation created until the last iteration. These kinds of individuals, who represent the parents are feasible (workable) solutions for the problem studied. Workable individuals are solutions that met the needs of a problem.

So, the individual (parents) selected according to the Fitness function in each iteration are feasible solutions among others to discover probably in the following iterations. But instead of choosing both parents, we decided to compare them by choosing the best among them (the one that has the greatest value of Fitness, if they have equal values, then both will be chosen as feasible solutions). And each feasible solution represents one of the best-suggested ways to place containers in a minimum number of hotels without any waste of resources or energy consumed.

Crossover: The choice of the best individuals in the selection phase is the starting point of the crossover. These selected individuals are genetically better according to the function of Fitness, and they contain characteristics that will improve each population by producing new individuals called sons. The creation of new individuals is done by a crossover applied to the parents explained by the following two points:

- Divide each parent individual into two parts (from the middle), if the size of the individual is even, then both parts will have the same length, otherwise, the first part will exceed the size of the second with an element.
- Concatenate the first part of the first parent with the second of the other to have the first son, then the second part of the first parent with the first of the other parent to have the second son.

After creating the new individuals (sons), the parent individuals are kept to build a new generation of four chromosomes as the initial population after the application of the mutation operation on the sons. Keeping parents is not optional, but it has two very important roles in the algorithm process.

First, it avoids the case of having a bad generation, because parents are already good solutions according to the Fitness function applied in the selection, on the other hand, there is no information about the nature of new individuals created, whether they are good solutions to the problem or not, until the application of the mutation phase and build a new population and apply the selection operator to it to indicate the individuals who will be the new parents who may be one of the created sons of the previous iteration or both or none.

Secondly, the conservation of parents and combining them with the new individuals makes it possible to respect the rule of construction of a population in the genetic algorithm which indicates that all generations must be of the same size as the one at the beginning, as well as this combination guarantees diversity in the best solutions and increase the proportion of having several that are feasible.

Mutation: Generally, the mutation principle is used to apply more change to the sons created in the crossover phase to obtain a rich and different population from the previous one. These changes will apply to some genes of one or more chromosomes. In our case, the genes on each chromosome are the identifiers of the hosts chosen to host the containers, as shown in the figure 4.

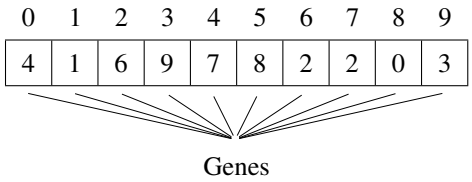


Figure 4: Individual-level active hosts represent their genes

So the change in the hosts of an individual involves a modification in the location of the containers. For this reason, the mutation phase for our approach will be slightly different compared to those of other problems that use the genetic algorithm. More precisely, The main role of the mutation will be corrective, that is to correct the faults produced by the crossing at the level of the son individuals. This crossing can cause false locations for containers at these individuals. For this reason, we must first check the location of the containers in each child's individual if one or more exceed the host that hosts them in terms of RAM. In this way, the two sons will be corrected by ensuring that they respect the constraints of the approach before combining them with the parents to build a new generation.

This "corrective" mutation has another impact because the correction made at the level of the sons will have a great chance of bringing back new characteristics. After all, the containers in those chromosomes have probably different locations to those in the individual parents, and therefore the possibility of having other better solutions. With all these characteristics, the mutation phase will have a great impact on the realization of a good dynamic container placement which is one of the big goals of this optimization problem, meaning the appearance of the migration notion of virtual instances

between active hosts, which is absent in the static placement which prevents the migration of a container from one host to another after having placed it.

Algorithm 1: Container Placement based on Genetic Algorithm

Data: List of container requests and list of hosts
Result: Deploy containers on hosts
$HL \leftarrow HostsList$;
$CL \leftarrow ContainersList$;
for $i = 1$ *to* 4 **do**
  $list_{InitialSolutions} \leftarrow createChromosome(HL, CL)$;
  $map_{InitialPopulation}.add(i, list_{InitialSolutions})$;
  $list_{InitialSolutions} \leftarrow [\,]$;
end
for $i = 1$ *to max* $= 100$ **do**
  $map_{bests} \leftarrow Selection(map_{InitialPopulation})$;
  $map_{feasibles} \leftarrow Workable(map_{bests}, HL, CL)$;
  for $key : map_{feasibles}.keySet()$ **do**
    $list_{feasibleSolutions}.add(map_{feasibles}.get(key))$;
  end
  $map_{crossover} \leftarrow Crossover(map_{bests})$;
  $map_{mutation} \leftarrow Mutation(map_{crossover}, HL, CL)$;
  $map_{NG} \leftarrow NewGeneration(map_{bests}, map_{mutation})$;
  $map_{InitialPopulation} \leftarrow map_{NG}$;
end
$list_{res} \leftarrow optimize(list_{feasibleSolutions}, HL, CL)$;
Deploy list of feasible solutions ($list_{res}$) that represent the best container placements in the host;

Algorithm 1 shows the principle of the genetic algorithm and the main methods used to achieve the objective of the problem. Input data is set before starting initialization. This entry will help the different operations related to the algorithm to determine the possible feasible solutions that will be the result, which represent several best container placements.

The algorithm starts by checking the selected input data, which are the hosts and the containers to be placed. This operation checks whether the total sum of Rams of all outbound hosts is greater than that of containers, to ensure that there is sufficient space for the placement of the container request. Then, if the verification is done and the input data is accepted, then the construction of the initial population begins by creating four chromosomes. Each individual built (as a list) will be added to the Map which represents the initial population until the end of the operation by obtaining an initial generation of four individuals.

The selection phase will be applied to the initial population obtained to determine the best individuals who will be the parent chromosomes. The best individuals (parents) selected will be the entry of the method responsible on the crossing to produce new individuals (the sons), the result of this method will be after the entry of mutation.

In the end, a new generation will be created, after the end of the mutation phase, which will be the population of the second iteration. The final list of feasible solutions will go through the optimization phase to further verify the placement of the containers and improve it if possible.

3.1.2 First-Fit Decreasing

First-Fit Decreasing is one of the best-known algorithms for the classic problem of Bin Packing [31]. The FFD's strategy for placement is defined by three points. At first, we set the elements in descending order of size. Secondly, we put each item we get there in the oldest bin (opened the earliest) into which it fits (whenever an item fits the capacity of bin 1, put it there, otherwise, it fits into bag 2, if it fits). Thirdly, the opening of a new bag or bin is only done if the item does not fit into a bag that already contains something [30, 32]. For our problem, the items to put in the bin will be the containers, while the bags or bins will be the hosts.

The following algorithm shows the process of placing containers in servers with the FFD.

Algorithm 2: FFD-based container placement

Data: List of container requests and list of hosts
Result: Deploy containers on hosts
Set the container list in decreasing order of RAM;
$HS \leftarrow$ The size of the host list;
$CS \leftarrow$ The size of the container list;
for $i = 0$ *to* $CS - 1$ **do**
  for $j = 0$ *to* $HS - 1$ **do**
    The *Resource* is a vector that contains the values of four resources (RAM, CPU, Storage, Bandwidth);
    if $Resource_{host_j} >= Resource_{container_i}$ *and* $!PrevHostList.contains(i)$ **then**
      $PlacementList.add(j)$;
      $Resource_{host_j} \leftarrow Resource_{host_j} - Resource_{container_i}$;
      $PrevHostList.add(i)$;
    end
  end
end
Deploy the *PlacementList* that contains the placement of the container list in the hosts;

3.2 Forecasting Energy Consumption

To predict the energy consumption in the data centers, we will calculate the energy consumed by the system of active hosts and storage. This energy will help us predict the energy consumed by other equipment such as the cooling system. The following figure shows the distribution of energy consumed by a data center based on studies and previous work.

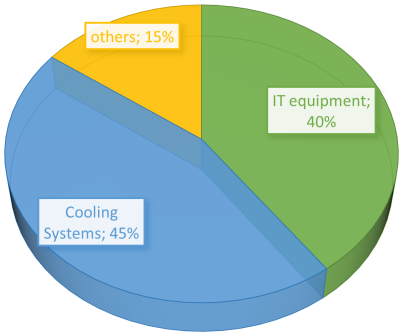


Figure 5: The distribution of energy consumption of a data center

To predict the energy consumed by active hosts, we rely on the placement of containers in these servers. This workload placement will help us determine the percentage used for hardware resources (RAM, CPU, Storage, Bandwidth) of each active host. The calculated percentages will be sorted between 0 (0%) and 1 (100%) to simplify their use in the CloudSim power model. Based on these percentages, we will define for each host the energy consumed by each resource in watts. The following equation represents the energy consumed in watts by a resource according to its percentage of use for a host.

$$Power_{resource_i} = getPower(Use_{resource_i}) \tag{3}$$

Predicting the energy consumed by each resource will help us to define the energy consumption for a single host and for all active servers that are defined by the following two equations.

$$Power_{host} = \sum_{i=1}^{4} getPower(Use_i), Use_i \in [0, 1] \tag{4}$$

The Use_1 is the used value of the RAM, Use_2 is the used value of CPU, Use_3 is the used value of the BW, and Use_4 is the used value of the Storage.

$$Power_{AllHosts} = \sum_{i=1}^{N} Power_{host} \tag{5}$$

We represent below our algorithm for predicting the energy consumed by a data center based on container placement.

Algorithm 3: Forecasting energy consumption

Data: List of container requests and list of hosts
Result: Predict the energy consumed by the data center
Apply the Genetic Algorithm or FFD to predict container placement;
Group the percentages of use of each resource for each server in a map named $UseMap$;
for $key : UseMap.keySet()$ **do**
  for $i = 0$ *to* $UseMap.get(key).size()$ **do**
    Apply the Equation (3) to calculate the energy consumed by each resource (RAM, CPU, Storage, Bandwidth); Apply Equation (4) to calculate the total energy consumed by a single host;
  end
  Apply Equation (5) to predict the energy consumed by all active servers;
end
$DataCenterPower(Power_{AllHosts})$;

So, to predict the total energy consumed by a data center, we will rely on the energy predicted for the active host system. Algorithm (3) predicts the energy consumed by the active hosts in the data center based on the placement of workloads. Based on Figure 5 and previous work, we noted that the energy consumption of a data center is divided into three classes. The cooling system takes a large proportion with a value between 45% and 50% (47% on average). For the active host system and storage equipment, their consumption is between 36% and 40% (38% on average). Other systems such as lighting and communication equipment consume 15% of the total energy.

4 Experiments and Results

To evaluate our approach, we used CloudSim 4.0 to apply our algorithms to homogeneous and heterogeneous cloud systems. Our experiments will present the prediction of container placement, then the energy consumption for each system. In this section, we will perform three different experiments. In the first two applications, we will have two scenarios.

4.1 Application 1 : Homogeneous System

The homogeneous system we used in this experiment consists of 50 identical servers in terms of material resources and 3000 containers of different classes.

Table 1: Details of the Application 1

System	Number	RAM (MB)	Number of CPUs	Bandwidth	Storage (MB)
Hosts	50	65536	512	2000000	2000000
Containers (class 1)	1000	512	1	2500	1024
Containers (class 2)	1000	256	1	2500	1024
Containers (class 3)	1000	128	1	2500	1024

4.1.1 SCENARIO 1 : When all hosts are active and the percentage of use of each resource = 50%

In this scenario, we decided to predict the energy consumed by this data center in the case where the use of each material resource equals 50%. The following table shows the results obtained in detail.

Table 2: Energy consumption for the scenario 1 of Application 1

RAM use (watts)	CPU use (watts)	Bandwidth use (watts)	Storage use (watts)
5800	5800	5800	5800
Number of active hosts : 50			
Energy consumption of active hosts (38%)	Energy consumption of cooling system (47%)	Energy consumption of others (15%)	Energy consumption of the data center
23200.0 watts	28694.736 watts	9157.895 watts	61052.63 watts

4.1.2 SCENARIO 2 : Genetic Algorithm application

This scenario represents the application of our genetic algorithm for the prediction of container placement and the energy consumption of the different systems of the data center. The results obtained are as follows.

Table 3: Energy consumption for the scenario 2 of Application 1

RAM use (watts)	CPU use (watts)	Bandwidth use (watts)	Storage use (watts)
2288.932	1920.636	1828.676	1738.72
Number of active hosts : 18			
Energy consumption of active hosts (38%)	Energy consumption of cooling system (47%)	Energy consumption of others (15%)	Energy consumption of the data center
7776.964 watts	9618.877 watts	3069.8542 watts	20465.693 watts

4.1.3 SCENARIO 3 : FFD application

In this scenario, we applied the First-Fit Decreasing algorithm to place the container list and predict the energy consumed by this data center after completing the workload placement. The following table shows the results obtained.

Table 4: Energy consumption for the scenario 3 of Application 1

RAM use (watts)	CPU use (watts)	Bandwidth use (watts)	Storage use (watts)
2288.932	1826.91	1830.826	1739.208
Number of active hosts : 18			
Energy consumption of active hosts (38%)	Energy consumption of cooling system (47%)	Energy consumption of others (15%)	Energy consumption of the data center
7782.606 watts	9625.8545 watts	3072.0813 watts	20480.543 watts

The objective of the first scenario is to see the rate of energy that will be consumed if a data center hosts a given workload but without any strategy to define its conception so as not to fall into the problem of waste of resources which implies an expansion of energy consumption. In this scenario, each host consumed 464 watts because they are identical. The fact that the 50 hosts are all active, involved an expansion of the total energy consumed that exceeded 61052 watts, which is ordinary because all servers are active. This scenario shows the importance of defining a strategy for the placement of data center workloads and seeing the impact when a strategy is applied to place containers in an minimal number of servers. The application of the genetic algorithm gave two solutions of 18 servers for each that will be active to host the 3000 containers. We noticed in each solution that the power consumption of the hosts is close to each other (between 400 and 476 watts). As well as workloads are well distributed among active servers (a good load balancing). For total energy consumption, the first solution predicted a value of 20465.693 watts with 40586.937 saved energy compared to the first scenario. The second solution estimates the energy consumption with 20466.234 watts and 40586.396 watts of energy saved when compared with the solution of scenario 1.

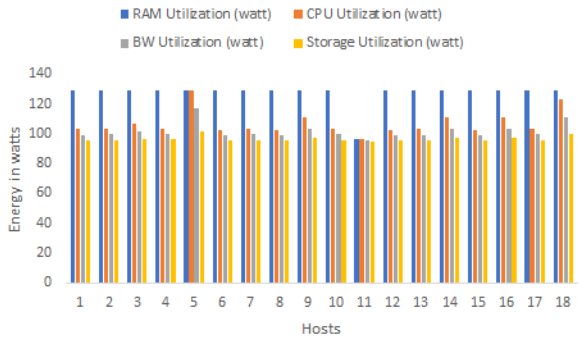


Figure 6: The energy consumed in watts by each resource for the GA first solution - Application 1

Like the genetic algorithm, the FFD proposed a single arrangement of 18 hosts among the starting 50 to place the list of containers. So, 32 will be retained, contrary to the first scenario. Of the 18 active servers, 9 hosts each consumed more than 423 watts, 8 others consumed between 427 and 476 watts for each, and only one server that is the last one consumed less than 400 watts (383.59146 watts). The placement proposed by the FFD also gave good results for the total energy consumption (20480.543 watts) which is lower than in the first scenario with a value of 40572.087 watts of energy saved. But the solutions of the genetic algorithm are better in terms of the energy consumed but identical to that of the FFD in the number of active hosts.

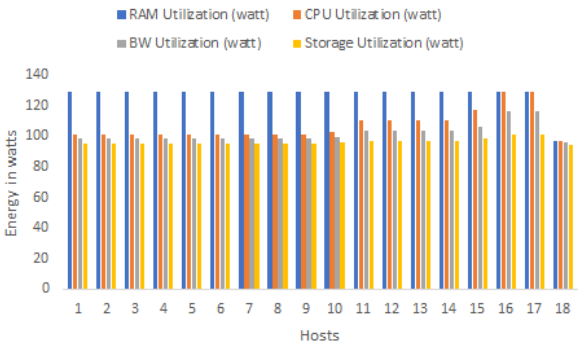


Figure 7: The energy consumed in watts by each resource for the FFD solution - Application 1

4.2 Application 2 : Heterogeneous System

In this application, we used a heterogeneous system that consists of three different types of hosts with a total number of 50, and 3000 containers.

Table 5: Details of the Application 2

System	Number	RAM (MB)	Number of CPUs	Bandwidth	Storage (MB)
Hosts (class 1)	17	32768	256	2000000	2000000
Hosts (class 2)	17	65536	512	2000000	2000000
Hosts (class 3)	16	131072	640	2000000	2000000
Containers (class 1)	1000	512	1	2500	1024
Containers (class 2)	1000	256	1	2500	1024
Containers (class 3)	1000	128	1	2500	1024

4.2.1 SCENARIO 1 : When all hosts are active and the percentage of use of each resource = 50%

In the first scenario of application 2, we predicted the energy consumed for a heterogeneous system when all servers are active.

Table 6: Energy consumption for the scenario 1 of Application 2

RAM use (watts)	CPU use (watts)	Bandwidth use (watts)	Storage use (watts)
6202	6202	6202	6202
Number of active hosts : 50			
Energy consumption of active hosts (38%)	Energy consumption of cooling system (47%)	Energy consumption of others (15%)	Energy consumption of the data center
24808.0 watts	30683.578 watts	9792.632 watts	65284.21 watts

4.2.2 SCENARIO 2 : Genetic Algorithm application

Scenario 2 represents the application of the genetic algorithm on our heterogeneous system. The results are as follows.

Table 7: Energy consumption for the scenario 2 of Application 1

RAM use (watts)	CPU use (watts)	Bandwidth use (watts)	Storage use (watts)
1944.637	1594.332	1450.27	1193.698
Number of active hosts : 11			
Energy consumption of active hosts (38%)	Energy consumption of cooling system (47%)	Energy consumption of others (15%)	Energy consumption of the data center
6182.937 watts	7647.3164 watts	2440.633 watts	16270.887 watts

4.2.3 SCENARIO 3 : FFD application

For this scenario, we applied the FFD algorithm. Below is the prediction of the energy consumed by our system after this application.

Table 8: Energy consumption for the scenario 3 of Application 2

RAM use (watts)	CPU use (watts)	Bandwidth use (watts)	Storage use (watts)
3049.891	2592.923	2448.035	2358.184
Number of active hosts : 26			
Energy consumption of active hosts (38%)	Energy consumption of cooling system (47%)	Energy consumption of others (15%)	Energy consumption of the data center
10449.033 watts	12923.805 watts	4124.6187 watts	27497.457 watts

In the second application, we used a heterogeneous system to see if the placement of the containers and the rate of energy consumption will be influenced by the nature of the system. In the first scenario, the 50 hosts consumed a different energy rate because they are heterogeneous. The first 17 servers consumed 408 watts each, and the last 16 consumed the high value (624 watts each). The other servers consumed 464 watts each. For the total energy consumption, the system consumed 65284.21 watts with superiority of 4231.58 compared to scenario 1 of the first application.

The application of the genetic algorithm for this heterogeneous system gave two different solutions in contrast to scenario 2 of Experiment 1 which proposed two identical solutions at the level of the proposed number of active hosts. The heterogeneous nature of this system has increased the chance of having diversified solutions in terms of the hosts used and their number.

The first solution obtained by the genetic algorithm has proposed 13 servers that will be active to host the container list, which is the best number compared to the first scenario. The energy consumption of the 13 active servers is between 350 and 800 watts because of the hosts' diversity in the hardware resources. This diversity influenced the rate of energy consumed by the data center which did not exceed 17623.83 watts with 47660.38 watts of energy saved compared to scenario 1.

The second solution was the best with 11 active hosts and less than 2 servers compared to the first solution to host workloads. The values consumed of energies differ from one host to another because of the diversity in material resources. This number of active servers influenced the energy consumption rate, which was minimized at the level of all the data center systems and 1352.943 watts for the total consumption.

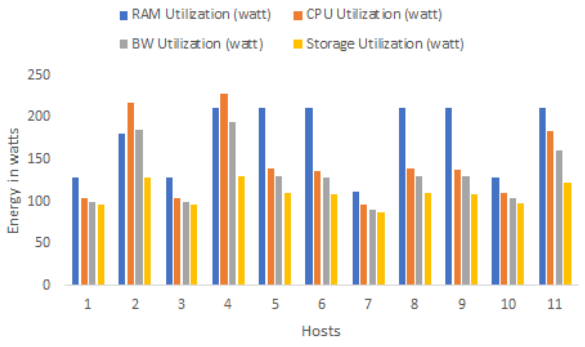


Figure 8: The energy consumed in watts by each resource for the GA second solution - Application 2

The unique solution of the FFD is different from those of the

genetic algorithm. The solution offers 26 servers for container placement which is not optimal when compared with the solutions of scenario 2. This increase in the number of active hosts for the FFD is because of its way of assigning containers to servers which is classic and avoids waste of space but is not always effective if there is a list of containers in a defined order (increasing or decreasing). For this reason, most of the active hosts (17) belong to class 1 and the others belong to the second class.

So workloads were not well distributed between the active hosts about the solution of the genetic algorithm, which is a weak point in the FFD algorithm. The high number of active servers consumed between 370 and 480 watts for each, implied an increase in the rate of energy consumed by the data center with a value of 27497.457 watts.

We note that the genetic algorithm guarantees best container placement and minimal energy consumption due to its functioning which is adaptive to different types of systems. But in general, both algorithms offer a good solution for virtualization in a data center, which is important for the whole system to operate without any waste of resources that can increase energy consumption

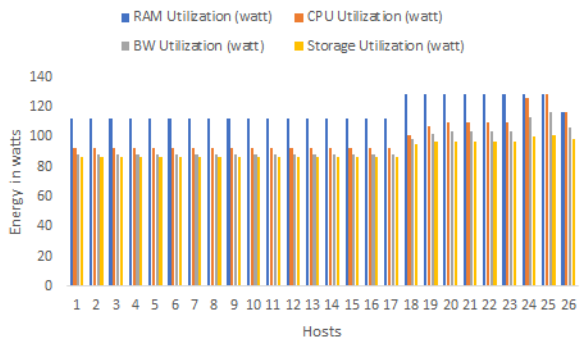


Figure 9: The energy consumed in watts by each resource for the FFD solution - Application 2

4.3 Application 3 : Global Comparison

After testing our algorithms, we decided to compare them with other previous heuristic or meta-heuristic methods to properly evaluate our algorithms. For this, we chose the approach in [19] which uses the Simulated Annealing (SA) algorithm to predict the placement of virtual machines based on the estimate of energy consumption to choose the best placement. In our case, we will evaluate this algorithm using container instances to compare it with our genetic algorithm (GA). More of this we will compare the approach proposed in [33] which applies the Ant Colony Optimization (ACO) for container placement with our genetic algorithm (GA).

The choice of the Simulated Annealing and ACO to examine the effectiveness of our genetic algorithm was not random, but due to several reasons. First of all these three algorithms belong to the class of metaheuristics which is inspired by nature to solve optimization problems. More of that, they can adapt to problems with a high complexity and which require a huge calculation. Another point that encouraged us more to compare our Genetic Algorithm with Simulated Annealing and the ACO is that the works [19] and [33] have the same vision regarding the placement of virtual instances by optimizing the use of material resources and uses the same procedure, except that the work [19] are based on virtual machines instead of containers. For the FFD, we will evaluate it with First-Fit (FF) and Random-Fit (RF) which are algorithms of the Bin Packing problem such as the FFD. Table 9 shows the resources of the different cloud systems used on which we will perform our comparison.

Table 10 represents the results obtained for the energy consumption of several data centers for each algorithm.

Table 10 represents three main results for six algorithms. For each system, we applied six algorithms to predict the energy consumed by each data center based on container placement. As well as, we calculated the energy saved or conserved for each system by comparing the predicted energy consumption with the energy consumed when all the hosts of a system are active and the percentage of use of each resource = 50%. In addition, we calculated the execution time of each algorithm.

For the first system which is homogeneous, we notice that the energy consumption obtained by the genetic algorithm does not exceed 13460 watts with 47594.73 watts of saved energy. For the ACO algorithm, the energy consumption exceeded the value predicted by the genetic algorithm twice with 32931.009 watts of stored energy. Similarly, for the SA algorithm, the energy consumed exceeded that of GA but with a big difference which influenced the rate of energy saved which was less than 10380 watts. On the other hand, the execution time of the genetic algorithm was very high (more than 31 seconds) compared to that of ACO which gave its result in 0.297 seconds and the SA in 5.62 seconds. The reason why the execution time of the genetic algorithm is very high is because of its execution process that operates under several iterations (in our case 100 iterations).

For heuristics, the three Bin Packing problem-solving algorithms (FFD, FF, and RF) predicted energy consumption values close to each other. The values proposed by the FFD and FF were identical to the superiority of the RF algorithm which gave a value of 13447.468 watts which is optimal compared to those of the FFD and FF. For the energy saved, the three heuristics provided better values with the genetic algorithm, but they are fast in execution time because they produce a single solution without using several iterations.

In the second system, which is heterogeneous, the results obtained by the genetic algorithm for the energy consumed and saved are the best when compared with those of ACO and SA and also with the three heuristics. On the other hand, its execution time which reached 31 seconds is very high compared to the other algorithms. We also notice that the FFD, FF, and RF gave the same energy consumption value (31874.584 watts), but the RF surpasses them in the execution time (0.08 seconds). The same goes for the third system, which has a superiority of the genetic algorithm over the others in terms of the energy consumed which is optimal, but the execution time has increased greatly (301.769 seconds) because of the size of the number of instances in this system. The SA also took 103.123 seconds which is high compared to other methods.

For other systems, our genetic algorithm was the best in systems 5 and 6 in terms of optimal energy consumption, but it takes a long time to finish its execution and give the final results. In system 4, the FFD and FF heuristics proposed optimal solutions with an energy

Table 9: Details of the Application 3

	Hosts					Containers				
System	Number	RAM (GB)	Number of CPUs	Bandwidth	Storage (MB)	Number	RAM (MB)	Number of CPUs	Bandwidth	Storage (MB)
1	50	32	256	2000000	2000000	334	512	1	1024	2500
						334	256	1	1024	2500
						332	128	1	1024	2500
2	17	16	128	2000000	2000000	1000	512	1	1024	2500
	17	32	256	2000000	2000000					
	16	64	512	2000000	2000000					
3	25	16	128	2000000	2000000	667	128	1	1024	2500
	25	32	256	2000000	2000000	667	256	1	1024	2500
	25	64	512	2000000	2000000	666	512	1	1024	2500
4	100	64	512	2000000	2000000	1667	512	1	1024	2500
						1667	256	1	1024	2500
						1666	128	1	1024	2500
5	34	32	256	2000000	2000000	1667	128	1	1024	2500
	34	64	512	2000000	2000000	1667	256	1	1024	2500
	32	128	1024	2000000	2000000	1666	512	1	1024	2500
6	30	256	1024	4000000	4000000	3500	256	1	1024	2500
	30	128	1024	4000000	4000000					
	30	64	512	4000000	4000000	3500	512	1	1024	2500
	30	32	256	4000000	4000000					

Table 10: Comparison of different algorithms for predicting the energy consumption of different cloud systems

	GA			ACO			SA		
System	Data Center Energy (watt)	Energy Saved (watt)	Execution time (second)	Data Center Energy (watt)	Energy Saved (watt)	Execution time (second)	Data Center Energy (watt)	Energy Saved (watt)	Execution time (second)
1	13457.9	47594.73	31.954	28121.621	32931.009	0.297	50678.336	10374.294	5.62
2	18792.02	42260.61	31.004	39339.375	21713.255	0.179	51927.41	9125.22	4.034
3	20413.207	71165.743	301.769	53369.266	38209.684	0.264	77419.08	14159.87	103.123
4	43210.434	121000.086	5304.5	67590.805	96619.715	1.496	100156.78	64053.74	3870.694
5	28867.098	101701.322	3886.164	77247.85	53320.57	1.17	98288.49	32279..93	1165.504
6	19218.874	148149.546	11634.408	31827.826	135540.594	1.126	116079.95	51288.47	1317.876

	FFD			FF			RF		
System	Data Center Energy (watt)	Energy Saved (watt)	Execution time (second)	Data Center Energy (watt)	Energy Saved (watt)	Execution time (second)	Data Center Energy (watt)	Energy Saved (watt)	Execution time (second)
1	13462.325	47590.305	0.194	13462.325	47590.305	0.122	13447.468	47605.162	0.118
2	31874.584	29178.046	0.116	31874.584	29178.046	0.062	31874.584	29178.046	0.08
3	40164.184	51414.766	0.351	40546.266	51032.684	0.2	40333.258	51245.692	0.287
4	43202.652	121007.868	7.788	43202.652	121007.868	8.441	43329.03	120881.49	2.856
5	48174.152	82394.268	2.93	48301.848	82266.572	2.594	48233.105	82335.315	2.186
6	21229.072	167368.42	12.186	21242.38	146126.04	8.91	21098.488	146269.932	10.894

value of 43202.652 watts which is less than 7782 watts compared to the value proposed by the genetic algorithm. Generally, our GA has proposed best values for energy consumption for most systems ahead of other metaheuristics (ACO and SA) and even heuristics. On the other hand, execution time remains its main weakness because of its way of solving the problem. FFD was best with RF in the heuristics used. In terms of execution time, the ACO was the best with an average time of 0.75 seconds. More of this the SA algorithm proposed large energy values and ranked second before the genetic algorithm at the level of execution time.

5 Conclusion

In this paper, we have presented an approach based on heuristics and metaheuristics for predicting the energy consumed by different data centers based on the placement of workloads. Our approach has proposed a genetic algorithm and a First-Fit Decreasing using

cloud containers to predict the best and optimal placement for these instances in several servers without wasting hardware resources. The results obtained showed that the genetic algorithm guarantees a good placement for containers and minimizes the energy consumption in the data centers, after comparing it with other metaheuristics such as Ant Colony Optimization and Simulated Annealing.

Conflict of Interest The authors declare no conflict of interest.

References

[1] A. Bouaouda, K. Afdel, R. Abounacer, "Forecasting the Energy Consumption of Cloud Data Centers Based on Container Placement with Ant Colony Optimization and Bin Packing," in 2022 5th Conference on Cloud and Internet of Things (CIoT), 150–157, 2022, doi:10.1109/CIoT53061.2022.9766522.

[2] A. Greenberg, J. Hamilton, D. A. Maltz, P. Patel, "The Cost of a Cloud: Research Problems in Data Center Networks," SIGCOMM Comput. Commun. Rev., **39**(1), 68–73, 2009, doi:10.1145/1496091.1496103.

[3] G. Wu, M. Tang, Y.-C. Tian, W. Li, "Energy-Efficient Virtual Machine Placement in Data Centers by Genetic Algorithm," in T. Huang, Z. Zeng, C. Li, C. S. Leung, editors, Neural Information Processing, 315–323, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[4] A. Mosa, N. Paton, "Optimizing virtual machine placement for energy and SLA in clouds using utility functions," Journal of Cloud Computing, **5**, 2016, doi:10.1186/s13677-016-0067-7.

[5] "Cloud Computing Energy Efficiency, Strategic and Tactical Assessment of Energy Savings and Carbon Emissions Reduction Opportunities for Data Centers Utilizing SaaS, IaaS, and PaaS," Technical report, Pike Research, 2010.

[6] M. Dayarathna, Y. Wen, R. Fan, "Data Center Energy Consumption Modeling: A Survey," IEEE Communications Surveys Tutorials, **18**(1), 732–794, 2016, doi:10.1109/COMST.2015.2481183.

[7] "Energy efficiency policy options for australian and new zealand data centres," Technical report, The Equipment Energy Efficiency (E3) Program, 2014.

[8] R. Brown, E. Masanet, B. Nordman, B. Tschudi, A. Shehabi, J. Stanley, J. Koomey, D. Sartor, P. Chan, J. Loper, S. Capana, B. Hedman, "Report to Congress on Server and Data Center Energy Efficiency: Public Law 109-431," 2007, doi:10.2172/929723.

[9] D. Boru, D. Kliazovich, F. Granelli, P. Bouvry, A. Y. Zomaya, "Energy-efficient data replication in cloud computing datacenters," in 2013 IEEE Globecom Workshops (GC Wkshps), 446–451, 2013, doi:10.1109/GLOCOMW.2013.6825028.

[10] W. Van Heddeghem, S. Lambert, B. Lannoo, D. Colle, M. Pickavet, P. Demeester, "Trends in worldwide ICT electricity consumption from 2007 to 2012," Computer Communications, **50**, 64–76, 2014, doi:https://doi.org/10.1016/j.comcom.2014.02.008, green Networking.

[11] "Facts & stats: Data architecture and more data," Technical report, Info-Tech, 2010.

[12] Greenpeace, "Make It Green: Cloud Computing and its Contribution to Climate Change," Technical report, Greenpeace International, 2010.

[13] A. Sharma, N. Nitin, "A Multi-Objective Genetic Algorithm for Virtual Machine Placement in Cloud Computing," International Journal of Innovative Technology and Exploring Engineering (IJITEE), **8**, 2278–3075, 2019.

[14] G. Dasgupta, A. Sharma, A. Verma, A. Neogi, R. Kothari, "Workload Management for Power Efficiency in Virtualized Data Centers," Commun. ACM, **54**(7), 131–141, 2011, doi:10.1145/1965724.1965752.

[15] G. Warkozek, E. Drayer, V. Debusschere, S. Bacha, "A new approach to model energy consumption of servers in data centers," in 2012 IEEE International Conference on Industrial Technology, 211–216, 2012, doi:10.1109/ICIT.2012.6209940.

[16] L. Liu, H. Wang, X. Liu, X. Jin, W. He, Q. Wang, Y. Chen, "GreenCloud: a new architecture for green data center," 2009, doi:10.1145/1555312.1555319.

[17] X. Fan, W.-D. Weber, L. A. Barroso, "Power Provisioning for a Warehouse-Sized Computer," SIGARCH Comput. Archit. News, **35**(2), 13–23, 2007, doi:10.1145/1273440.1250665.

[18] D. Meisner, T. F. Wenisch, "Peak power modeling for data center servers with switched-mode power supplies," in 2010 ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED), 319–324, 2010, doi:10.1145/1840845.1840911.

[19] K. Dubey, S. C. Sharma, A. A. Nasr, "A Simulated Annealing based Energy-Efficient VM Placement Policy in Cloud Computing," in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 1–5, 2020, doi:10.1109/ic-ETITE47903.2020.119.

[20] Y. Wu, M. Tang, W. Fraser, "A simulated annealing algorithm for energy efficient virtual machine placement," in 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 1245–1250, 2012, doi:10.1109/ICSMC.2012.6377903.

[21] J. Xu, J. A. B. Fortes, "Multi-Objective Virtual Machine Placement in Virtualized Data Center Environments," in 2010 IEEE/ACM Int'l Conference on Green Computing and Communications and Int'l Conference on Cyber, Physical and Social Computing, 179–188, 2010, doi:10.1109/GreenCom-CPSCom.2010.137.

[22] Y. Gao, H. Guan, Z. Qi, Y. Hou, L. Liu, "A multi-objective ant colony system algorithm for virtual machine placement in cloud computing," Journal of Computer and System Sciences, **79**(8), 1230–1242, 2013, doi:https://doi.org/10.1016/j.jcss.2013.02.004.

[23] S. Pang, W. Zhang, M. Tongmao, Q. Gao, "Ant Colony Optimization Algorithm to Dynamic Energy Management in Cloud Data Center," Mathematical Problems in Engineering, **2017**, 1–10, 2017.

[24] A. Kansal, F. Zhao, J. Liu, N. Kothari, A. A. Bhattacharya, "Virtual Machine Power Metering and Provisioning," in Proceedings of the 1st ACM Symposium on Cloud Computing, SoCC '10, 39–50, Association for Computing Machinery, New York, NY, USA, 2010, doi:10.1145/1807128.1807136.

[25] P. Dziurzanski, S. Zhao, M. Przewozniczek, M. Komarnicki, L. S. Indrusiak, "Scalable distributed evolutionary algorithm orchestration using Docker containers," Journal of Computational Science, **40**, 101069, 2020, doi:https://doi.org/10.1016/j.jocs.2019.101069.

[26] C. Zaher, "For CTO's: the no-nonsense way to accelerate your business with containers," Technical report, Canonical Limited 2017. Ubuntu, Kubuntu, 2017.

[27] D. Bernstein, "Containers and Cloud: From LXC to Docker to Kubernetes," IEEE Cloud Computing, **1**(3), 81–84, 2014, doi:10.1109/MCC.2014.51.

[28] O. Kramer, Genetic Algorithms, 11–19, Springer International Publishing, Cham, 2017, doi:10.1007/978-3-319-52156-5_2.

[29] I. Boussaïd, J. Lepagnot, P. Siarry, "A survey on optimization metaheuristics," Information Sciences, **237**, 82–117, 2013, doi:https://doi.org/10.1016/j.ins.2013.02.041, prediction, Control and Diagnosis using Advanced Neural Computations.

[30] N. Janani, R. Jegan, P. Prakash, "Optimization of Virtual Machine Placement in Cloud Environment Using Genetic Algorithm," Research Journal of Applied Sciences, Engineering and Technology, **10**, 274–287, 2015, doi:10.19026/rjaset.10.2488.

[31] A. Wolke, B. Tsend-Ayush, C. Pfeiffer, M. Bichler, "More than bin packing: Dynamic resource allocation strategies in cloud data centers," Information Systems, **52**, 83–95, 2015, doi:https://doi.org/10.1016/j.is.2015.03.003, special Issue on Selected Papers from SISAP 2013.

[32] R. Panigrahy, K. Talwar, L. Uyeda, U. Wieder, "Heuristics for Vector Bin Packing," 2011.

[33] O. SMIMITE, K. AFDEL, "Hybrid Solution for Container Placement and Load Balancing based on ACO and Bin Packing," International Journal of Advanced Computer Science and Applications, **11**(11), 2020, doi:10.14569/IJACSA.2020.0111174.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 1-12 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

The Perceptions of Students and Teachers When using ICTs for Educational Practices Matter: *A Systematic Review*

Angela Pearce*

College of Doctoral Studies, Grand Canyon University, Phoenix, AZ, City, 85061, United States

ARTICLE INFO

Article history:
Received: 01 August, 2022
Accepted: 18 October, 2022
Online: 13 November, 2022

Keywords:
Information and Communication Technologies
ICTs
Educational Practices
Technology Acceptance Model
TAM2

ABSTRACT

Before succumbing to the 2019 Coronavirus pandemic, information and communication technologies (ICTs) have sustained a ubiquitous presence in human lives and society. ICTs have changed the standards and dynamics of educational practices (EPs). Many academic institutions had already integrated technological-based pedagogical instructions into their educational practices but, in various cases, faced challenges of failing to consider the perceptions of chief users, students, teachers, and subjective norms. This paper is an extension of work originally presented at the 2022 11th International Conference on Education and Information Technology (ICEIT). This paper aims to demonstrate and provide future directions regarding the effects of ICTs and how such usage proliferates and disharmonizes learning and teaching experiences and academic achievement. The expanded version of the technology acceptance model (TAM2) is the theoretical foundation for this research. TAM2 provides insight into how the perceptions of students and teachers matter when adopting and using ICTs in educational practices. Depending on these perceptions of perceived ease of use and usefulness, using ICTs in educational practices can impact intentional and behavioral use, currently and futuristically. Subjective norms also influence individuals' perceptions and willingness to use ICTs for educational practices. Limitations, strengths, and future recommendations and directions are discussed.

1. Introduction

This paper is an extension of work originally presented at the 2022 11th International Conference on Education and Information Technology (ICEIT) [1]. The purpose of this paper is to demonstrate the effects of ICTs and how such usage proliferates and disharmonizes learning and teaching experiences and academic achievement. There is a need for this study, especially as research [1] suggested social networking sites (SNSs) have stated a claim in educational practices while losing the battle to the 2019 coronavirus pandemic and onward situations. It was suggested that future researchers take a student-center approach and to study teachers' and students' perceptions regarding the use of ICTs for education-based involvements [1].

Although the 2019 coronavirus and its variants are still omnipresent, onward situations remain inevitable. Outside of SNSs, information and communication technologies (ICTs) hold great promise in educational practices (EPs) [2], [3]. Education remains to be a significant factor in human connections and relationships and has become instrumental in increasing access to various opportunities [4]. It is at the heart of what, how, where, and when individuals learn and teach [5]. ICTs play a leading role in academic achievement and future endeavors, especially as it has become a considerable agent for changes in EP [5]. However, when deciding to place ICTs in educational institutions, many decision-makers and policymakers failed to consider students' and teachers' perceptions. This paper suggests that while ICTs for EPs are both beneficial and detrimental, these outcomes are influenced by students' and teachers' perceptions of acceptance and adoption, which affects their attitudes toward technology and intentional and behavioral use.

As the primary stakeholders in education, students', and teachers' perceptions of ICTs for EPs matter. Other factors that matter when using ICTs for EPs are subjects, types of digital apparatuses, the persons using the devices, and geographical locations [2], [6]. Material access, digital skills, technology literacy, and equality also matter [7], [8]. Many academic institutions have already embraced and implemented technology in their institutions. Much of this resulted from the inexplicable

*Corresponding Author: Angela Pearce, apearce4@my.gcu.edu

onset of the covid pandemic, forcing schools to use technology, thus accelerating the adoption and usage levels of ICTs for EPs. This implementation has resulted in a mega shift and has become revolutionary for improving the qualities and efficiencies of EPs [9].

In the 21st century, technology pedagogy and content knowledge (TPACK) is a requirement that teachers must encompass when engaging in new learning environments [10]. This requirement assists teachers restructure educational practices, thus minimizing the today versus the future learning and teaching gaps that currently exist [11]. The Organisation for Economic Co-Operation and Development (OECD) suggested that students and teachers could benefit from using innovative ICTs [3], primarily as teachers are critical assets to the qualities of EPs [12]. However, these benefits will result from how well the frontliners are trained and prepared to use the technologies provided, especially as ICTs have become the lifeline for EPs [2] and will continue to be post-pandemic and while facing onward situations. The benefits will also result from how accessible and easy ICTs are to use and navigate and their usefulness to students' and teachers' educational experiences and academic outcomes.

Governmental entities play a significant role in ensuring quality education [13]. Yet, digital and educational gaps between communities exist, which influence students' and teachers' intentional use of ICTs for EPs [14], [15]. It is essential to provide equal access to physical materials and resources to students and teachers regardless of their geographical location. It is equally crucial to equip classrooms with quality ICTs, maintain the facilities, keep the systems up to date, provide technical support, and train teachers and students to use them effectively and efficiently [9], [14], and [16]. For students and teachers to fully participate in EPs, they must be granted access to quality equipment and broadband connections [11], particularly as ICTs for EPs have become mandatory, not necessarily voluntary. They must not be subjected to digital divides or geographical disparities where they lack access to the necessary materials and resources to successfully engage in educational practices [8], [14], and [17], and [18]. The quality of training in ICTs for EPs is significant and can substantially affect learning and teaching experiences [3], [11], and [14]. Technology training can help students and teachers overcome their reluctance to learn and use ICTs for EPs [10]. Furthermore, it will enable them to perceive ICTs for EPs as easy to use and useful [10], resulting in intentional and behavioral use and positive attitudes.

Significant findings in some studies [10], [19] suggested that teachers' acceptance and self-efficacy were associated with behavioral and intentional use of technology. Teachers are a significant asset in the educational field [11]. So, when teachers feel they lack proper training, feel unprepared, and are forced to use ICTs in EPs, it reduces teaching quality [3], [14]. It also diminishes the chances of accepting, adopting, intentional use, and contributing to negative attitudes toward ICTs for EPs. Studies [6], [9] showed evidence that ICTs could improve students' academic achievement, particularly in science, but not necessarily for practicing skills, math, and reading. Students reportedly had higher academic achievement when teachers used ICTs for EPs than teachers who did not [2]. For the types of ICTs, the system functionality should meet the standards to increase acceptance and adoption [13], which means being easy to use and navigate, which leads to usefulness and intentional use. Digital content should be attractive and interactive, encompassing audio, video, and animated simulations related to the learning content [6], [13]. Several scholars demonstrated positive effects of using visual content in learning environments [5], [20], and [21]. These visual and auditory aids were said to capture learners' attention, make learning more engaging and cooperative, and improve academic experiences [20], especially in the attempt to boost language proficiency [5], [21]. Additionally, response time should be fast and consistent in ways that cultivate users' interest [13], whereas working from slow and outdated systems can result in lower use and nonuse.

ICTs have shifted societal and life dynamics, where the influences are felt in schools [11], [22]. Now more than ever, ICTs have gained a leading role in EPs for knowledge acquisition and academic achievement [1], [14], and [16]. It has reshaped education into more flexible and efficient learning in developing nations, providing optimal learning, and teaching experiences [13], [16], and [22]. In OECD countries, the use of ICTs for EPs has increased significantly and has become an amalgamation into traditional education to provide cost-effective education [6], [13]. There are positive and negative associations of using ICTs in OECD countries to look up ideas and information [9]. However, the usage of ICTs in EPs in non-OECD countries remains marginal, as the effects may be less pronounced due to low levels of effective teaching [9].

There continues to be controversy and mixed views about how beneficial ICTs are in EPs and how it influences students' performance and academic success [4], [6]. It remains questionable whether introducing ICTs in EPs guarantees the acceptance and continued use or whether the material being taught, and the digital infrastructures are ample enough to implement in academic institutions [13]. In a study, [6] illustrated the differences of opinions in studies investigating the use of ICTs for EPs. Something researchers term the *null* effect [9]. Some advantages were related to improving students' academic outcomes by placing a wealth of information at their fingertips and more comprehensive resources [9].

Positive associations have been linked to using ICTs to look up information and explore new ideas that are otherwise unavailable in traditional settings [9]. They provide students and teachers with flexibility and autonomy, thus improving their attitudes and experiences toward usage [6]. Using ICTs in EPs can foster digital competencies required to reap the benefits that innovative technologies ICTs offer [3]. ICTs in EPs can optimize learning and teaching experiences and increase well-being [16], [23].

Furthermore, using ICTs for EPs could increase social presence, collaboration, and cooperation and promote active, diverse, and inclusive learning environments [1], [24]. Multiple studies [6], [11], and [25] argued that using ICTs can allow teachers to augment their teaching materials by making them more attractive and engaging, thus improving academic outcomes. Information and communication technologies can support teachers in assessing students' progress and levels of engagement at a group and individual level, monitor their

behaviors, and provide instant feedback or get assistance from teachers and other students [1], [16], and [25]. It also provides the means to form quality student- teacher dyads, increase communication, cooperation, and collaborative efforts, and boost social enrichment and academic and personal well-being [1], [11], [16], and [23]. When ICT activities fail to improve students learning experiences, it takes away from other activities that could potentially harvest students' academic outcomes [9]. It also removes their willingness to engage and accept ICTs for EPs.

While ICTs are instrumental in EPs, research has suggested that ICT use cannot determine whether students have learned what was taught, as good grades do not predict academic success [26]. In contrast, research suggested that ICT use can influence students' academic achievement, where specific environmental and family dynamics and using ICTs passively and recreationally resulted in low grades and markings [27]. ICTs in EPs can distract students from learning [1], [6], [16], and [17]. Placing an abundance of information at students' fingertips can result in cognitive and information overloads and technoference, thus negatively affecting their ability to process information [1]. Furthermore, ICTs can result in social disengagement, exclusion, or diminished human interactions [1], [6], and [14]. Studies have suggested that students are less creative when using ICTs for EPs and can be easily targeted for cyberbullying [1], [6], and [16].

A significant barrier associated with using ICTs for EPs is that many school districts and students, mainly those in rural and undeveloped areas, face the triple digital divide of not having physical, skill, and usage accessibility [14], [16], [17], [28], and [29]. Also, they cannot afford the required ICTs or home broadband connectivities to engage in EPs, which contributes to poor academic achievement, reduced social presence, attrition or dropout [14], [16]. More importantly, persons residing in rural communities face geographical disparities of lacking quality broadband and internet connections [14], [17], [23], and [30].

Using ICTs for EPs is also faced with endogeneity problems- unobservable characteristics that can dramatically impact students' and teachers' willingness to accept and intentionally use ICTs for EPs [6]. Many ICT interactions are done in private and unobservable settings. Moreover, individuals use technology to engage in activities other than for EPs. Therefore, it would be difficult to determine how often students and teachers intentionally engage in ICTs for EPs. It may also be nuanced to suggest that technology acceptance results from using it for EPs, whereas technologies offer many activities that could facilitate acceptance and intentional and behavioral use.

This study extends work originally presented at the 2022 11th International Conference on Education and Information Technology (ICEIT), where a study [1] suggested that social networking sites (SNSs) have stated a claim in educational practices while losing the battle to the 2019 Coronavirus pandemic and onward situations. This study examines the use of ICTs for EPs. The technology acceptance model (TAM2) is used to ground this research [31]. Adopting elements from this model will show how students' and teachers' perceptions matter when decisions are made to implement and use ICTs for EPs. Many of these perceptions result from perceived ease of use, usefulness, attitudes toward technology, and social influences, which may lead to intentional and behavioral use. Although technology holds a dominating presence in human lives and society, and many students already use ICTs for reasons other than for academic achievement, using ICTs for EPs may not directly affect their intentional use. Limitations and future recommendations are discussed.

2. Theoretical Foundation

2.1. Technology Acceptance Model - 2

The technology acceptance model- TAM2 created in 2000 is an expansion of the 1989 technology acceptance model (TAM) and is used to guide this paper. Figure 1 displays the TAM2 model. It expands the model by incorporating the perceptions of students and teachers, information and communication technologies (ICTs), and educational practices (EPs) related to perceived ease of use (PEOU), perceived usefulness (PU), intentional use (IU) or behavioral use (BU), and social influences (SI). Since ICTs dominate most human lives and society and now play a significant role in communication methods, socialization, knowledge acquisition, and academic success, students' and teachers' perceptions matter when deciding to use ICTs for EPs. Additionally, the information provided below defines the concepts of TAM2 and its external variables and demonstrates their association with ICTs and EPs.

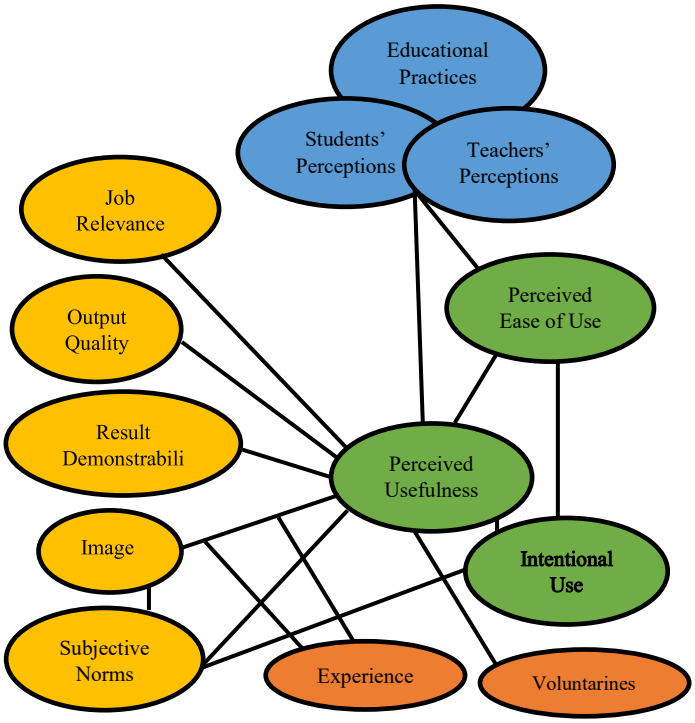


Figure 1: Technology Acceptance Model-2 and teachers' and students' perceptions of information and communication technologies for educational practices.

TAM stemmed from the theory of reasoned action (TRA) and the Theory of Planned Behavior (TPB) [32]. The TRA comprised three pillars, attitudes, subjective norms, and behavioral intentions, and focuses on two groups of variables: subjective norms and attitudes [33], [34]. TRA was designed to explain human behavior. It is also based on the premise that intention is

the primary determinant of an individual's actions and behavior [35]. As such, behavior is a direct function of intention [35].

As an extension of TRA, the central factors of TPB are individuals' intention to perform a given behavior [36]. These intentions are assumed to capture motivational factors that influence behavior, how hard a person is willing to try, and their effort in performing the behavior. The stronger the intention to engage in a specific behavior, the more likely it will be performed [35]. TPB comprises attitude toward the behavior, subjective norms, and perceived behavioral control (actual behavior) [32], [36].

The TAM model explains the behavior, attitudes, and intentions to use and adopt technology [37], [38], [39], and [40]. Constructs of TAM encompass various components, such as perceived ease of use (PEOU), perceived usefulness (PU), attitudes toward use (ATU), behavioral intentions to use (BIU), and actual use (AU) [39], [40]. PEOU is hypothesized to strongly influence PU, which directly impacts the ATU of various technologies [38], [41]. However, PEOU indirectly affects ATU [38]. Subsequently, BIU is influenced by AU [41]. Additionally, intentionality predicts users' willingness to adopt and use ICTs [39], [42]. These intentions are also influenced by an individual's attitudes [42]. The TAM model has been proven to be robust, the most influential, and has been widely applied to various regions examining technology acceptance and usage [13], [31], and [42]. Similarly, it is the most popular theory applied to research investigating students' behaviors and using and accepting ICTs for EPs [43], [44]. When individuals perceive ICTs as easy to use, useful, and beneficial, it increases their willingness to adopt, adapt, and accept technology [1]. More importantly, they develop positive attitudes toward technologies, resulting in continued and intentional use [39].

Due to the significant progress made over the past few decades in explaining and predicting users' acceptance and adoption of ICTs, the TAM model was expanded, constructing TAM2 [31]. The expansion aimed to determine how intentional use changes when individuals gain experience using ICTs [31]. While TAM rejected subjective norms due to their insignificant impact on usage, the TAM2 model incorporates these social aspects [40]. TAM2 theoretical constructs spanned from social influences (i.e., subjective norms, voluntariness, and image) and cognitive instrumental processes (i.e., job relevance and output quality) [31], [45]. The extension identifies and theorizes the social and cognitive influences that expand and determine usage perceptions [31], [45]. Job relevance, output quality, result demonstrability, and perceived ease of use capture to influences of cognitive instrument processes on perceived usefulness [45]. Social and cognitive influences are the most influential aspects that may result in students' and teachers' willingness to adopt ICTs for EPs. For example, when examining social influences or forces, many students are teachers face geographical disparities resulting from a lack of accessibility to physical materials and resources. Geographical disparities are also a result of poor broadband connectivity and internet access or no connections at all. Demographic disparities result from residing in rural rather than urban communities [14], [23]. Another contributing factor is financial stability, determining whether students, teachers, or other educational stakeholders can afford the essential ICTs needed to increase academic experiences and performances.

TAM2 also measures adoption behavior [31], [38], and [39]. Reputable researchers [35] suggested that behavior is a function of intentions. However, intentional behaviors are attributes of attitudes and beliefs [41]. Frequently, behaviors and attitudes are not factored into adopting and accepting ICTs for EPs. Moreover, this framework focuses on three interrelated social forces which impact an individual's decision to adopt, reject, or intentionally use innovative systems [31]. When students and teachers lack access to vital resources, digital skills, or the wherewithal to acquire quality equipment and internet connections, these social normalcies make the decisions for them.

2.2. TAM2 Concepts Defined

TAM has been used in various studies examining individuals' technology acceptance and usage.

- ***Perceived Ease of Use***. Perceived ease of use (PEOU) is a cognitive instrumental process that determines perceived usefulness and refers to the degree to which individuals believe that using new and innovative ICTs is free of effort or effortless [31], [39], [41], and [46]. It is also defined by the degree to which usage lacks mental and physical effort [38]. Job and academic performances increase when little effort is put into using ICTs. [31]. PEOU is defined as the degree to which students believe using ICTs for EP is effortless and will simplify the learning process [5]. Therefore, PEOU is associated with students and teachers perceiving ICTs for EPs as easy to use, navigate, effortless, and user- friendly. When technologies are not user-friendly and users lack the necessary training, support, and technical knowledge, they are less inclined to adapt and accept ICTs for EPs [47]. Perceived knowledge about technology integration of ICTs for EPs directly influenced PEOU [48]. PEOU directly affects PU for both teachers and students [48]. Furthermore, when individuals find technology easy to use, they may find it useful, resulting in intentional usage and positive attitudes towards technology.
- ***Perceived Usefulness.*** Perceived usefulness (PU) is based on individuals believing that ICTs will improve their job performance [39], [41]. PU is the degree to which students and teachers perceive that using ICTs for EPs will enhance their performance in various ways [5]. Researchers [39], [48] posited that PEOU influences PU. That means using ICTs will assist individuals in performing or completing a task they set out to do. Several studies proved that PU strongly predicts intentional use [5], [20], and [38]. Therefore, the perceived usefulness of ICTs for educational practices is associated with students and teachers finding ICTs to be valuable tools that will assist them in effectively and efficiently completing specific educational tasks, inside and outside the academic settings. Even when students and teachers find ICTs for EPs useful, this does not necessarily contribute to their intentional use.
- ***Intentional Use.*** PEOU and PU are strong determinants of intentional use [31], [39]. However, intentional use can change over time with experience [31]. Attitudes toward

technology influence intended use [5], [48]. Several studies have demonstrated a positive effect that PEOU and PU can have on usage intentions [15]. Furthermore, researchers have suggested that system quality and trust in systems and institutions are critical factors affecting acceptance and adoption and users' satisfaction and intentional use [47], [49]. The program or content and the learning and teaching tools also influence the intention to use ICTs in EPs [5]. Therefore, it is essential to understand better which strategies influence the acceptance and adoption of ICTs for EPs [47]. Scholars have suggested fundamental ways to effectively implement innovative ICTs into EPs to improve learning and teaching experiences [11].

- ***Behavioral Use.*** The actual use of ICTs is a behavior [49]. Behavioral or actual use of ICTs for EPs is significantly associated with technology acceptance [5]. Intentional use of ICTs is influenced by an individual's attitude toward technology [49], [50]. Attitudes toward technologies are related to behavioral use [41]. The quality of services, satisfaction, trust in systems, and knowledge sharing were significant predictors of usage behaviors of ICTs for EPs [47]. When individuals set out to act, they do so without limitations or coercion [41]. This means peers or other social norms do not influence behavioral use. Therefore, it can be suggested that when teachers and students perceive ICTs for EPs to be easy to use and useful, they develop positive attitudes toward technologies, ultimately leading to intentional and behavioral use.

2.3. External Variables of TAM2 Defined

- ***Image.*** Image is a social process that refers to an individual's perception that using and accepting innovators will enhance their social status [31]. TAM2 is a supposition to subjective norms in that image is enhanced when individuals of a prominent group believe they can perform specific behaviors and that performing those behaviors will strengthen their standing and status within that group [31]. Therefore, image is students' and teachers' perceptions that innovative technologies will enhance their learning and teaching experiences or academic stance or status. Additionally, the image component can be seen as a significant contributor to increasing knowledge acquisition and academic achievement (students) and job performance (teachers), which others will recognize in ways that would increase their academic and professional status.

- ***Job Relevance.*** Job relevance is a cognitive instrumental process that determines perceived usefulness and refers to individuals' perceptions that ICTs are suitable for their job [31], [46]. The job relevance variable of TAM2 is a function of importance within one's job and using ICTs will support a person in completing specific tasks [31]. ICTs are advantageous when they enhance performance and assist in achieving functions [39]. Therefore, when ICTs can perform certain functions within the scope of educational practices, whereby enhancing academic experiences, the systems are categorized as being relevant to the job. Moreover, job relevance will ultimately lead to positive attitudes toward technology and continued usage.

- ***Output Quality.*** Just like most things, output quality is a result of input quality. Like job relevance, output quality is a cognitive instrumental process that predicts perceived usefulness. It refers to individuals perceiving ICTs as good enough to perform specific functions; however, it is also based on whether the tasks performed by ICTs fit the desired goal [31], [46]. ICT users are the ones that consider how well systems perform, thus determining their output quality [31]. An information system (IS) model suggests system and information qualities impact user satisfaction [49]. Therefore, system and information quality may result from output quality. When students and teachers find that the output of ICTs for EPs is of quality, it may increase their likeliness to develop positive attitudes, wherein increasing intentional and behavioral use. While output quality is significantly related to input quality and the quality of the machines and software, it may also depend on the quality of internet and broadband connections. How teachers and students interact with ICTs for EPs can also establish output quality.

- ***Result Demonstrability.*** Result demonstrability (RD) is a cognitive instrumental process that determines perceived usefulness and is based on physical results and being able to attribute ICT use to increase performance [31], [46]. RD tangible results are positively associated with perceived usefulness [31]. Moreover, individuals will develop positive perceptions about how useful ICTs are when outcomes are of quality and noticeable [31]. Therefore, when ICTs are used in EPs, and the physical results can be attributed to increased performance academically, this demonstrates RD. Furthermore, all elements that have led to intentional and behavioral use may result in RD.

- ***Subjective Norms.*** Subjective norms (SNs) are social processes that refer to individuals perceiving social influences or social forces as the reason for performing a behavior or not [31], [36]. External variables and others influence many individuals' behaviors and actions. Internalization is based on individuals taking on the perceptions of others they see as crucial in their lives [46]. Furthermore, SNs are when inferior persons believe superior others approve or disapprove of specific behaviors. These beliefs of the inferior persons will result in them internalizing the perceptions of the majority figure in which they base their decisions on whether to engage in certain behaviors or believe that ICTs are useful [31], [36]. TAM2 reflects three interconnected social strengths that play a critical role in individuals' willingness to adopt or reject innovative ICTs when given the opportunity [31]. These social dynamics are subjective norms, voluntariness, and image [31]. Because subjective norms focus on social experiences and influences, these social dynamisms can relate to facilitating conditions, environmental factors, affordability, and accessibility. The usefulness of ICTs can be based on the perceptions of others and internalized processes [31], [36]. However, when individuals set out to perform specific behaviors, they do so

without limitations or coercion, as peers or other subjective norms do not influence their behaviors [41].

- ***Voluntariness.*** According to the TAM2 model, voluntariness is associated with subjective norms and intention to use. Voluntariness is a social process that refers to how technology adopters and adapters perceive their decisions to adopt new and innovative technologies as voluntary and not mandatory; freedom to choose [31], [46], and [51]. Moreover, it represents an individual's state to act or behave in a specific manner [51]. The use of ICTs voluntariness is significantly associated with TAM elements that lead to intentional and behavioral use [46]. Therefore, when using ICTs for educational practices, the voluntariness aspect refers to how teachers and students use digital devices of their own volition; usage is not mandatory or forced, and users do not take on the perceptions of essential others or social influences. Individuals who take the initiative to learn new and innovative technologies increase their experience and confidence levels in navigating the various functions [46] related to job relevance and evaluating the output quality.

There are two primary factors associated with actual or intentional use [51]. These factors are ease of use (technological contacts) when using ICTs in mandatory environments (low voluntariness) and 2) organization facilitating conditions (implementation context) and voluntary environment (high voluntariness) [51]. Demonstrating such relationships, when teachers and students find ICTs easy to use and usage or adoption is mandatory, it contributes to low voluntariness, which indicates that they were forced to use ICTs and were not provided with a choice [51]. Some students and teachers are exposed to specific organizational and facilitating conditions where ICTs are readily implemented. However, when using ICTs is mandatory, they are free to use or not use ICTs for educational practices, thus demonstrating high voluntariness in ICT use [51]. In sum, facilitating conditions predicted voluntariness and the actual use of ICTs in EPs [51].

3. Methodology and Study's Focus

This paper is an extension of work originally presented at the 2022 11th International Conference on Education and Information Technology (ICEIT) [1].

A researcher [1] suggested that social networking sites (SNSs) have stated a claim in educational practices while losing the battle to the 2019 coronavirus pandemic and onward situations. This study focuses on previous literature on accepting and adopting information and communication technologies (ICTs) in educational practices (EPs). ICTs have stated a claim in EPs and become a facilitating factor in academic engagements and success. However, when making decisions to implement ICTs in educational institutions, decision-makers and policymakers fail to consider the perceptions of students and teachers. This research seeks to understand the perceptions of students and teachers regarding the usefulness of ICTs for EPs as they relate to social influences and intentional use. The questions addressed in this research are:

1. What factors influence a student's perception of using ICTs in EPs that results in intentional use?
2. What factors influence a teacher's perception of using ICTs in EPs that results in intentional use?

4. Students' Perceptions of ICTs for EPs

Several dynamics influence students' perceptions of ICTs for EPs. Students are more likely to continually and intentionally use educational-based infrastructures through individual acceptance and personal willingness to using ICTs in EPs [5], [16], and [52]. In an article, a scholar [1] suggested various intricacies of using technology for learning. Some intrinsic factors include beliefs and attitudes toward ICTs [31], [39] cognitive abilities, learning styles [21], ICT literacy and competence [5], [52], and [53], academic intellect, motivation [5], [54], and gender [54], [55]. Moreover, previous experience, learning content and subject, and grade level [54] also contributed to students' perceptions of using ICTs in EPs.

While internal and self-motivating factors influence students' perceptions of using ICTs in EPs, students' willingness to use ICT and EPs depend on their learning styles [21]. More importantly, external, and contextual factors also shape their perceptions. These underlying forces are associated with social and peer influences, access to quality equipment, internet connections, training, academic settings, and learning environments [5], [23], and [52]. Many students feel they could use ICTs for EPs, especially as technology has become a significant part of their lives [54]. Children who started using technologies at younger ages demonstrated higher competence with using ICTs for EPs than those who began using technology later [56], resulting in positive perceptions of using ICTs in EPs and suggesting that ICT competency matters. Contrastingly, students from low socioeconomic and poor households who did not have previous exposure to ICTs performed at lower academic rates than students with prior ICT exposure and experiences [20]. However, this is not to suggest that students with later or no exposure to various technologies will develop negative perceptions about using ICTs, but that more experienced users were less confident in using ICTs for EPs than students with less experience [54]. Relatively, those students in lower grade levels were more satisfied with using ICTs in EPs than those in higher grade levels [54].

When considering gender differences, researchers [54], [55] revealed that male students were more confident in using ICTs in EPs than females. This finding could be based on some students' having equitable access and independently making decisions regarding the use of technology. Nonetheless, females may be less inclined to use technology because they perceive ICT usage as a male-centric activity [54].

Geographically, students in Portugal and Ukraine favored using ICTs for EPs compared to United Arab Emirate (UAE) students [57]. Students in all three countries, Portugal, Ukraine, and UAE, reported favorable perceptions and attitudes toward using ICTs in EPs [57]. This is because it allowed them to manage their time, work independently, and at their own pace and from their desired learning environments [57]. While some students in the United States found that educational-based technologies were too restrictive, their perceptions were based on teachers' knowledge, skills, and abilities (KSAs) to use ICTs for course instructions effectively and efficiently. American students' perceptions also resulted in whether teachers could assist in resolving technical issues [58].

Additionally, the learning environments, whether teachers cared about students' academic success, and whether students experienced a disconnect or disassociation between in-class and out-of-the-classroom settings also influenced their perceptions of how beneficial ICTs were for engaging in EPs [58]. In the United States and France, the quality, satisfaction, and loyalty of services provided to students through e-learning played a significant role in students constructing their perceptions of using ICTs in EPs [59], [60]. The quality of educational technologies that can assist with optimal learning experiences and successful outcomes also influenced students' perceptions of using ICTs in EPs [61].

With innovative technologies holding a ubiquitous presence in human lives and society, it is essential to account for how students use ICTs in everyday life and how such usage differs from educational purposes [53]. This is mainly because some students reported that educational-based technologies and digital infrastructures are too restrictive [58], particularly as they have unlimited access and usage of everyday technologies outside academia. While a few studies [53], [57], and [62] reported that past and previous experiences synchronously and asynchronously impact students' perceptions of using ICTs in EPs, others [52] suggested that this is not the case. Instead, some researchers [52] advocated that past experiences are not predictors of a student's willingness to accept ICTs for EPs simply because students' ICT usages and engagements differ from one context to the next.

Scholars [53] proclaimed that Chinese students' perceptions of using ICTs were based on past experiences, as they spent numerous hours on their devices which increased their comfort levels with using ICTs for EPs. Because of technology's flexibility and convenience, students in India developed positive perceptions of using ICTs in EPs [62]. Indonesian students' perceptions resulted from ICTs providing them with the capability to improve their motivation to learn, level of independence, understanding of the topics being taught, time management skills, self-discipline, and interactions. Information and communication technologies (ICTs) also increased their exposure to vital resources that would have been otherwise inaccessible without technology [52]. Some researchers [62], found that students preferred to use their smartphones, pre-recorded instructions, and take online quizzes [62]. Also, in an analysis based in India and South Korea, researchers [63] found that students' perceptions of using ICTs in EPs were influenced by classroom and social interactions, motivation, course structures and layout, instructors' knowledge, and facilitation [57]. Not long ago, Turkish students' perceptions of using ICTs resulted in their basis and motivation to learn [54].

Studies in several countries demonstrated students' positive perceptions and attitudes toward using ICTs in EPs [52], [53], [54], [57], [62], and [63] and how using ICTs are attractive but can be both beneficial and detrimental [1], [14], [16], and [23]. However, Canadian students demonstrated negative perceptions of ICTs in EPs due to cognitive overload, poor communication and interactions between students and teachers, technical issues, and inability to follow and navigate course structures [64], [65]. Using ICTs in EPs is not always beneficial, whereas some students, experienced and inexperienced, will face challenges and complexities related to cybersecurity, hacks, viruses, cyberbullying, and internet instability and outages. When using ICTs in EPs, some students face concerns about interacting with teachers and peers [57]. Some of these issues are associated with technical problems. However, technological disruptions and complexities were not always significant factors influencing a student's perception of using ICTs in EPs [52], [57]. Skills or digital literacy, time management, language, behavior, motivation, and their objectives and goals to learn were [52], [54], [57], and [66].

Digital knowledge and skills were primary determinants of a student's willingness to participate fully and accept ICTs in EPs [52], [66]. Students are not motivated to learn, accept, or use technology if they lack the necessary confidence to use ICTs for EPs [55]. This can result in students' fear of learning and unwillingness to engage or position themselves in novel situations due to negative technological experiences [55]. It can also increase attrition, enrollment, and academic disengagement rates [1], [23]. Some students' motivation to use ICTs in EPs deals with their ability to strategically communicate and interact with other students and teachers [52]. For novice language learners, systematic communication practices are essential, especially as they need to develop their language skills and competencies [5], [21], and [57] or have a desire for academic success [61]. Students that developed positive perceptions toward ICTs for EPs were those who found ICTs to be useful and easy to use [52]. The usefulness aspect allowed them to improve their understanding and academic competencies in a specific subject [52]. It also provided them the means to study independently and increased self-discipline and motivation to learn [52], [54]. Those who developed negative perceptions and attitudes toward ICTs found that the digital infrastructures and platforms were not designed to meet students' specific needs, presented navigation complexities, were not user-friendly, and produced cognitive overloads [65].

As some students continually face digital divides, such as access to physical materials and resources and geographical disparities [1], [14], [16], [18], [23], and [30], other students have access and can afford ICTs [57]. Academicians [1], [23], and [67], illustrated some advantages and disadvantages of using ICTs in EPs, a significant benefit of using ICTs in EPs is that students can actively, authentically, and cooperatively engage in the learning process, which ultimately resulted in them developing a positive perception of using ICTs in EPs. Overall, students' positive perceptions of using ICTs in EPs outweighed negative perceptions, but this may not be the same when considering teachers' perceptions.

5. Teachers' Perceptions of ICTs for EPs

Teachers' perceptions of using ICTs in EPs matter, especially as they are responsible for integrating technology into their classrooms, curricula, and pedagogies [68]. While there are several advantages and disadvantages of using ICTs in EPs [1], teachers' positive or negative perceptions or attitudes toward ICTs result from how easy and useful they use them in EPs [69]. When teachers perceive ICTs as useful and easy to use, they develop positive attitudes toward technology, which increases their intentional use [15]. Factors that motivated teachers to use ICTs in EPs resulted from self-efficacy, educational values, impact on teaching, and the quality of training in ICTs to be used in EPs [68].

Furthermore, educationalists accepted educational-based technologies because they provided them with the flexibility to deliver instructions [1], [70], and [71]. ICTs assist teachers with obtaining information easily and swiftly and make teaching and learning more interesting [72]. While ICTs offer significant benefits in teaching practices, teachers' gender [71] and the subjects they teach [73] played an essential role in their willingness to adopt and use ICTs in EPs. Furthermore, based on their own experiences, many teachers will form their individual attitudes and beliefs about how valuable and effective ICTs are when used for educational purposes.

Endogenous and exogenous factors influenced teachers' perceptions of using ICTs in EPs [74]. Endogenous factors were based on nonmanipulative conditions, while exogenous factors were manipulative and changeable [74]. Teachers and school levels are both endogenous and exogenous dynamics [74]. However, when considering the use of digital pedagogies and instructions, the digital platforms must not only be understandable and available to teachers and within the academic milieu, but students must be able to comprehend the material they are being taught through digital instructions.

Some teachers are apprehensive about using ICTs in EPs when they lack the training to use technological infrastructures, platforms, software, and hardware effectively and efficiently. There are two sides to teachers' perceptions of using ICTs in EPs, whereas some teachers had positive attitudes toward technologies and felt that ICTs were beneficial in teaching practices [72]. However, some teachers had opposing views and felt there was no value or benefit in using ICTs in EPs where language learning is concerned [72]. Despite such trepidations, academics [68], [72] advocated that those teachers who taught in Indonesia's rural school districts felt using ICTs in EPs was beneficial. Teachers found that using ICTs in EPs effectively increased students' motivation to learn, fostered positive attitudes toward ICTs, and made learning and teaching activities more exciting and enjoyable [68]. Additionally, teachers found that using ICTs in EPs helped students better understand how various technologies affected their lives. Regardless, teachers perceived ICTs in EPs to improve teaching performances and assisted them with developing new teaching skills [68].

In the West Indies, teachers' perceptions of ICTs significantly influenced how beneficial ICTs were in EPs [74]. These considerations were based on cooperation, job satisfaction, and self-confidence [74]. For Malaysian teachers, evidence showed that these teachers faced various challenges in integrating and using ICTs in EPs [69]. They found that using ICTs was helpful in ways that increased their job performance and productivity and allowed them to work more quickly and effectively. A teacher's perception of using ICTs heavily dealt with them finding that the provided digital infrastructures were straightforward, understandable, and easy to remember and control [69]. In Sweden, teachers had positive perceptions of using ICTs in EPs and used the digital platforms for various reasons [71]. Teachers felt that ICTs provided them with the flexibility needed in the time and space and allowed them to provide students with online instructions and readily available assignments [71]. Palestinian teachers perceived ICTs to positively influence their teaching and educational practices [24]. In Ethiopia, some researchers [75] found a positive relationship between teachers' perceptions toward ICTs in teaching and learning practices, especially when they were encouraged to do so and received the necessary support. Although using ICTs in EPs increased the quality of courses and productivity and allowed teachers to prepare and deliver course material for students more efficiently, most teachers did not use ICTs as a teaching tool [75]. Significant barriers to using ICTs in EP were that teachers lacked the necessary technical knowledge and experienced substantial shortages of resources and materials [75].

In East African countries, students and teachers continually face digital transformation challenges of not having access to various technologies, physical materials, and resources. African countries, among other developed and undeveloped countries, incessantly lack vital resources for teaching and learning, resulting in educators' and pupils weakened digital knowledge and technology fluencies [12], [76]. Sub-Saharan African teachers who actively use ICTs in EPs are those who acquired equipment and training at their own expense [77]. While students are excited to use ICTs in EPs, teachers are open to such usability provided they receive the appropriate training and support needed to operate the devices effectively and efficiently [77]. Human, physical, technical, system, and policy environmental readiness' are some significant caveats that African countries face [77].

Insofar, there are several barriers that impact Indian teachers' perceptions of using ICTs in EPs [78]. Moreover, researchers found that while a large portion of teachers was ready and willing to use ICTs in EPs, many teachers were not [78]. Teachers with unfavorable attitudes toward ICTs in EPs felt that using technology was not beneficial to the teaching practices [78]. For example, several studies demonstrated challenges teachers faced when implementing ICTs into EPs. Challenging encounters were adequate training, teacher and technical competencies, accessibility and broadband/internet connectivity issues, lack of technical support, quality of equipment, updated software, and inability or time constraints to shift from formal/traditional teaching paradigms to informal teaching practices [68], [69], [71], [72], [75], and [78]. Some academic investigators [75] conjectured that equipping teachers' and academic institutions with ICTs are not enough to successfully integrate ICTs into EPs. Teachers and students must receive adequate training so that the ICTs for EPs afforded for learning and teaching can easily be understood.

6. Research Strengths, Implications, and Future Recommendations

This study demonstrates various strengths and limitations regarding the perceptions of students and teachers as they relate to their willingness to accept and intentionally use information

and communication technologies in educational practices. Furthermore, this study provides recommendations that future researchers could conduct to expand this study or add to existing bodies of literature regarding the use of ICTs in EPs. Strengths, limitations, and future recommendations are as follows:

6.1. Research Strengths

Some significant strengths of this research are that it illustrates how students and teachers are the primary stakeholders and frontliners when using ICTs in EPs. Yet, decision-makers and policymakers fail to consider their perceptions of how valuable ICTs are in EPs, especially regarding job relevance, output quality, result demonstrability, perceived ease of use, and all cognitive instrumental processes that contribute to perceived usefulness [46]. Another strong point is that the information presented is timely and relevant, mainly because ICTs have stated a claim not only in human lives, but in EPs and have become essential in educational practices. Students' and teachers' perceptions matter. This research demonstrates how using ICTs provides flexibility in teaching and learning practices and how they can increase the potential of academic experiences and achievement given that students and teachers are provided with adequate training and vital resources and equipment. Most importantly, this study demonstrates how teachers' and students' perceptions of using ICTs in EPs matter, as they are not harvested based on opinion but real-world circumstances and actualities.

6.2. Study Limitations

This study comes with limitations. A limitation associated with this study is that it is based on previous literature. At the same time, this study demonstrated the perceptions of students and teachers and how their perspectives matter when making decisions to integrate information and communication technologies into educational practices. However, examining specific learning management systems (LMS) and digital platforms while considering students' and teachers' perspectives jointly and individually may yield precise results. Comparing students and teachings in rural school districts versus urban academic settings may produce other meaningful outcomes, those that are distinct from what is revealed in this study.

6.3. Future Recommendations

Although this study provides evidence to show how students' and teachers' perceptions matter when using ICTs in EPs, it is suggested that future researchers examine students' and teachers' educational experiences as they relate to using ICTs in EPs while taking a diverse approach. A qualitative or quantitative approach that produces robust information can help understand how ICTs influence EPs among students and teachers. Future researchers should consider various demographics, socioeconomic status (SES), environmental and social influences, age, culture, geographical positions, and accessibility when trying to understand the perceptions of students and teachers as they relate to their intentional use of ICTs in educational practices. All educational stakeholders must realize that students and their learning styles are not homogenous and should be viewed differently. Researchers should also note that teachers and their teaching styles are heterogeneous, as these aspects should be considered when examining the perceptions of students and teachers and their willingness to using ICTs in EPs, where technology acceptance paradigms are concerned.

7. Conclusion

The main objective of this paper was to expand the previous work presented at the 2022 11th International Conference on Education and Information Technology (ICEIT) [1]. This study [1] suggested that social networking sites (SNSs) have stated a claim in educational practices while losing the battle to the 2019 coronavirus pandemic and onward situations. However, when making critical decisions on whether to implement technologies into educational practices, the perceptions of students and teachers matter, but in frequent cases, they are not always considered. Therefore, this paper addressed two pressing questions a) what factors influence students' perceptions when using ICTs in EPs that result in intentional use, and b) what factors influence teachers' perceptions when using ICTs in EPs that result in intentional use. Understanding these factors and knowing what it takes to implement ICTs into EPs successfully, primary stakeholders will start to value the perceptions of students and teachers and not just focus on the affordability and geographical aspects.

While many governmental entities and policymakers encourage school districts to implement technologies into their learning and teaching practices, they do not bring students and teachers to the forefront to acknowledge what is doable. Instead, they take it upon themselves to purchase and decide whether to implement technology into schools, leaving students and teachers in the background. Yet, students and teachers are the primary users.

There are numerous advantages and disadvantages of using ICTs in EPs. Meanwhile, various endogenic and exogenic factors influence the perceptions of students and teachers about using and accepting educational-based digital infrastructures and platforms that will result in intentional use. However, to increase intended use, it is imperative that frontline users, such as school districts, students, and teachers, be provided with the necessary resources. The properties include funding, adequate training, access to quality equipment, software, internet connections, technical support, and unending support from primary decision-makers and educational and governmental stakeholders. Policymakers must pay attention to endogenous and exogenous factors and value students’ and teachers' perceptions when deciding whether to implement ICTs in EPs [74]. Otherwise, it can result in students and teachers developing negative attitudes toward technology use in educational practices. These negative attitudes can stem from teachers and students facing the challenge of not being heard or receiving adequate training or resources.

As ICTs have stated a claim in educational practices, the perceptions of students and teachers matter. Their perceptions go beyond the scope of elements described in the technology acceptance model regarding ease of use and perceived usefulness and spill over into the extended model-TAM2. Teachers and students must receive essential materials and resources to successfully integrate digitally-based pedagogical instructions into their learning and teaching curricula and reap the full benefits technologies offer. Training will increase students' and teachers' confidence levels, comfortability in use, and knowledge acquisition. Studies have suggested that using ICTs in EPs can optimize learning and teaching experiences, mainly when training is provided. However, the training process cannot be a one-time thing, as training must be ongoing, as hardware and software packages and licenses are updated and supported frequently. Teachers' and students' perceptions are valuable assets in discovering the practicality of how beneficial and easy ICTs are to use EPs. Given that, this study suggests that primary decision-makers account for teachers' and students' perceptions when deciding what technologies are implemented into learning and teaching practices. In short, teachers and students are the frontliners when using ICTs in EPs, as their perceptions are not a matter of opinion but a matter of fact.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. Pearce, "Socially networking sites stating claim in educational practices while losing the battle to COVID-2019 and onward situations", in 2022 11th International Conference on Education and Information Technology (ICEIT), IEEE, 105-111, 2022, doi:10.1109/ICEIT54416.2022.9690736.

[2] J. Bryant, F. Child, E. Dorn, S. Hall, New Global Data Reveal Education Technology's Impact on Learning, McKinsey & Company, 2020.

[3] Organization of Economic Co-operation and Development (OECD), "ICT Resources in School Education: What Do We Know from OECD Work?," OECD, 2020, https://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/EDPC/SR/RD(2020)2&docLanguage=En

[4] The World Bank, Understanding Poverty / Education Overview, World Bank Group, 2022. https://www.worldbank.org/en/topic/education/overview

[5] Wiyaka, J. Mujiyanto, D. Rukmini, "Students' perceptions on the usefulness of ICT-based language program," English Language Teaching, **11**(2), 53-60, 2018, doi:10.5539/elt.v11n2p53.

[6] M. Fernández-Gutiérrez, G. Gimenez, J. Calero, "Is the use of ICT in education leading to higher student outcomes? Analysis from the Spanish autonomous communities", Computers & Education, **157**, 1-15, 2020, doi:10.1016/j.compedu.2020.103969.

[7] J. Adhikari, A. Mathrani, C. Scogings, "Bring your own devices classroom: Exploring the issue of digital divide in the teaching and learning contexts," Interactive Technology and Smart Education, **13**(4), 323-343, 2016, doi:10.1108/ITSE-04-2016-0007.

[8] J. A. G. M. van Dijk, "The evolution of the digital divide – The digital divide turns to inequality of skills and usage," in Digital Enlightenment Yearbook 2012, IOS Press, 57-78, 2012.

[9] O. Falck, C. Mang, L. Woessmann, "Virtually no effect? Different uses of classroom computers and their effect on student achievement", Oxford Bulletin of Economics and Statistics, **80**(1), 1-38, 2018, doi:10.1111/obes.12192.

[10] Y. J. Joo, S. Park, E. Lim, "Factors influencing preservice teachers' intention to use technology: TPACK, teacher self-efficacy, and technology acceptance model," Educational Technology & Society, **21**(3), 48-59, 2018.

[11] L. Ratheeswari, "Information communication technology in education," Journal of Applied and Advanced Research, **3**(S1), S45-S47, 2018, doi:10.21839/jaar.2018.v3iS1.169.

[12] S. Kommers, M. van der Zijde, A. Elfferich, E. Thravalou, Digitalisation of education in East Africa: Needs, experiences, and opportunities for the future, Nuffic, 2021. https://www.nuffic.nl/sites/default/files/2021-09/digitalisation-of-education-in-east-africa-needs-experiences-and-opportunities-for-the-future.pdf

[13] F. Kanwal, M. Rehman, "Factors affecting e-learning adoption in developing countries-empirical evidence from Pakistan's higher education sector," in 2017 IEEE Access, **5**, 10968-10978, 2017, doi:10.1109/ACCESS.2017.2714379.

[14] A. Pearce, "The digital gap still exists, generationally, rurally, and academically," in 2020 12th International Conference on Education, Technology, and Computers (ICETC), ACM, 215-222, 2020 doi:10.1145/3436756.3437048.

[15] F. Weng, R-J. Yang, H-J. Ho, H-M. Su, "A TAM-based study of the attitudes towards use intention of multimedia among school teachers," Applied System Innovation, **1**(3), 1-9, 2018, doi:10.3390/asi1030036.

[16] A. Pearce, "Optimal learning using technology: Amplifying students collaboration and social networking," in 2021 PUPIL: International Conference of Teaching, Education, and Learning (ICTEL), **5**(2), 19-32, 2021, doi:10.20319/pijtel.2021.52.1932.

[17] B. Warf, "Teaching digital divides," Journal of Geography, **118**(2), 77-87, 2019, doi:10.1080/00221341.2018.1518990.

[18] A. J. van Deursen, J. A. van Dijk, "The first-level digital divide shifts from inequities and physical access to inequities in material access," New Media & Society, **21**(2), 354-375, 2019, doi:10.1177/1461444818797082.

[19] R. Sherer, F. Siddiq, J. Tondeur, "All the same or different? Revisiting measures of teachers' technology acceptance", Computers & Education, **143**, 1-17, 2020, doi:10.1016/j.compedu.2019.103656.

[20] E. T. Maziriri, P. Gapa, T. Chuchu, "Students perceptions toward the use of Youtube as an educational tool for learning and tutorials," International Journal of Instructions, **13**(2), 119-138, 2020, doi:10.29333/iji.2020.1329a.

[21] G. Sakkir, S. Dollah, J. Ahmad, "Students' perceptions toward using Youtube in EFL classrooms," Journal of Applied Sciences, Engineering, Technology, and Education, **2**(1), 1-10, 2020, doi:10.35877/454RI.asci2125.

[22] Y. Jung, J. Lee, "Learning engagement and persistence in massive open online courses (MOOCs)," Computers & Education, **122**, 9-22, 2018, doi:101016/j.compedu.2018.02.013.

[23] A. Pearce, "Achieving academic satisfaction through the use of education-based technologies: Strengthening students personal well-being while facing online learning mandates and digital disparities," London Journals Press, **21**(5), 19-36, 2021, doi:10.34257/LJRHSSVOL21IS5PG21.

[24] H. Qaddumi, B. Bartram, A. L. Qashmar, "Elevating the impact of ICT on teaching and learning: A study of Palestinian students' and teachers' perceptions," Education and Information Technologies, **26**, 1865-1876, 2021, doi:10.1007/s10639-020-10339-5.

[25] S. K. D'Mello, "Improving student engagement in and with digital learning technologies," in OECD Digital Education Outlook 2021: Pushing the Frontier with Artificial Intelligence, Blockchain and Robots, Paris, OECD Publishing, 67-92, 2021.

[26] A. Yli-Jyrä, "Optimal learning experiences in language technology education," ResearchGate, 1-14, 2014.

[27] S. Simões, T. Oliveira, C. Nunes, "Influence of computers on students' academic achievement," Heliyon, **8**(1), 1-13, 2022, doi:10.1016/j.heliyon.2022.e09004.

[28] J. A. G. M. van Dijk, "Digital divide research, achievements and shortcomings," Poetics, **34**(4-5), 221-235, 2006, doi:10.1016/j.poetic.2006.05.004

[29] J. A. G. M. van Dijk, "Digital divide: Impact of access," The International Encyclopedia of Media Effects, 1-11, 2017, doi:10.1002/9781118783764.wbieme0043.

[30] C. G. Reddick, R. Enriquez, R. J. Harris, B. Sharma, "Determinants of broadband access and affordability: An analysis of a community survey on the digital divide," Cities, **106**, 1-12, 2020, doi:10.1016/j.cities.2020.102904.

[31] V. Venkatesh, F. D. Davis, "A theoretical extension of the technology acceptance model: Four longitudinal field studies," Management Sciences, **46**(2), 186-204, 2000, doi:10.1287/mnsc.46.2.186.11926.

[32] I. Ajzen, "From intentions to actions: A theory of planned behavior," in Action Control: SSSP Springer Series in Social Psychology, Springer, 11-39, 1985.

[33] M. Fishbein, I. Ajzen, Belief, attitude, intention, and behavior: An introduction to theory and research, Addison-Wesley, 1975.

[34] B. Salgues, Health Industrialization. ISTE Press Ltd, 2016.

[35] G. D. Sideridis, A. Kaissidis, S. Padeliadu, "Comparison of the theories of reasoned action and planned behavior," British Journal of Educational Psychology, **68**, 563-580, 1998.

[36] I. Ajzen, "The Theory of Planned Behavior," Organizational Behavior and Human Decision Processes, **50**(2), 179-211, 1991, doi:10.1016/0749-5978(91)90020-T.

[37] F. Abdullah, R. Ward, E. Ahmed, "Investigating the influences of the most commonly used external variables of TAM on students' perceived ease of use (PEOU) and perceived usefulness (PU) of e-portfolios," Computers in Human Behavior, **63**, 75-90, 2016, doi:10.1016/j.chb.2016/05.014.

[38] H. Ahmad, A. Basden, "Non-disciplinary use of information system and the technology acceptance model," University of Salford Manchester, 2008.

[39] F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," MIS Quarterly, **13**(3), 319-340, 1989.

[40] P. Legris, J. Ingham, P. Collerette, "Why Do People Use Information Technology? A Critical Review of the Technology Acceptance Model," Information & Management, **40**(3), 191-204, 2003, doi:10.1016/S0378-7206(01)00143-4.

[41] S. R. Sakarji. K. B. M. Nor, M. M. Razali, N. Talib, N. Ahmad, W. A. A. W. M. Saferdin, "Investigating Students Acceptance of E-Learning Using Technology Acceptance Model Among Diploma in Office Management and Technology Students at Uitm Melaka," Journal of Information System and Technology Management, **4**(13), 13-26, 2019, doi:10.35631/JISTM.413002.

[42] P. Reddy, K. Chaudhary, B. Sharma, R. Chand, "The two perfect scores for technology acceptance," Education and Information Technologies, **26**, 1505-1526, 2021, doi:10.1007/s10639-020-10320-2.

[43] J. Henderson, N. B. Milman, "The technology acceptance model: Considerations for online educators," Distance Learning, **17**(3), 104-107, 2021.

[44] B. Šumak, M. Heričko, M. Pušnik, "A meta-analysis of e-learning technology acceptance: The role of user types and e-learning technology types," Computers in Human Behavior, **27**, 2067-2077, 2011, doi:10.1016/j.chb.2011.08.005.

[45] V. Venkatesh, H. Bala, "Technology acceptance model 3 and a research agenda on intervention," Decision Sciences, **39**(2), 273-315, 2008, doi:10.1111/j.1540-5915.2008.00192.x.

[46] M. Kaur, "Integration of ICT in education issues and challenges," in New Paradigm in Business and Education,. National Press Associates, 75-80, 2020.

[47] S. A. Salloum, M. Al-Emran, K. Shaalan, A. Tarhini, "Factors affecting the E-learning acceptance: A case study from UAE," Education and Information Technologies, **24**, 509-530, 2019, doi:10.1007/s10639-018-9786-3.

[48] P. Luik, M. Taimalu, "Predicting the intention to use technology in education among student teachers: A path analysis," Education Sciences, **11**, 1-14, 2021, doi:103390/educsci11090564.

[49] W. H. Deleon, E. R. McLean, "The Deleon and McLean model of information systems success: A ten-year update," Journal of Management Information Systems, **19**(4), 9-30. 2003, doi:10.1080/07421222.2003.11045748.

[50] S. R. Joseph, "Students' perspectives on ICTs acceptance and use in higher educational institutions of Botswana: Limkokwing University," International Journal of Engineering & Scientific Research, **3**(1), 185-198, 2015.

[51] B. Brevell, V. Arkorful, "LMS-enabled blended learning utilization in distance tertiary education: establishing the relationships among facilitating conditions, voluntariness of use and use behavior," International Journal of Educational Technology in Higher Education, **17**(6), 1-16, 2020, doi:10.1186/s41239-020—183-9.

[52] L. Vitoria, M. Mislinawati, N. Nurmasyitah, "Students' perceptions on the implementation of e-learning: Helpful or unhelpful?," Journal of Physics: Conference Series, **1088**(1), 1-6, 2018, doi:10.1088/1742-6596/1088/1/012058.

[53] A. Popovici, C. Mironov, "Students' perceptions of using learning technologies," Procedia-Social and Behavioral Sciences, **180**, 1514-1519, 2015, doi:10.1016/j.sbspro.2015.02.300.

[54] M. Kahveci, "Students perceptions to use technology for learning: Measurement integrity of the modified Fennema-Sherman Attitudes Scales," TOJET: The Turkish Online Journal of Educational Technology, **9**(1), 188-201, 2010.

[55] H. K. Yau, L. F. Cheng, "Gender differences of confidence in using technology for learning," The Journal of Teaching Studies, **38**(2), 74-79, 2012.

[56] L. Juhañák, J. Zounek, K. Záleská, O. Bárta, K. Vlčková, "The relationship between the age at first computer use and students' perceived competence and autonomy in ICT usage: A mediation analysis," Computers & Education, **141**(1), 1-11, 2019, doi:10.1016/j.compedu.2019.103614.

[57] P. Fidalgo, J. Thormann, O. Kulyk, J. A. Lencastre, "Students' perceptions on distance education: A multinational study," International Journal of Educational Technology in Higher Education, **17**(18), 1-18, 2020, doi:10.1186/s41239-020-00194-2.

[58] J. Stefl-Mabry, M. Radlick, W. Doane, "Can you hear me now? Student voice: High school & middle school students' perceptions of teachers, ICT and learning", International Journal of Education and Development using ICT, **6**(4), 64-82, 2010.

[59] S. Ivanaj, G-B. Nganmini, A. Antoine, "Measure e-learners' perceptions of service quality," Journal of Organizational and End User Computing (JOEUC), **31**(2), 83-104, 2019, doi:10.4018/JOEUC.2019040105.

[60] L. Pham, S. Williamson, R. Berry, "Student perceptions of e-learning service quality, e-satisfaction, and e-loyalty," International Journal of Enterprise Information Systems (IJEIS), **14**(3), 19-40, 2018, doi:10.4018/IJEIS.2018070102.

[61] D. Petko, A. Cantieni, D. Prasse, "Perceived quality of educational technology matters: A secondary analysis of students' ICT use, ICT-related attitudes, and PISA 2012 test scores", Journal of Educational Computing Research, **54**(8), 1070-1091, 2017, dio:10.1177/0735633116649373.

[62] T. Muthuprasad, S. Aiswarya, K. S. Aditya, G. K. Jha, "Students' perception and preference for online education in India during COVID-19 pandemic," Social Sciences & Humanities Open, **3**(1), 1-11, 2021, doi:10.1016/j.ssaho.2020.100101.

[63] H. Baber, "Social interactions and effectiveness of the online learning-A moderating role of maintaining social distance during the pandemic COVID-19," Asian Education and Development Studies, **11**(1), 159-171, 2021, doi:10.1108/AEDS-09-2020-0209.

[64] A. Pearce, "Socially-oriented technologies attractability influence on individuals academically and neurologically," in 2020 12th International Conference on Education, Technology, and Computers (ICETC), ACM, 165-171, 2020, doi:10.1145/3436756.3437040.

[65] C. Conrad, Q. Deng, I. Caron, O. Shkurska, P. Skerrett, B. Sundararajan, "How student perceptions about online learning difficulty influenced their satisfaction during Canada's Covid-19 response", British Journal of Educational Technology, **53**(1), 534-557, 2022, doi:10.1111/bjet.13206.

[66] N. Selwyn, O. Husen, "The educational benefits of technology competence: an investigation of students' perceptions" Evaluation & Research in Education, **23**(2), 137-141, 2010, doi:10.1080/09500790.2010.483515.

[67] R. Faizi, R. Chiheb, A. E. Afia, "Students' perceptions toward using Web 2.0 technologies in education", International Journal of Emerging Technologies in Learning (IJET), **10**, 32-36, 2015, doi:10.3991/ijet.v10i6.4858.

[68] M. Mahdum, H. Hadriana, M. Safriyanti, " Exploring teachers perceptions and motivations to ICT use in learning activities in Indonesia," Journal of Information Technology Education: Research, **18**, 293-317, 2019, doi:10.28945/4366.

[69] S. Ghavifekr, T. Kunjappan, L. Ramasamy, A. Anthony, "Teaching and learning with ICT tools: Issues and challenges from teachers' perceptions," Malaysian Online Journal of Educational Technology (MOJET), **4**(2), 38-58, 2016.

[70] S. Amir, M. S. Kamal, M. D. T. Shahria, L. Iftekhar, "Undergraduate engineering courses: A case study of emergency remote teaching amid large digital divide," 2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), 235-242, 2020, doi:10.1109/TALE48869.2020.9368316.

[71] J. O. Lindberg, A. D. Olofsson, G. Fransson, "Contrasting views: Student and teacher perceptions in ICT in education," The Proceedings of the International Conference on Information and Communication Technologies in Education (ICICTE), 1-10, 2016.

[72] C. V. Katemba, "Teachers' perceptions in implementing technologies in language teaching and learning," Acuity: Journal of English Language Pedagogy, Literature, and Culture, **5**(2), 38-51, 2020.

[73] Z. Walker, H. H. Kho, D. Tan, N. Lim, "Practicum teachers' use of mobile technology as measured by the technology acceptance model, Asia Pacific Journal of Education, **40**(2), 230-246, 2020, doi:10.1080/02188791.2019.1671808.

[74] S. Li, S. Yamaguchi, J-i. Takada, "Understanding factors affecting primary school teachers' use of ICT for student-centered education in Magnolia," International Journal of Education and Development using Information and Communication Technology (IJEDICT), **14**(1), 103-117, 2018.

[75] M. A. Gebremedhin, A. Fenta, "Assessing teachers' perceptions of integrating ICT in teaching-learning process: The case of Adwa College," Journal of Education and Practice, **6**(4), 114-124, 2015.

[76] J-P. Niyigena, Q. Jiang, D. Ziou, R-S. Shaw, A. S. M. T. Hasan, "Modeling the measurements of the determinants of ICT fluency and evolution of digital divide among students in developing countries—East Africa case study," Applied Sciences, **10**(1), 1-27, 2020, doi:10.3390/app10072613.

[77] M. Burns, M. I. Santally, R. Halkhoree, K. R. Sungkur, B. Juggurnath, Y. B. Rajabalee, "Information and communication technologies in secondary education in Sub-Saharan Africa: Policies, Practices, Trends, and Recommendations," 2019, https://www.edu-links.org/sites/default/files/media/file/ICT-in-Secondary-Education.pdf

[78] C. Singhavi, P. Basargekar, "Barriers perceived by teachers for use of information and communication technology (ICT) in the classroom in Maharashtra, India," International Journal of Education and Development using Information and Communication Technology (IJEDICT), **15**(2), 62-78, 2019.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 4, 09-14 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Maintainability Improving Effects such as Insulation Deterioration Diagnosis in Solitary Wave Track Circuit

Takayuki Terada[1,*], Hiroshi Mochizuki[2], Hideo Nakamura[2]

[1]Technology Production Headquarters, Daido Signal Co., Ltd., Tokyo, 105-8650, Japan

[2]College of Science and Technology, Nihon University, Chiba, 274-8501, Japan

ARTICLE INFO

Article history:
Received: 02 June, 2022
Accepted: 01 July, 2022
Online: 18 July, 2022

Keywords:
Track circuit
Solitary wave
Maintainability

ABSTRACT

This paper is an extended version of the journal presented at ICECCME2021. In ICECCME2021, the authors presented that we have developed a solitary wave track circuit (SW-TC), and it is energy-saving compared to existing track circuits. Furthermore, we also explained that it can realize advanced train control at a low cost, equivalent to digital automatic train control. After that, we have conducted research to improve preventive maintenance, which is a problem of existing track circuits, by using SW-TC. In this extended paper, we explain that we can further expand the functions of SW-TC, added new functions such as insulation deterioration diagnosis of the track circuit. With these new functions, the SW-TC can improve reliability, availability, maintainability, and safety. Especially, because of the effect of the insulation deterioration diagnosis function, so railway operators can significantly reduce the time required to identify the cause, when a track circuit failure occurs.

1. Introduction

This paper is an extended version of the journal presented at ICECCME2021 [1]. A track circuit is used as a train detection sensor in railway signals, and consists of a transmitter, receiver, and rails connecting them. In 1872, William Robinson invented a track circuit that detects trains using rails as a circuit. The advent of track circuits enabled automatic signaling systems and contributed to the modernization of train control. Since 1904, various track circuits have been introduced in Japan, and they have contributed to ensuring the safety of railway signals for more than 100 years [2-5]. Since the latter half of the 1990s, research on digital track circuits using microcomputers for track circuits in station premises has been conducted with the aim of saving energy, reducing hardware, and improving maintenance performance, and sending micro-electronics track circuit (SMET) was developed [6-8]. An SMET is capable of time-division processing using digital processing by a microprocessor, can reduce energy and hardware.

On the other hand, in the track circuit between the stations where the distance of the track circuit is long, when processing is performed by one device such as an SMET, the amount of cable is large and the cost is high, so the existing track relay has been used for a long time. Therefore, there were problems such as reduction of cables and energy saving.

To improve these problems, the authors have been developing a new track circuit method, and named it as a solitary wave track circuit (SW-TC). We presented in ICECCME2021 that significant energy savings can be achieved comparing with existing track circuits [1,9]. Furthermore, we explained that it can realize advanced train protection control equivalent to D-ATC [10-18].

After that, we have conducted research to improve preventive maintenance, which is a problem of existing track circuits, by using SW-TC. Currently, railway operators regularly drive inspection vehicles to check the condition of rails in order to prevent track circuit failure. However, there are still many cases of rail breakage and etc., and if adverse conditions overlap, it may lead to dangerous accidents such as derailment. In addition, due to cost issues, inspections using track inspection vehicles are carried out only several times a year, and many small and medium-sized railway operators have not been able to perform sufficient track inspections.

When the track circuit is unexpectedly cut off due to a rail failure, the railway operator walks on the rail and investigates the failure site using a dedicated measuring instrument such as a search

*Corresponding Author: Takayuki Terada, terada@daido-signal.co.jp

https://dx.doi.org/10.25046/aj070402

coil. However, it takes a considerable amount of time to investigate the cause and recover.

To address these problems, we have expanded the functions of SW-TC, added new functions such as diagnosing deterioration of the insulation of the track circuit. Hence, it can realize preventive maintenance of the track circuit. In this paper, we explain that SW-TC can improve not only maintainability but also reliability, availability, and safety by expanding the functions, compared to existing track circuit.

2. Materials and Method

In this chapter, we explain the basic principles of SW-TC.

2.1. Solitary wave

A solitary wave is a waveform obtained by cutting out a part of a continuous signal wave, and refers to a signal having only one wavelength of a certain frequency. Figure 1 shows an example of a solitary wave and its interval (no current), respectively. SW-TC can save energy by transmitting a few wave sources (WSs) within a cycle instead of continuous alternating current (AC) signals as in existing track circuits.

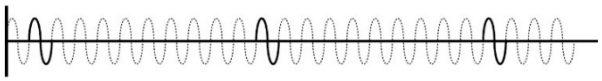


Figure 1: Example of solitary waves

2.2. Example of information allocation

We decided to use the interval between WSs of SW-TC as information. Figure 2 shows an example of information allocation of the SW-TC. Here we prepare 2 WSs and set the WS to a 25Hz sinusoidal waveform for the sake of clarity.

In Figure 2, there are 23 non-current spaces (broken dotted line) in one cycle excluding the 2 WSs (solid line). In the example at the third of Figure 2, there are non-current spaces of 3 waves and 20 waves between the 2 WSs, which is defined as (3:20) or (20: 3). However, (3:20) and (20:3) are considered the same allocation. The small intervals distance between the 2 WSs is defined as signal number (Signal No.), and in the above example, it is defined as Signal No.3. As a result, 11 types of information can be acquired in Figure 2.

If we set the WS to 3 WSs, the information quantity will increase significantly from 11 types to 70 types. In this way, the information quantity can be expanded by the number of WSs, and can be further expanded by increasing the frequency from 25Hz to 50Hz or 100Hz. There is no restriction on the shape of the WS, and it is possible to use a triangle waveform instead of a sinusoidal waveform.

Figure 3 shows an example of the Signal No. transmitted to the SW-TC of each track circuit. First, the state where the train exists on the track circuit and the signal current cannot be received is defined as Signal No.0. Next, the case where a train exists on the front track circuit is defined as Signal No.1, and Signal No. corresponding to the position of the front train is transmitted to the rear track circuit, as shown in Figure 3. Signal No.11 is transmitted to the track circuit that is more than 11 tracks away from the front train.

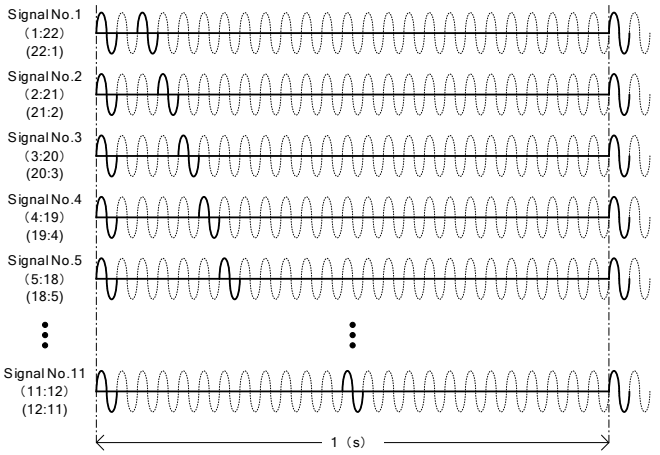


Figure 2: Example of information allocation

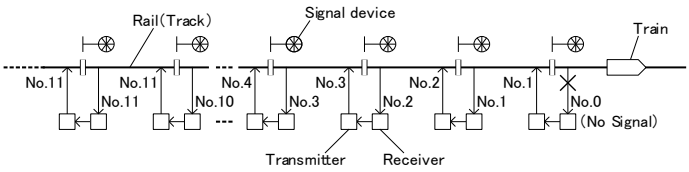


Figure 3: Example of information allocation

2.3. Enhancement by defining solitary wave frame

We have defined the solitary wave frame (SWF) so that SW-TC can be equipped with various functions except for the train position. The structure of the SWF is shown in Figure 4. In this paper, we assume that one SWF is configured at the position of 25WSs.

First, we defined 2 consecutive WSs as a starting element (SE) and placed it at the beginning of the SWF, and set the start position of the frame. If space is secured before and after the SE and one WS is assigned in the remaining positions, 21 types of information can be acquired. If 2 WSs are assigned, the information quantity increases to 190 types, and because of the effect of the SE, the information quantity of SW-TC increases significantly, and the function of SW-TC can be expanded.

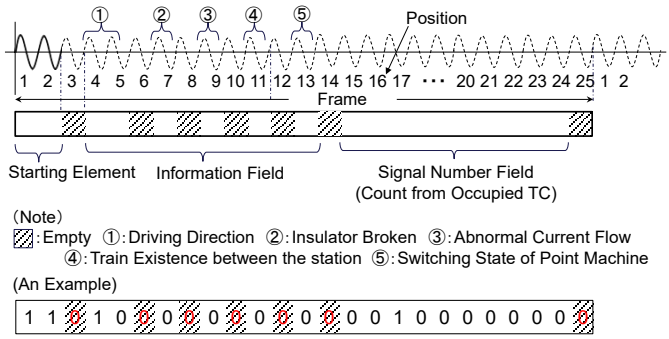


Figure 4: Structure of an SWF

Next, we defined the 4th-13th positions on the SWF as the information field. SW-TC can expand the function by assigning each position of the information field a function, however the details will be explained in the following chapters.

Finally, we defined the remaining 15th-24th positions on the SWF as the signal number field. When the state where the train exists on the track circuit, the 15th-24th positions are assigned 0. When the case where a train exists on the front track circuit, the 15th position is set. When the case where a train exists on the track circuit that is more than 9 tracks away, the 23th position is set. When a train is not exist on the track circuit at the station premises and a route related to that SW-TC is not set , the 24th position is set. Furthermore, in the SWF, there is a restriction that the WS of 2 consecutive waves are assigned only to the SE, and are not assigned in the information field and the signal number field.

3. Current Issue

3.1. Periodic inspection by track inspection car

The track circuit has an important role in detecting trains to ensure safe train operation. Thus, railway operators run inspection cars such as track inspection cars to regularly verify the condition of the tracks [19]. However, despite the inspection, there are still many railroad damages such as rail breakage, and if adverse conditions overlap, it may lead to dangerous accidents such as derailment. Furthermore, since inspection by a dedicated track inspection vehicle is expensive, and there are many small and medium-sized railway operators that cannot perform sufficient track inspections.

3.2. Condition monitoring by track circuit monitor

In the existing track circuit, preventive maintenance has been achieved by introducing a track circuit monitor, a condition monitoring system, etc., and constantly measuring the transmission/reception level of the track circuit. These track monitoring system must work in different weather conditions. Figure 5 shows an example of the screen of the track circuit monitor for the SMET (SMET monitor) of the train detection device for station premises.

The SMET monitor can display the reception level of each track circuit accumulated in the past, the leakage voltage of the adjacent track circuit, etc., which is effective in identifying the cause of a track circuit failure [20]. However, fluctuations in the transmission/reception level vary by track circuit or weather, and the technology for detecting signs of track circuit failure from the tendency of level fluctuations is not completely developed.

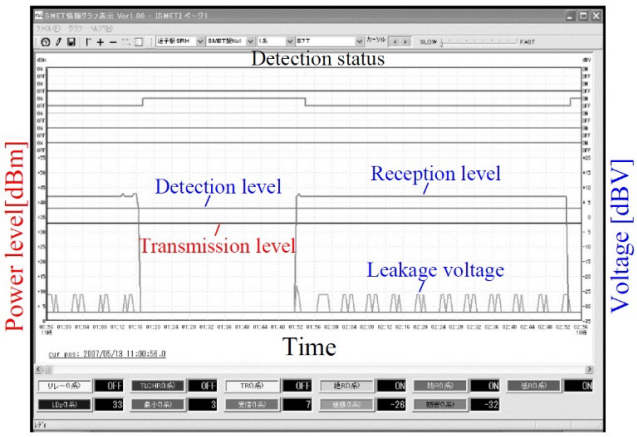


Figure 5: Sample of SMET monitor screen

3.3. Investing the cause with a dedicated measuring instrument

Currently, when an unexpected interruption in the track circuit occurs owing to a failure caused by the rail, in addition to measuring the voltage, a dedicated measuring instrument such as a search coil is used to measure the signal current flowing through the rail by walking on the site. Because there is a high possibility that some type of failure has occurred near the place where the current is changing, the investigation is conducted while grasping the current distribution, but it takes a considerable time to identify the cause and recover the track.

Furthermore, the signal and return currents flow in the track circuit, however, if a large return current flows in the left or right rails, the signal current is affected, and the track circuit becomes unbalanced. When track circuit becomes unbalanced, railway operators investigate the cause using a dedicated measuring instrument such as a return current measuring device. This device measures the return current flowing through the rail with a clamp-type current sensor and measuring device, as shown in Figure 6.

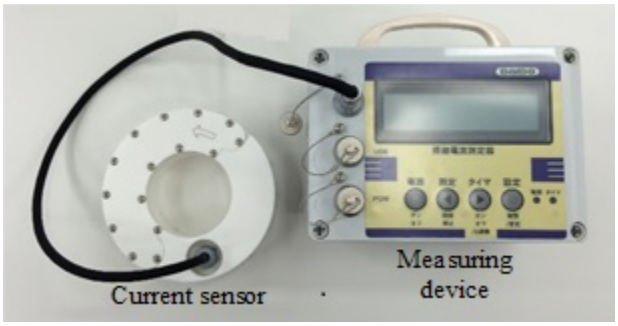


Figure 6: Return current measuring device

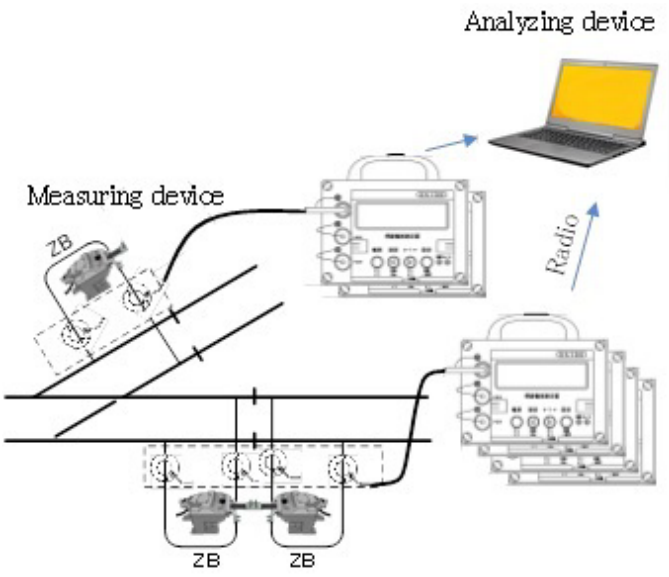


Figure 7: On-site measurement method

Figure 7 shows the on-site measurement method [21]. A current sensor and measuring device are installed on-site to measure the current flowing through the left and right rails. The measurement data are stored in a plurality of measuring devices, and the analyzing devices wirelessly collects the measurement data from the measuring devices in a batch, analyzes the data, and

identifies the cause. However, it takes time and effort to investigate, such as installing devices at the site after a track circuit failure occurs, measuring data for a long time, and removing the devices from the site after measurements are completed.

4. Results and Discussion

In this chapter, we present that SW-TC can diagnose the insulation deterioration of the track circuit and detect the deterioration of the reception level. As a result, we explain that railway operators can identify signs of track circuit failure and improve maintainability using SW-TC. Furthermore, we also explain that SW-TC can also improve reliability, availability, and safety, compared to existing track circuits.

4.1. Insulation deterioration diagnosis of the track circuit

Ordinally, track circuits other than non-insulated track circuits have track insulation inserted to separate their boundaries [22]. When the insulation breakdown occurs at the boundary of the track circuit, if the track circuit current is short-circuited because of the train approaching, the current of the rear track circuit may wrap around and the track relay will operate, preventing the detection of the train. As a countermeasure, an orbital current with the phase inverted is passed through the rear track circuit, and when there is insulation breakdown regardless of the state of the train, the track relay is dropped, and the block signal is stopped. However, it is not easy to identify the faulty part in the field because the track circuit failure has various factors such as the insulation breakdown of the track circuit, failure of the receiving equipment, rail breakage, and failure of the track circuit transmission equipment. Figure 8 shows the track circuit boundary and track circuit transmission/reception equipment (TC-TRE). Ordinally, there is a dead section between the bonding points where the impedance bond is connected across the rail insulation, and no current flows. When insulation breakdown occurs at the boundary of the track circuit, leakage current flows in the dead section at the point where the track insulation is broken, and the faulty part can be detected by performing on-site measurements using a dedicated measuring instrument. However, during that time, the train cannot operate and maintenance personnel needs to move to the site.

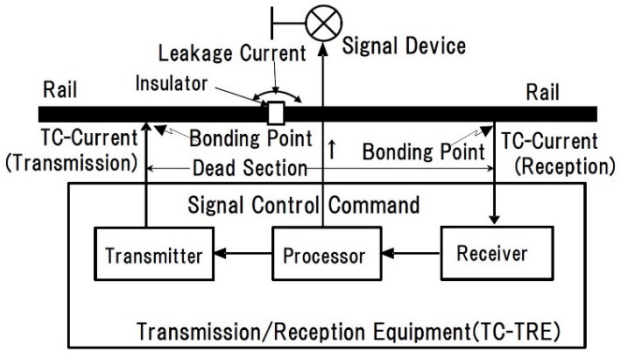


Figure 8: On-site measurement method

As SW-TC assigns SE to the beginning of SWF, if the track circuit insulation is broken, the rear track circuit current wraps around and is superimposed. As a result, the SE generation interval is disturbed in the receiving equipment of the track circuit. At this time, if the current transmission of the transmission equipment to the rear SW-TC is stopped, the wraparound and SE interval disturbance in the SWF are eliminated. The SW-TC can clearly detect the breakdown of the track circuit insulation even during operation by using this mechanism.

Figure 9 shows a deterioration diagnosis method for track insulation of the SW-TC using the principle of the track circuit boundary insulation breakdown detection. Ordinally, the received current input to the TC-TRE has a waveform of [Line No.1], as shown in Figure 9. For the transmission signal to the rear track circuit, an SWF of [Line No.2] is generated, indicating that the Signal No. shifts by one position. In fact, SW-TC transmits SWF at different timings after a certain offset time, as in [Line No.3].

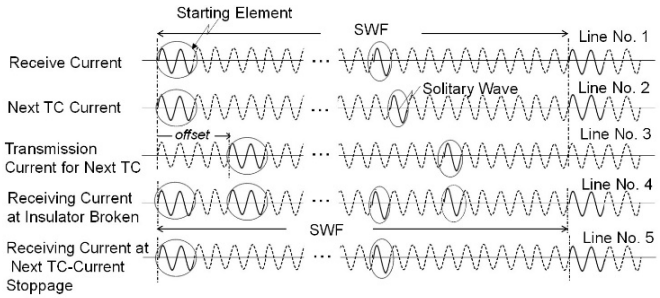


Figure 9: Fault detecting method at a broken insulator

If insulation breakdown occurs at the boundary of the track circuit in this state, SW-TC receives a signal that is a mixture of [Line No.1] and [Line No.3], in which two SEs exist in the SWF. This signal has a waveform similar to that of [Line No.4], and, as explained in section3 of chapter2, TC-TRE can detect insulation breakdown because only SE has two consecutive SWs in the SWF. At this time, if the transmission to the rear track circuit is temporarily stopped, the track circuit current waveform becomes [Line No.5], which is the same as [Line No.1], and the SW-TC can diagnose that the waveform disturbance is due to insulation breakdown. As a result, SW-TC can significantly reduce the time required to identify the insulation breakdown location of the track circuit, and improve maintainability.

Especially in the existing track circuit, railway operators need to measure the received signal level of the track circuit on a regular basis. On the other hand, the SW-TC can automatically performs maintenance measurement for each track circuit and request maintenance, so maintainability of the track circuit can be significantly improved.

4.2. Output of preventive maintenance information

The SW-TC obtains track signals based on the position of WS in SWF; therefore, even when the reception level drops, processing can be continued if the required solitary wave can be detected. However, it is thought that the decrease in the reception level eventually shifts to a state in which the solitary wave itself cannot be detected. Therefore, it is possible to request maintenance at an early stage by outputting the state in which the reception level has dropped as preventive maintenance information. As a result, SW-TC can automatically perform preventive maintenance

measurements at the track circuit level and request maintenance while maintaining its function even when the reception level drops, and maintainability of the track circuit can be improved.

4.3. Transmission method

When the SW-TC detects failure sign information of insulation breakdown of the track circuit or deterioration of reception level, as shown in Figure 4, ② (7position: Insulator Broken) or ③ (9position: Abnormal Current Flow) of the in the SWF is set. With this track signal, it is possible to transmit preventive maintenance information via the crew, for example, by blinking the aspect of the block signal.

If the on-board device is mounted on the train, it can detect insulation breakdown or a decrease in the reception level of the track circuit by decoding the failure bit of the information field in the received SWF. With this information, for example, it is possible to turn on the LED of the on-board device and request maintenance.

4.4. Improved reliability

In the existing track circuit, in order to control block signal of multiple aspects, it was necessary to lay a cable and obtain the information of the forward block signal. On the other hand, SW-TC can obtain the necessary train location information with cableless, and can control block signal of multiple aspects. Furthermore, SW-TC is a simple configuration that eliminates the need for transformers, resistors, phase adjusters, etc., which are required depending on the type of existing track circuit.

With the existing D-ATC, it is necessary to install a large-scale ground equipment to generate digital telegrams and detect trains. On the other hand, SW-TC does not require ground equipment and can realize advanced train protection control equivalent to D-ATC, so significant cost reduction can be expected. Therefore, SW-TC can realize the same functions as the existing track circuit with a simple configuration, and thus the reliability is improved.

4.5. Improved availability and safety

In the field environment of an actual railway signal, the reception level of the track circuit signal fluctuates due to the influence of electric rolling stock current flowing in the track circuit and rainfall. Therefore, in the SMET, which is an existing digital track circuit, a highly reliable measure that follows changes in environmental conditions, such as automatic tracking of threshold levels, have been adopted and are effective. SW-TC judges reception based on the digital sampling data obtained by analog-to-digital (A/D) conversion. Therefore, SW-TC records the received waveform when adjusted at the time of installation as a template, and discriminates WS from the correlation between the received waveform and the waveform of the template. Furthermore, SW-TC can separate from noise by confirming the validity of the number of WS in one cycle.

SW-TC can filter noise judging the track circuit information based on the input waveform itself obtained through A/D conversion, not just the level, and confirming the validity such as the shape check of the isolated wave. As a result, it can improve the noise resistance performance, and reduce dangerous accidents due to disturbing waves, as compared with the existing track circuit method. In this way, the SW-TC can reduce the probability that the train will stop due to track circuit failure due to noise, and improves availability. Furthermore, safety can be improved by reducing dangerous accidents.

5. Conclusion

In this paper, we explained the principle of the SW-TC, and introduced application examples to realize preventive maintenance. The proposed SW-TC scheme can not only realize advanced train protection control equivalent to D-ATC, but also significantly improve maintainability compared to the existing track circuits by detecting failure sign information such as insulation breakdown and deterioration of the reception level. As a result, it can significantly reduce the cause identification and recovery time for railway operators when track circuit failures occur.

Track circuits are a proven technology that has ensured the safety for many years, however, in recent years, new train control signal systems that utilize radio have increased, and the systems tend not to use the track circuits. However, it is difficult for small and medium-sized railway operators to introduce a wireless train control system in terms of costs. The proposed SW-TC method is an extension of the existing track circuit, however, cableless and energy saving compared to the existing track circuit, and can realize advanced train protection control at a low cost. Furthermore, it has excellent reliability, availability, maintainability, and safety, its introduction effect is high, and it is a cheaper and more manageable system for small and medium-sized railway operators. SW-TC has the potential to continue to develop its functions, and we expect that the track circuit can be regenerated by using this method.

Conflict of Interest

The authors declare no conflicts of interest.

Acknowledgment

This research has been funded with a grant from Daido Signal co., Ltd. I would like to thank the three anonymous reviewers for their valuable suggestions for improvement.

References

[1] T. Terada, H. Mochizuki, H. Nakamura, “Development of New Track Circuits for Energy Conservation and Signal Control Innovation," 2021 International Conference on Electrical, Computer, Communications and Mechatronics Enginnering (ICECCME2021), 1-5, 2021, doi:10.1109/ICECCME52200.2021.9591127.

[2] E. Itakura, Kidoukairo, Signal Safety Association of signal technology series of Japan, 1971.

[3] S. Egusa, Shingou niokeru system kaihatsu, Railway and Electrical Engineering of Japan, 2016.

[4] T. Kawano, M. Fukuda, Atarashii tougougata kidoukairo no kaihatsu, Japan Railway Engineer's Association of Japan, 2012.

[5] Y. Hirao, “Safety Technologies on Railway Signalling and Functional Safety," Fundamentals Review of Japan, **7**(2), 124-132, 2013, doi:10.1587/essfr.7.124.

[6] S. Masutani, H. Utsumi, A. Minami, Scanning siki kidoukairo nitsuite,

Railway and Electrical Engineering of Japan, 2011.
[7] A. Minami, Y. Youda, T. Suga, Ressya kenti souti (SMETgata), Daido signal corporation quarterly publication of Japan, 2003.
[8] T. Mizuno, T. Yamamoto, Shingoukiki no shou energy ka no torikumi, Railway and Electrical Engineering of Japan, 2016.
[9] T. Terada, Y. Matsuwaki, T. Fuse, A. Minami, H. Mochizuki, H. Nakamura, "Regeneration of Track Circuit and a Proposal for a New Signal Control System," The transactions of the Institute of Electrical Engineers of Japan.D, **141**(3), 206-211, 2021, doi:10.1541/ieejias.141.206.
[10] N. Terada, Digital ATC, Railway and Electrical Engineering of Japan, 1998.
[11] T. Igarashi, K. Tashiro, Digital ATC system niokeru RAMS kikaku heno taiou, Railway and Electrical Engineering of Japan, 2004.
[12] T. Takashige, kiki bunsangata ATC (digital ATC), Railway and Electrical Engineering of Japan, 1993.
[13] H. Nakamura, Hoan setsubi no sugata to tenbou, Railway and Electrical Engineerig of Japan, 2012.
[14] M. Matsumoto, S. Kitamura, D. Watanabe, Zairaisen digital ATC niokeru assurance gijutsu, Railway and Electrical Engineering of Japan, 2001.
[15] M. Fukuda, H. Arai, Ressya no anzenunkou wo sasaeru gijutsu, Railway Technical Research institute of Railway Research Review of Japan, 2010.
[16] M. Fukuda, Kishikata Yukusue, Railway Technical Research institute of Railway Research Review of Japan, 2012.
[17] H. Arai, N. Terada, Singou Ressyaseigyo gijutsu no Hensen to Doukou, Railway Technical Research institute of Railway Research Review of Japan, 2015.
[18] H. Nakamura, Recent Trends of ICT Application to Railway Operation and Signaling Systems: The Innovation of Railway Signaling Systems, Information Processing Society of Japan, 2014.
[19] M. Matsumoto, "Trend of Sensing Technology on Railway Operation," The transactions of the Institute of Electrical Engineers of Japan.E, **127**(11), 461-466, 2007, doi:10.1541/ieejsmas.127.461.
[20] T. Senoo, Ressya kenti souti (SMETgata) you monitor souti, Daido signal corporation quarterly publication of Japan, 2007.
[21] T. Noguchi, M. Suzuki, T. Kobayashi, The development of new measuring device of return current, Technical review of JR East of Japan, 2015.
[22] A. Taguchi, "Improvement of Border Characteristics of Jointless Track Circuit," The transactions of the Institute of Electrical Engineers of Japan.D, **118**(2), 243-252, 1998, doi:10.1541/ieejias.118.243.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 13-19 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Matching TCP Packets to Detect Stepping-Stone Intrusion using Packet Crossover

Lixin Wang*,[1], Jianhua Yang[1], Austin Lee[1], Peng-Jun Wan[2]

[1]TSYS School of Computer Science, Columbus State University, Columbus, GA 31907, USA

[2]Department of Computer Science, Illinois Institute of Technology, Chicago, IL, USA

ARTICLE INFO

Article history:
Received: 01 August, 2022
Accepted: 18 October, 2022
Online: 13 November, 2022

Keywords:
Stepping-stone
Intrusion Detection
Packet Matching
Packet Crossover
Host-based Detection
Connection Chain

ABSTRACT

Hackers on the Internet often send attacking commands through compromised hosts, called stepping-stones, for the purpose to be hidden behind a long interactive communication session. In a stepping-stone attack, an intruder uses a chain of stepping-stones as relay machines and remotely login these machines using a remote login program such as SSH (secure shell). A great number of detection methods for SSI have been proposed since 1995. Many of these existing detection approaches are either not easy to implement, or not efficient as a great number of packets have to be monitored and analyzed. Some of these detection methods for SSI are even not effective as their capabilities to detect SSI are very limited. In this paper, we propose an effective detection method for SSI by using packet crossover. Packet crossover ratios can be easily computed, and thus our proposed detection method for SSI cannot only be easily implemented, but also efficient. Well-designed network experiments are conducted and the effectiveness of the developed SSID algorithm is verified through the experiments.

1. Introduction

Today hackers usually send attacking commands through compromised hosts in order to be hidden behind a long interactive communication session. These compromised hosts involved in attacks are referred to as stepping-stone ones. While launching a stepping-stone intrusion (SSI) attack, the intruder operates on a local host and sends attacking packets that will be relayed through the intermediate stepping-stones before they reach the final target system.

The TCP protocol was designed in a way that every interactive TCP connection between the attacker host and the final target is independent of one another, even though they are relayed connections. Therefore, the target machine is only able to get information from the last stepping-stone host in the connection chain. That is, it is notoriously hard for the final victim host to obtain information about the geographic region of the origin of the intrusion.

Figure 1 shows a sample of a connection chain that can be exploit to send attacking commands with an SSI. Host 0 in the figure serves as the intruder machine, Host N the final victim machine, and Host 1, Host 2, . . . , Host i-1, Host i, Host i+1, . . . , and Host N -1 serve as the stepping stones for the attack. The purpose of SSI detection (SSID) is to determine whether a host in a network is employed as a stepping-stone one for an attack. In the process of SSID, any intermediate machine within the chain could be chosen as the sensor. A packet sniffing program such as TCPdump or Wireshark must be available on a sensor host. In this figure, Host i is employed as the sensor.

Next, we introduce some important concepts that are needed design detection algorithms for SSI. An incoming connection to Host *i* is defined to be a connection from Host i-1 to Host i. An outgoing connection from Host i is defined to be a connection from Host i to Host i+1. There is possibly an intrusion if an incoming connection of the sensor matches with one of its outgoing connections.

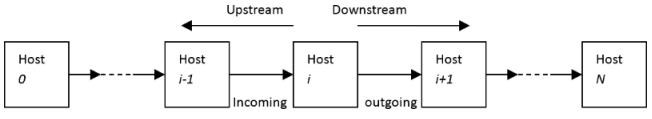


Figure 1. A sample connection chain

*Corresponding Author: Lixin Wang, 4225 University Ave., Columbus, GA 31907, USA. Contact No: 001-706-507-8190. Email: Wang_Lixin@ColumbusState.edu

https://dx.doi.org/10.25046/aj070602

A great number of methods for SSID have been proposed since the first seminar work [1] in 1995. These detection approaches for SSI can be divided into two different types. Compare the incoming connections to a machine with its outgoing connections, and then make a decision to determine whether there is an intrusion based on the comparison. This type of SSI detection approach is referred to as host-based detection [1]-[6]. It is well-known that stepping-stone hosts can be used by some applications to access a remote server legally. Therefore, high false-positive errors could be generated by using host-based detection methods for SSI.

In order to reduce the false-positive errors produced by host-based detection approaches, another category of detection methods was developed to overcome the challenge by counting the number of stepping-stone machines contained in a connection, which is called the length of the connection chain. This category of SSID methods is called connection-chain based or network-based detection [7]-[11], [14], [15]. Typically, there is no need to remotely access a sever via more than three stepping-stones as a lot of unnecessary network traffic will be produced and make the data communication much slower. The threshold number here is three as applications only uses one or two stepping-stone machines to access a remote server legitimately [12], [13].

Many of these existing detection approaches are either not easy to implement, or not efficient as a lot of packets have to be monitored and processed. Some of these SSID methods are even not effective as their detection capabilities for SSI are very limited. There is a need to propose an innovative detection algorithm for SSI that can be easily implemented without having to monitor a lot of TCP packets. Thus, such an algorithm for SSID is efficient in terms of processing time. In this paper, we propose an effective detection method for SSI by using packet crossover ratios. Packet crossover ratios can be easily computed, and thus our proposed detection method for SSI cannot only be easily implemented, but also efficient.

In [16], the authors used the idea of packet crossovers to identify a "long" connection chain. However, this paper made two assumptions (1) there exist packet crossovers in a long connection chain, and (2) a shorter connection chain produces less packet crossovers and a longer connection chain generates more packet crossovers. The conclusion made in [16] was based on these two assumptions. In this paper, we verify the following important statements through network experiments: if the packet crossover ratio of an incoming connection of a sensor host is almost equal to that of an outgoing connection of the sensor, then these two connections are relayed ones, and vice versa.

The remaining of this paper is organized as follows. A literature review for existing approaches for SSID is discussed in section 2. Preliminary knowledge needed in this paper is presented in section 3. In Section 4, an innovative algorithm to match TCP packets using packet crossover is proposed. In Section 5, we design and conduct network experiments to verify the correctness of Proposition 1 described in Section 4. Section 6 gives a conclusion and future research direction for this paper.

2. Literature Reviews

Let us begin our reviews with the existing host-based detection methods that have been proposed for SSI since 1995. The content thumbprint approach to detect SSI was proposed in [1]. This method determines whether a communication session is an intrusion by comparing the content of the packets from the outgoing connection of the sensor host with that of the packets from its incoming connection. It is highly possible that the session is an intrusion if there is a relayed pair between them. However, this approach cannot be used to detect SSI for computer networks with encrypted traffic. To overcome the drawback of this method, a time-thumbprint based approach for SSID was developed in [2]. This approach utilized the timestamps of the monitored packets. This approach for SSID can be used for networks with encrypted traffic as packet timestamps are not encrypted. An example of encrypted traffic is to login to a remote server using SSH.

If the network traffic is encrypted, it is much harder to detect SSI, and even more difficult if a communication session is manipulated by attackers using hacking tools. The packet counting method proposed in [5] was to address such a challenge in detecting SSI through counting the number of packets in both the incoming and outgoing connections. The methods proposed in this paper were to identify stepping-stone connections when the traffic is encrypted and the packet timestamps are jittered. This method also allows an attacker to inject certain amount of meaningless chaff packets into an attacking stream. However, in order for this method to work effectively, both the amount of chaffed meaningless packets and the percentage of the packets with jittered timestamps must be small. Therefore, the capability of the detection approach proposed in [5] is very limited to handle session manipulation by hackers.

Next, we provide a literature review on network-based approaches for SSID that estimate a connection-chain length. In 2002, the authors of [7] developed the 1st network-based approach for SSID to estimate the length of a connection chain. Yung's method calculated the ratio between the Send-Echo RTT and of Send-Ack RTT. The RTTs of an Echo and a Send packet stand for the length of the connection chain from the sensor to the final victim host. On the other hand, an RTT for a Send packet and an Ack packet can only stand for the length from the sensor to its next adjacent machine in the downstream connection sub-chain. Therefore, a high false-negative error was generated by Yung's method in [7] because of the adoption of the acknowledgement packets. The issues of the method proposed in [7] was addressed in the work [8], which is the 2nd network-based approach for SSID proposed in 2004. The detection algorithm proposed in [8] uses the step-function approach to calculate the length of a connection chain in a local area network (LAN). An improvement of the approach in [8] over Yung's one in [7] is that the authors of [8] changed the way to set up the connection chain so that every Send packet can be possibly matched with a corresponding Echo packet.

In a LAN, the step-function method for SSID worked well and reduced both the false positive error and the false negative error, compared to Yung's method proposed in [7]. However, a major drawback of the detection method proposed in [8] is that this method only worked well within a LAN, but wasn't working in the Internet environment. With the context of the Internet, a conservative and greedy packet matching algorithm for SSID was proposed [14]. However, this method in [14] only very few Send packets can be matched with corresponding Echo packets. Therefore, the method in [14] did not work effectively either in the Internet environment.

To address this issue, a clustering and partitioning data mining detection approach for SSI was proposed in [9]. In [9], the packet RTTs were computed by utilizing the clustering and partitioning data mining approach. The packet matching method proposed in this paper is accurate as it went through all the possible Echo packets for every Send packet to be matched. A major issue of the detection method in [9] is that we have to monitor a great number of TCP packets. Therefore, in terms of packet processing time, the detection method proposed in [9] is not efficient.

One of our earlier works [10] addressed the issue of the SSID method proposed in [9] and developed a detection approach via mining network traffic by utilizing the *k*-Means clustering data mining algorithm. This k-Means based approach proposed in [10] does not need to capture and analyze a huge number of packets, and thus it is more efficient than MMD based approach proposed in [9]. But because of the use of the k-Means clustering, the length of a connection chain must be pre-determined for this approach, which makes its performance and capability very limited for SSID. Also, this *k*-Means based detection approach is ineffectively if large fluctuations of the TCP packets exist.

In a recent work [11], we developed an effective network-based SSID approach by calculating the packet crossover ratios. This packet-crossover based method is easy to implement as we can easily compute the packet crossover ratios. With a modification of the k-Means clustering algorithm, [12] proposed an improved algorithm for SSID based detection approach developed in [10] by eliminating some of the packet RTT outliers. However, this paper did not provide any technical analysis regarding whether or not the SSID method is resistant to session manipulations by intruders. Another algorithm proposed for removing packet-RTT outliers is the work [13]. This outlier detection algorithm in [13] can be used to design new approaches to detect SSI. A major drawback of the outlier detection algorithm in [13] it that the accuracy of discovering the RTT outliers is low.

3. Preliminaries

Let us introduce the preliminaries that are needed for our detection algorithm design for SSID in this section.

3.1. Definitions of Send/Echo Packets

Refer to [11] for the definitions of an Echo and a Send packets. For example, when a command is entered on a terminal window in a host running Linux, such as "cd", we assume that the command "cd" is delivered to the remote server in two separate TCP packets: one holding the letter "c" and the other holding the letter "d". Both of these packets are Send ones. When the letter "c" is entered on the user's machine, the packet holding "c" will be delivered to the remote server. Once this packet is received and processed at the server, an Echo packet is sent back to the user, the letter "c" displays on the command line of the user's screen. In such a scenario, the Send packet "c" and the Echo packet "p" are matched. Similarly, a Send packet holding "d" and its corresponding Echo packet holding "d" are also matched.

3.2. Packet Crossover

Packet crossover occurs when a newly Send packet meets an Echo packet of a previous Send packet along the connection chain between a client host and a server host (see Fig. 2). In Fig. 2, we have a connection chain starting from the client (Host 1), to Host 2, then to Host 3, and finally to the server (Host 4), where Host 2 and Host 3 are the stepping-stones in this chain. The red packets S1, S2 and are Send packets, and the green packets E1, E2, and E3 are their Echo packets, respectively. First let us assume that packet crossover is observed at Host 1. In this case, the sequence of these six packets is S1, S2, E1, S3, E2, and E3. Therefore, there are two occurrences of packet crossovers in this case. Now let us observe packet crossover at Host 2. The Send and Echo packets from the connection between Host 2 and Host 3 are monitored. The sequence of these six packets observed at Host 2 is S1, E1, S2, S3, E2, and E3. Thus, only one occurrence of packet crossover is observed in this case.

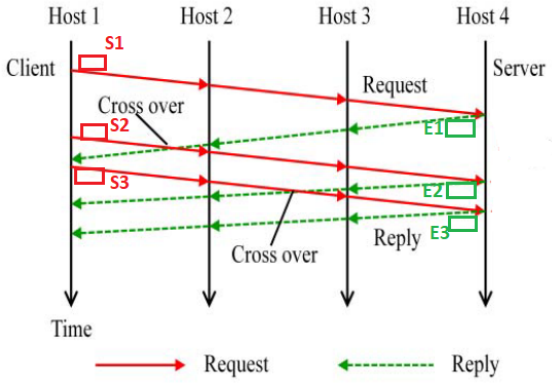


Figure 2: A sample of packet crossover in a connection chain of four hosts.

3.3. The Distribution of Packets' RTTs in a Connection Chain

It is well-known that the number of connections in a connection chain can be represented by utilizing the packet round-trip times computed using the Send packets with their matched Echo packets.

It is well-known that the packet RTTs computed from the TCP packets captured from a connection chain from the attacker host to a target host obey Poisson distribution. This statement was verified in the seminar work [2]. This discovery has been used in the design of SSID methods in the literature. The results of a well-designed network experiment conducted by the authors of [2] is shown in Figure 3. This figure shows that the packet RTTs follow

Poisson distribution. In this figure, the RTT values in micro-second are displayed on the X-axis, and the chances of the occurrences of RTT values are displayed on the Y-axis. In their network setup, the connection chain has a length of four. That is, the chain contains four connections with five machines in total. In their network experiment, μ stands for the mean of all the RTT values. From their observation, μ = 138,500. Since the packet RTTs follow Poison distribution, clearly, most values in the RTTs are very close to the value of μ (see Fig. 3).

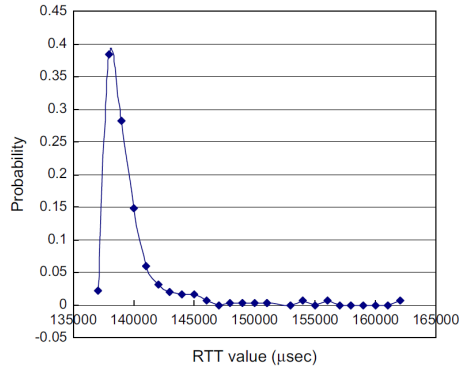


Figure 3: The Packets' RTTs Obey Poisson Distribution

4. Matching TCP Packets Using Packet Crossover

In this section, we first present a proposition that will be used to design our algorithm for SSID. Then we describe an effective algorithm to determine whether a host is used as a stepping-stone by utilizing packet crossover.

First, we present a proposition that will be verified through well-designed network experiments in Section 5. Our detection algorithm design for SSID using packet crossover is based on this proposition.

Proposition 1: If the packet crossover ratio of an incoming connection of a sensor host is almost equal to that of an outgoing connection of the sensor, then these two connections are relayed ones. If the two packet crossover ratios are not close to each other, then these two connections are not relayed.

Next, we describe our proposed host-based detection algorithm for SSI using packet crossover:

1). Pick a host of a network as the sensor host.

2). Adopt Algorithm 1 (Compute Packet Crossover Ratio) in [11] to calculate the packet crossover ratio for every **incoming** connection to the above sensor host.

3). Adopt Algorithm 1 (Compute Packet Crossover Ratio) in [11] to calculate the packet crossover ratio for every **outgoing** connection from the above sensor host.

4). If any of the packet crossover ratios calculated for an incoming connection at Step 2) is almost the same as one of the packet crossover ratios calculated for an outgoing connection at Step 3), then it is highly suspicious that these two connections are relayed ones, and the sensor host is used as a stepping-stone.

5). If none of the packet crossover ratios calculated at Step 2) for incoming connections is close to any of the packet crossover ratios calculated for outgoing connections at Step 3), then it is almost sure that the sensor is not used as a stepping-stone.

The correctness of our above detection algorithm for SSI is clearly asserted according to Proposition 1.

5. Network Experiments

In this section, we design network experiments to verify the correctness of Proposition 1 described in Section 3 by comparing the packet crossover ratios of incoming and outgoing connections. Relayed pairs will typically result in almost equal packet crossover ratios. On the other hand, non-relayed pairs will typically result in dissimilar packet crossover ratios.

To set up our experimental environment, we created two distinct connection chains that shared the same sensor host H3 (see Fig. 4 below). The first connection chain consisted of one local host and four geographically dispersed Amazon AWS servers; all the hosts in the experiment ran Ubuntu Linux operating system. We created a long connection chain by using Secure Shell (SSH) to sequentially connect to each host in the connection chain from the attacker host H1 to the victim host H5 (see Fig. 4). In other words, a single terminal appearing on the attacker host H1 was used to create the entire connection chain by using sequential remote SSH access. From our local PC H1 in Georgia, USA with IP address 168.27.2.105, we remotely accessed host H2 (the first stepping-stone host in the chain), located in Northern Virginia, USA with public IP address 54.226.83.33. We then extended the connection chain by using H2 as a stepping-stone to remotely access the host H3 (our sensor host), located in Northern California, USA with public IP address 54.215.55.31. We then extended the connection chain again by using H3 as a stepping-stone to remotely access host H4 (the last stepping-stone in the chain), located in Tokyo, Japan with public IP address 3.115.8.190. We then extended the connection chain for the final time by using H4 as a stepping-stone to remotely access host H5, located in Central Canada with public IP address 3.99.215.22.

After the first connection chain was established, both the incoming and outgoing connections of the sensor host will be monitored and the packets will be captured using tcpDump at the sensor host (labeled respectively i1 and o1 on in Fig. 4) at H3, the sensor. All data are captured at H3 in this entire network experiment. We entered the following standard Linux commands for about three minutes into a terminal at the attacker host (H1) and captured all packets from the indicated connections at H3:

```
ls
mkdir test
ls
cd test
cd ..
rmdir test
ls
pwd
touch test.txt
```

ls
rm test.txt

We captured ten datasets in total, with each data set comprising two files at the sensor host. After capturing the data, we ran our Packet Crossover Ratio algorithm to calculate the packet crossover ratio observed at H3 from both the incoming and outgoing connections.

We then created a second SSH connection chain that consisted of two local hosts and three geographically dispersed Amazon AWS servers, with each host again running Ubuntu Linux operating system. The sensor host H3 is the only common host shared by these two connection chains. From our local PC H6 in Georgia, USA with IP address 168.27.2.103, we remotely accessed host H7, which had an IP address of 168.27.2.106 and was co-located on the same LAN with our local PC in Georgia. We then extended the connection chain by using H7 as a stepping-stone to remotely access the host H3 (the same sensor host that was used for the first connection chain), located in Northern California, USA with public IP address 54.215.55.31. We then extended the connection chain again by using H3 as a stepping-stone to remotely access host H8 (the last stepping-stone in the chain), located in Frankfurt, Germany with public IP address 3.121.98.162. We then extended the connection chain for the final time by using H8 as a stepping-stone to remotely access host H9, located in London, England with public IP address 18.133.230.186.

After the second connection chain was established, we used tcpDump to capture both the incoming and outgoing connections (labeled i2 and o2 on in Fig. 4) at H3, the sensor. We entered the following standard Linux commands for about three minutes into a terminal at the attacker host (H1) and captured all packets from the indicated connections in the chain:

whoami
uname
groups *username*
who
lscpu
hwclock –verbose
sudo lshw
whatis -h
whatis -l netstat
man netstat (with ten seconds of scrolling)
netstat
help
ifconfig -a
ping google.com
traceroute google.com
sudo cat /etc/shadow
history
!lastCommand

We captured ten datasets in total, with each data set comprising two files at the sensor host. After capturing the data, we ran our Packet Crossover Ratio algorithm to calculate the packet crossover ratio observed in both the incoming and outgoing connections.

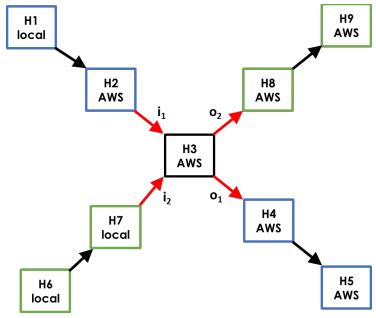


Figure 4: The experimental network setup. Depicts two distinct connection chains passing through the same sensor host H3. C=chain, i=incoming connection, o=outgoing connection, red arrows=connection captured from H3.

We then attempted to use the captured packet crossover ratios to match incoming and outgoing connections. Based on our previous research, we knew that the packet crossover ratios captured at a given sensor for the incoming and outgoing connections of a relayed pair should be close to 1. Therefore, we expected to see a matching of close to 1 for i1 and o1, as well as i2 and o2. Moreover, we expected to see a matching not close to 1 for non-relayed connection pairs such as i1 and o2. In Table 1, CR stands for Crossover Ratio, i1 for incoming connection 1, and o1 for outgoing connection 1. This table compares the CR of i1 to its respective outgoing connection. CR's of relayed pairs should be very similar. Therefore, the incoming connection's CR divided by the outgoing connection's CR should and does equal approx. 1.

Table 1: CR's of relayed pairs i1 and o1 close to 1

Matching Relayed Pair: i1/o1			
Dataset	i1 CR	o1 CR	i1/o1
1	0.3554	0.3554	1
2	0.5202	0.5202	1
3	0.3889	0.3879	0.9974
4	0.3864	0.3864	1
5	0.3431	0.3431	1
6	0.4898	0.4898	1
7	0.2879	0.2879	1
8	0.3279	0.3279	1
9	0.3725	0.3781	1.0150
10	0.3509	0.3509	1

Table 2: CR's of relayed pairs i2 and o2 close to 1

Matching Relayed Pair: i2/o2			
Dataset	i2 CR	o2 CR	i2/o2
1	1.7576	1.7452	0.9930
2	1.7615	1.7626	1.0006
3	1.9298	1.9240	0.9970
4	1.8020	1.8045	1.0013
5	1.8190	1.8212	1.0012
6	1.8362	1.8384	1.0012
7	1.8635	1.8656	1.0011
8	1.8563	1.8483	0.9957
9	1.7700	1.7683	0.9990
10	1.9382	1.9403	1.0011

In Table 2, i2 stands for incoming connection 2, and o2 for outgoing connection 2. This table compares the CR of i2 to its respective outgoing connection o2. CR's of relayed pairs should be very similar. Therefore, the incoming connection's CR divided by the outgoing connection's CR is approx. equal to 1.

Tables 1 and 2 depict a matching between relayed connections, where the packet crossover ratio of a given incoming connection is compared to the packet crossover ratio of its respective outgoing connection. Relayed pairs will typically result in almost equal packet crossover ratios. Therefore, the quotient of the packet crossover ratio of a given incoming connection divided by the packet crossover ratio of its respective outgoing connection should be approximately 1. All ten datasets for both connection chains demonstrate this general rule, with the quotients for connection chains 1 and 2 ranging from 0.9974 to 1.0150 and 0.9930 to 1.0013, respectively.

In Table 3, i1 stands for incoming connection 1, and o2 for outgoing connection 2. This table compares the CR of i1 to the CR of o2. Since these connections do not form a relayed pair, their crossover ratios should not be very similar. Therefore, the incoming connection's CR divided by the outgoing connection's CR should not approximate to 1.

In Table 4, i2 stands for incoming connection 2, and o1 for outgoing connection 1. This table compares the CR of i2 to the CR of o1. Since these connections do not form a relayed pair, their crossover ratios should not be very similar. Therefore, the incoming connection's CR divided by the outgoing connection's CR should not approximate to 1.

Table 3: CR's of non-relayed pairs i1 and o2

Matching Non-relayed Pair: i1/o2			
Dataset	i1 CR	o2 CR	i1/o2
1	0.3554	1.7452	4.9111
2	0.5202	1.7626	3.3882
3	0.3889	1.9240	4.9475
4	0.3864	1.8045	4.6703
5	0.3431	1.8212	5.3086
6	0.4898	1.8384	3.7536
7	0.2879	1.8656	6.4797
8	0.3279	1.8483	5.6366
9	0.3725	1.7683	4.7473
10	0.3509	1.9403	5.5299

Table 4: CR's of non-relayed pairs o1 and i2.

Matching Non-relayed Pair: i2/o1			
Dataset	i2 CR	o1 CR	i2/o1
1	1.7576	0.3554	0.2022
2	1.7615	0.5202	0.2953
3	1.9298	0.3879	0.2010
4	1.8020	0.3864	0.2144
5	1.8190	0.3431	0.1886
6	1.8362	0.4898	0.2667
7	1.8635	0.2879	0.1545
8	1.8563	0.3279	0.1767
9	1.7700	0.3781	0.2136
10	1.9382	0.3509	0.1810

Tables 3 and 4 depict a matching between non-relayed connections, where the packet crossover ratio of a given incoming connection is compared to the packet crossover ratio of another unrelated outgoing connection. Non-relayed pairs will typically

result in dissimilar packet crossover ratios. Therefore, the quotient of the packet crossover ratio of a given incoming connection divided by the packet crossover ratio of an unrelated outgoing connection should typically not be close to 1. All ten datasets demonstrate this general rule, with the quotients for i1/o2 and i2/o1 ranging from 3.3882 to 6.4797 and 0.1545 to 0.2953, respectively.

We can clearly see that matched connection pairs are very similar, and non-matched pairs are dissimilar. With a true-positive threshold of 0.99-1.02, 100% of the matched pairs would be recognized as a relayed pair. Furthermore, 100% of the non-relayed pairs would be identified correctly.

6. Conclusion

Many known SSID methods are either not easy to implement, or not efficient as a large number of packets have to be monitored and analyzed. Some of them are even not effective as their capabilities to detect SSI are very limited. In this paper, we proposed an effective SSID method by using packet crossover ratios. Packet crossover ratios can be easily computed, and thus the SSID method developed in this paper cannot only be easily implemented, but also efficient in terms of packet processing time. Through well-designed network experiments, we verified that for a given sensor, if an outgoing connection and an incoming connection are detected to be relayed with each other, then the packet crossover ratios obtained from these two connections should be very close, and vice versa.

As for future research direction, one may revise and improve our proposed detection method for SSI so that it will be resistant to session manipulation by intruders using hacking techniques such as chaff-perturbation of meaningless packets or time jittering.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work of Drs. Lixin Wang and Jianhua Yang is supported by National Security Agency (NSA) NCAE-C research grant H98230-20-1-0293 with Columbus State University, Columbus GA 31907, USA.

References

[1] S. Staniford-Chen, and L. T. Heberlein, "Holding Intruders Accountable on the Internet," Proc. IEEE Symposium on Security and Privacy, Oakland, CA, 39-49, 1995, DOI: 10.1109/SECPRI.1995.398921.

[2] Y. Zhang, and V. Paxson, "Detecting Stepping-Stones," Proc. of the 9th USENIX Security Symposium, Denver, CO, 67-81, August 2000, doi: https://dl.acm.org/doi/10.5555/1251306.1251319.

[3] D. Donoho, A. Flesia, U. Shankar, V. Paxson, J. Coit, and S. Staniford, "Multiscale stepping-stone detection: Detecting pairs of jittered interactive streams by exploiting maximum tolerable delay," in 5th International Symposium on Recent Advances in Intrusion Detection, Lecture Notes in Computer Science, 2516, 2002, DOI:10.1007/3-540-36084-0_2.

[4] T. He and L. Tong, "Detecting Stepping-stone Traffic in Chaff: Fundamental Limits and Robust Algorithms," the 9th International Symposium on Recent Advances in Intrusion Detection (RAID 2006), April 2006.

[5] T. He, L. Tong, "Detecting encrypted stepping-stone connections," In: Proceedings of IEEE Transaction on signal processing, **55**(5), 1612-1623, 2007, DOI: 10.1109/TSP.2006.890881.

[6] A. Blum, D. Song, And S. Venkataraman, "Detection of Interactive Stepping-Stones: Algorithms and Confidence Bounds", Proceedings of International Symposium on Recent Advance in Intrusion Detection (RAID), Sophia Antipolis, France, 20-35, September 2004, DOI:10.1007/978-3-540-30143-1_14.

[7] K. H. Yung, "Detecting Long Connecting Chains of Interactive Terminal Sessions," Proc. of International Symposium on Recent Advance in Intrusion Detection (RAID), Zurich, Switzerland, 1-16, October 2002, https://doi.org/10.1007/3-540-36084-0_1.

[8] J. Yang, S.-H. S. Huang, "A Real-Time Algorithm to Detect Long Connection Chains of Interactive Terminal Sessions," Proceedings of 3rd ACM International Conference on Information Security (Infosecu'04), Shanghai, China, 198-203, November 2004, DOI:10.1145/1046290.1046331; Corpus ID: 18061584.

[9] J. Yang, and S. S.-H. Huang, "Mining TCP/IP Packets to Detect Stepping-Stone Intrusion", Journal of Computers and Security, Elsevier Ltd., **26**, 479-484, December 2007, doi:10.1016/j.cose.2007.07.001.

[10] L. Wang, J. Yang, X. Xu, and P.-J. Wan, "Mining Network Traffic with the k-Means Clustering Algorithm for Stepping-stone Intrusion Detection", Wireless Communications and Mobile Computing, **2021**, Article ID 6632671, 2021, https://doi.org/10.1155/2021/6632671.

[11] L. Wang, J. Yang, and A. Lee, "An Effective Approach for Stepping-Stone Intrusion Detection Using Packet Crossover," the 23rd World Conference on Information Security Applications (WISA), August 24-26, 2022, DOI: 10.26599/TST.2021.9010041.

[12] L. Wang, J. Yang, M. Workman, and P.-J. Wan, "Effective algorithms to detect stepping-stone intrusion by removing outliers of packet RTTs," Tsinghua Science and. Technology. **2021**(27), 432-442, https://doi.org/10.26599/tst.2021.90100432-4421.

[13] O. Alghushairy, R. Alsini, X. Ma, and T. Soule, "Improving the Efficiency of Genetic-Based Incremental Local Outlier Factor Algorithm for Network Intrusion Detection," Advances in Artificial Intelligence and Applied Cognitive Computing. In Transactions on Computational Science and Computational Intelligence; Arabnia, H.R., Ferens, K., Fuente, D., Kozerenko, E.B., Olivas, J.A., Tinetti, F.G., Eds.; Springer, Cham: New York, NY, USA. **1**, 1011–1027, 2021, http://dx.doi.org/10.1007/978-3-030-70296-0_81.

[14] J. Yang, S.–H. S. Huang, "Matching TCP Packets and Its Application to the Detection of Long Connection Chains," Proceedings of 19th IEEE International Conference on Advanced Information Networking and Applications (AINA 2005), Taipei, Taiwan, China, 1005-1010, March 2005, 10.1109/AINA.2005.240.

[15] J. Yang, B. Lee, and S.S.-H Huang, "Monitoring Network Traffic to Detect Stepping-Stone Intrusion," the Proceedings of 22nd IEEE International Conference on Advanced Information Networking and Applications (AINA 2008), Okinawa, Japan, 56-61, March 2008, DOI: 10.1109/WAINA.2008.30.

[16] S. S.-H Huang, H. Zhang, and M. Phay, "Detecting Stepping-stone intruders by identifying crossover packets in SSH connections", the Proceedings of 30th IEEE International Conference on Advanced Information Networking and Applications, Fukuoka, Japan, 1043-1050, March 2016, DOI: 10.1109/AINA.2016.132.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 2, 14-23 (2023)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Multiple Criteria Decision-making: Risk Analyses for the Soft Target

Dora Kotkova*, Lukas Kralik, Lukas Kotek, Jan Valouch

Department of Security Engineering, Tomas Bata University in Zlín, Zlín, 760 05, Czech Republic

ARTICLE INFO

Article history:
Received: 31 August, 2022
Accepted: 29 January, 2023
Online: 11 March, 2023

Keywords:
Soft targets
Risk analysis
TOPSIS

ABSTRACT

This article focuses on risk analysis using a multi-criteria decision-making method. Due to many performed risk analyses for soft targets, we are constantly trying to find new methods for objective risk assessment. Many risk analyses are subjective, which is a problem when planning security measures and comparing results (different events, objects, places, etc.). In this text, we present our case study, which deals with the use of fuzzy TOPSIS. As a reference object, we have chosen one of the specific categories of soft targets – cultural events. The goal was to find the location most at risk of violent attacks on a selected cultural event - a music concert. We then established cooperation with three experts. The completed data in the risk analysis was then compared with practice. The selected fuzzy TOPSIS method was chosen as presumably more objective. Our hypothesis was confirmed. The results were objective and consistent with practical experiences.

1. Introduction

The article aims to present our research, which we have worked on for several years [1-12]. The goal is to research the protection of soft targets and find new methods to strengthen security. In this article, we focus on one necessary step, risk analysis. The biggest problem is selecting the appropriate method and minimalizing subjective view. Currently, many risk analysis methods exist, but none are specifically designed for soft targets. Our effort is to explore these methods and choose one or develop a new one that will provide objective results.

This article extends the contribution from The IEEE International Carnahan Conference on Security Technology 2021, where we presented an essential and necessary introduction and case study with the applied Fuzzy TOPSIS method [1]. In this extended version, we provide additional information on why we selected Fuzzy TOPSIS and a more detailed risk analysis of the whole workflow. Also, the introduction offers more detailed information about previous research and project on soft targets protection [1-12].

Our research is mainly concerned with the area of soft target protection. *"The term "Soft Targets" can be referred to those objects (buildings), (open) spaces, or events characterized by the accumulation of many people, the absence or low level of security measures against violent assaults, and their omission among critical infrastructure and hard target objects"* [13].

The protection of soft targets is a relatively new phenomenon in the geopolitical space of Europe. This issue became relevant after the events in 2014 when the so-called Islamic State declared a caliphate and began carrying out terrorist attacks worldwide [13], [14]. As the first figure (Fig. 1) shows, the number of terrorist attacks almost doubled. The data is obtained from our database created for our research[7], [9], [11], where violent attacks against soft targets are recorded from Europe and parts of Russia and Turkey. The existence and updating of the database are crucial for our research because we need to register this information and add details such as the location of the attack, time, modus operandi (Fig. 2), type of soft target (Fig. 3), and other information about the attack [14], [15].

This paper describes a significant part of our research. It is focused on security measures and, above all, on risk analysis. Nowadays, there are many risk analysis methods but just a few of them suits soft targets. The central lack of those methods is objectiveness. Some complex methods, such as Failure Mode and Effects Analysis (FMEA), are claimed as objective methods. But the result of this method depends on one evaluator's opinion [16]. The section 3 specifies some methods that are mostly used in security issues. Following section 4 goes into more detail about methods verified and tested in our research.

We asked three security managers to participate in our case study by finding the most exposed and weakest point in protecting soft targets at cultural events. Our previous research has shown the utility of multi-criteria decision-making (MCDM) in identifying

*Corresponding Author: Dora Kotkova, kotkova@utb.cz

soft targets. This case study aimed to verify the hypothesis that some methods of MCDM should be used for risk analysis, especially risk assessment. One of the significant parts of risk analysis is proposing security measures and making a plan of necessary activities; however, we focused just on the part of risk assessment and compared the final risk ranking from MCDM and typical risk analysis.

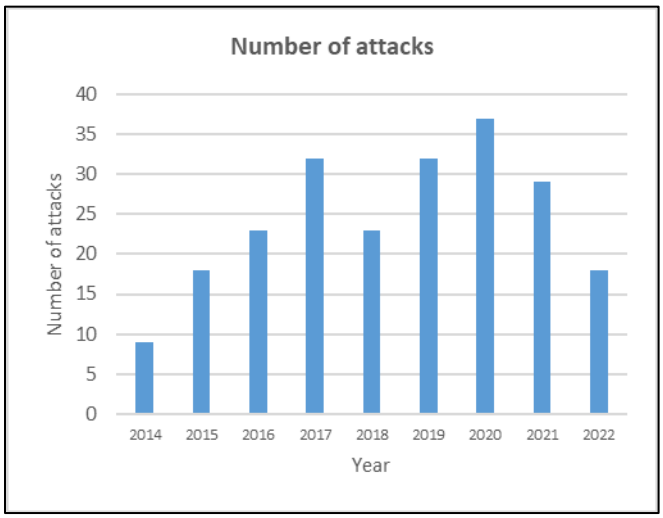


Figure 1: Number of terrorist attacks from 2014 to 2022 [15]

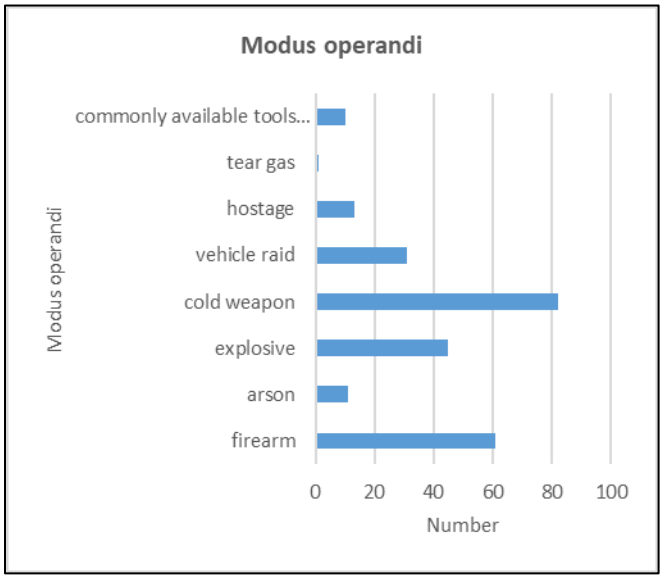


Figure 2: Modus operandi of terrorist attacks from 2014 to 2022 [15]

For this purpose, we selected some MCDM methods described deeply in section 4. Lately, we added the Fuzzy TOPSIS method to the list because this method perfectly fits our problem of objectives. Based on information from several publications [17 – 19], we decided to use the standard fuzzy TOPSIS version due to the existence of many hybrid approaches [20], [21], or updated methods [22], [23]. Some efforts are known to use Fuzzy TOPSIS for risk assessment [24], [25]. Classic fuzzy TOPSIS is only a segment of the developed methodology.

Soft targets has many categories and each category has own specifications. Cultural events is a category that we selected for our research and case study. As Fig. 3 depicts, cultural events are not targeted mostly, however they are very diverse. The quality of security measures that are highly variable, is the main reason for our choise.

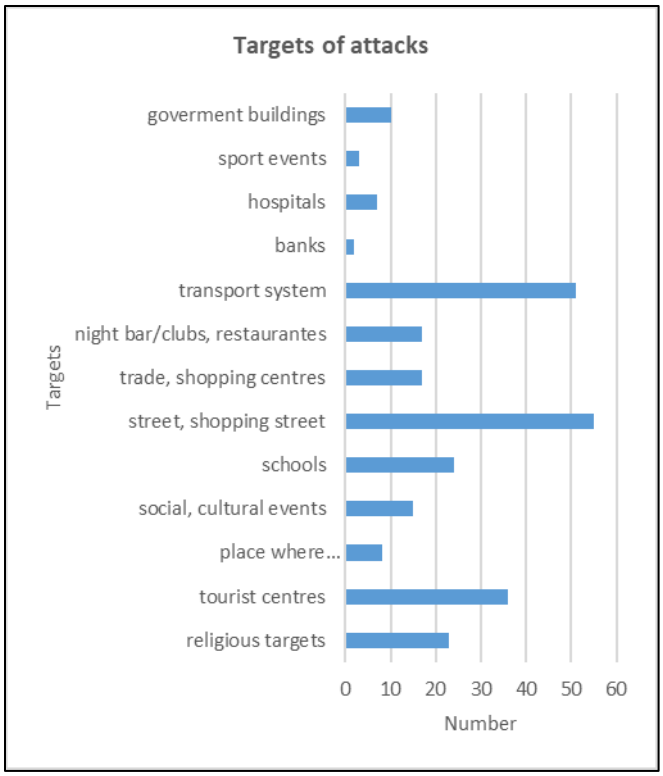


Figure 3: Targets of terrorist attacks from 2014 to 2022 [15]

2. Cultural events

Cultural events are one of the soft targets categories. *"Cultural events can be defined based on a combination of two words - culture and event or based on a combination of their defining concept. By combining these formulations, the following definition can be reached: cultural events are pre-planned events with a clearly defined place and time, the main goal of which is the presentation of spiritual and material values created by man."* [26]

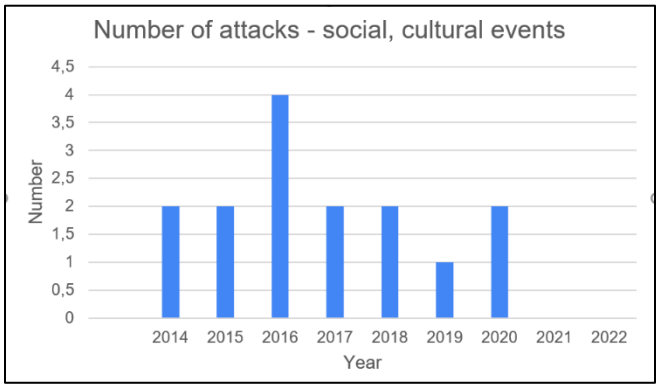


Figure 4: Number of terrorist attacks on social and cultural events from 2014 to 2022 [15]

Each mass event has its own specification, conditions and needs. Safety is a vital part of these mass events or gatherings. Differentiation of essential characteristics is absolutely necessary

to establish proper safety. Every successful terrorist attack on mass gathering may take many causalities. This is the reason why all mass gatherings must be well protected. The following figure (Fig. 4) shows that the average number of attacks is 2 and a half per year with a peak in 2016 [15].

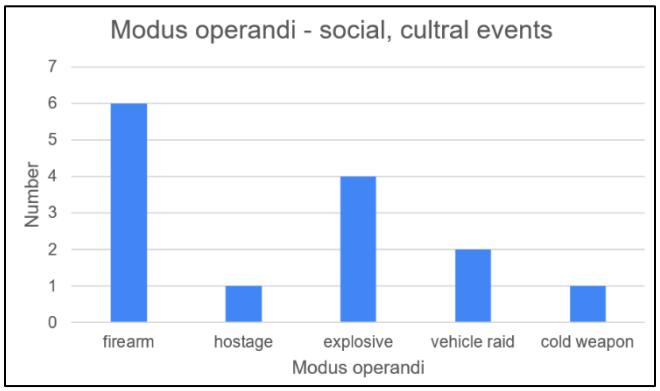


Figure 5: Modus operandi of terrorist attacks on social and cultural events from 2014 to 2022 [15]

3. Risk management

One of the integral parts of security planning is handling threats and their probability. We need to know what threats are in our facility. The objective of security planning is to take measures that lower the probability of emerging threats. The probability is also labeled as a risk.

Risk management (RM) is an endless process that subsists of three steps (fig. 6). Each step is described in more detail in the following sub-sections.

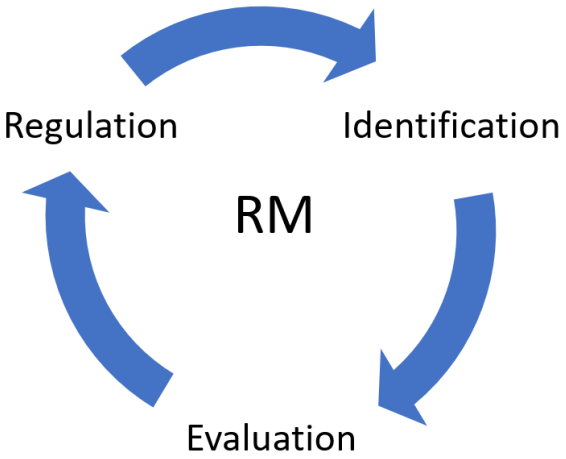


Figure 6: Risk Management steps [15]

3.1. Risk identification

Risk identification is ensured through risk analysis. The risk identification step usually includes the following activities:

1) Identification of sources of risk – this part includes:
   a) Determination of initial data – data on the analyzed entity, description of operation, location, etc.
   b) Identification of assets - definition and determination of the value of the assets owned by the subject;
2) Identification of threats – identification of events and actions that can negatively affect the value of assets;
3) Identification of vulnerabilities – every asset has its vulnerability or weakness to an identified threat. The level of this vulnerability affects the impact when a threat affects a protected asset;
4) Determining and assessment of impact.
5) Determination of risk - determination of the probability of the occurrence of threats and the degree of vulnerability of the subject [27]

There are three basic groups of risk analysis. The first one contains qualitative methods such as tree analysis – Event Tree Analysis (ETA), or Fault Tree Analysis (FTA). One of the most used and elemental methods is Safety/Security review. The group of quantitative methods stands on the other side, Failure Mode and Effect Analysis. (FMEA) is one of the well-known methods in this category. The last category is the combination of the two previous. It is a group of semi-quantitative methods that combine experience and numerical expression of risk [15].

3.2. Risk evaluation

Risk evaluation is an activity determined to enumerate risks and assign them to the correct category. The risks scale from the lowest negligible to the highest critical (Fig. 7). When planning security measures, we start with the necessary risk and need to spend the most funds in this category.

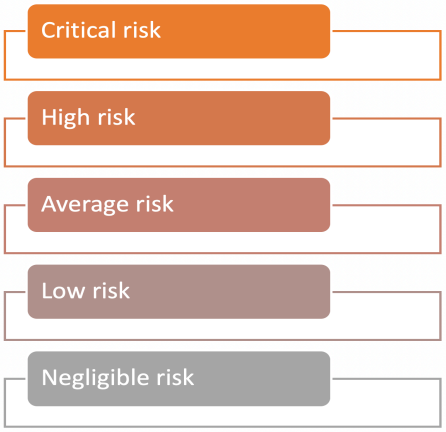


Figure 7: Risk scale

3.3. Risk regulation

The last step in risk management is risk regulation. This step includes six different approaches. Each approach has its own positives and negatives. It is very complicated to say which one is the best. It depends on the point of view. The possible approaches are:

- Ignoring the risk – conscious or unconscious
- Acceptance of the risk – this approach covers also issues with residual risk.
- Reduction of the risk – this is the most logical approach that implements corresponding security measures or reduces the severity of the impact.

- Risk avoidance
- Transfer of risk – transferring risk to another entity (e.g., outsourcing)
- Risk monitoring – the risk is not static, but it is variable and might change over time.

4. Methods

MCDM is a very complex field and provides many different methods and procedures. The most well-known methods such as the Analytic Hierarchy Process (AHP) [28], Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [29], [30] or Višekriterijumsko Kompromisno Rangiranje (VIKOR) [30] has been selected for our research.

4.1. AHP

Tasks, the number of elements (variants and criteria), and the complexity in more complicated decision-making increase exceptionally. Due to this, the decision-maker has a problem getting his bearings in the task. Therefore, in the 1970s, Professor Thomas L. Saaty created the AHP method, which he and his colleagues developed into a practical tool for solving these complicated tasks. [28]

AHP can be described as a method of breaking down a complex unstructured situation into simpler parts – the establishment of a hierarchical system (Fig. 8). The following part uses the subjective assessment of pairwise comparisons, which assigns numerical values to individual components that express their relative importance. The subsequent synthesis of these evaluations determines the element with the highest priority.

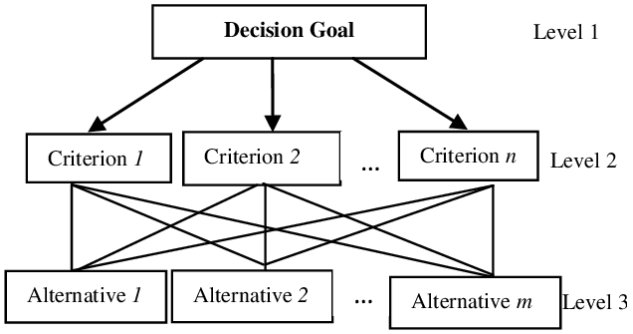


Figure 8: Structure (hierarchy) of AHP [31]

4.2. TOPSIS

The TOPSIS method, according to R.R. Venkata [32], was developed by Hwang and Yoon and is characterized by the fact that it requires only a minimum number of inputs from the user and also that the outputs are very understandable. The only subjective parameters are the weights of the criteria. The TOPSIS method is based on the idea that the optimal solution is the one that has the shortest distance from the ideal solution and is also the furthest from the solution that is negative [29], [30].

The method is based on five computational steps:

1. obtaining performance evaluation of alternatives on various criteria,
2. normalization of assessment,
3. normalized scores are assessed,
4. the distance between the ideal and negative solutions is calculated
5. the ideal solution selected based on ratios of calculated distances [29], [30]

4.3. VIKOR

The VIKOR method is another method that searches for the best solution on a set of alternatives according to the distance from the ideal solution [33]. This method was developed for multi-criteria optimization in complex systems and focused on evaluation and selection in the presence of conflicting criteria. Each alternative is evaluated according to the multi-criteria function, and the resulting compromise solution is obtained based on the proximity to the ideal alternative. The evaluation is obtained using five steps [30], [34].

1. The alternatives are labeled as $x1$, $x2$, … , xm; where m is number of alternatives.
2. Determination of the maximum and minimum values of all criterion functions.
3. Calculation of the utility and regret measure of the respective alternative.
4. Calculation of performance evaluation of alternatives.
5. Sorting of alternatives according to performance evaluation. A lower rating value represents a better alternative [30]

4.4. Comparison and selection

The results of the above methods were compared against results from semi-quantitive risk analysis PIE. We carefully selected security issues for risk analysis and compared only cases with exactly ten threats (criteria). We asked highly renewed experts for security and safety. Each of them made a PIE analysis, and then we made a threat list ranked according to the mean value of obtained results. This list served as a "standard" for comparing our selected methods. The method with the lowest ranking difference against our PIE "standard" was selected for further research. As shown in tab. I. Fuzzy TOPSIS (FTOPSIS) presents the best results. FTOPSIS was added in the progress of research because it allows the processing of opinion of more evaluators.

Table 1: Rank Comparison

	Threat 1	**Threat 2**	**Threat 3**	**Threat 4**	**Threat 5**
TOPSIS	0.20	0.11	0.20	0.19	0.18
AHP	0.28	0.15	0.19	0.23	0.27
VIKOR	0.18	0.12	0.19	0.18	0.23
FTOPSIS	0.08	0.14	0.10	0.12	0.15
	Threat 6	**Threat 7**	**Threat 8**	**Threat 9**	**Threat 10**
TOPSIS	0.11	0.14	0.19	0.12	0.15
AHP	0.28	0.18	0.21	0.30	0.26
VIKOR	0.11	0.15	0.10	0.15	0.12
FTOPSIS	0.10	0.13	0.11	0.14	0.13

Every change of rank by one position is reflected in comparison by value 0.1. The average value for comparison of all cases is in the above table (Tab. I.)

The most significant similarity with PIE (Tab. II.) shows FTOPSIS had similar threats rank for six threats. Similar methods, TOPSIS and VIKOR, has almost identical results. (Tab. I. and Tab. II.)

Table 2: Overal Similarity

	Similarity
TOPSIS	2/10
AHP	0/10
VIKOR	2/10
Fuzzy TOPSIS	6/10

5. Fuzzy TOPSIS

FTOPSIS is a modified standard TOPSIS, which belongs to the most used techniques for MCDM. It consists of adding fuzzy logic that allows the reflection opinion of all decision-makers. This technique utilizes two ideal solutions. Those solutions are called Fuzzy Positive Ideal Solution (FPIS) and Fuzzy Negative Ideal Solution (FNIS) [29]. The aim is to find a distance from both of these ideal solutions. The longest distance from FNIS and the shortest distance from the FPIS is the best result. It is calculated as a geometric distance from these solutions [16].

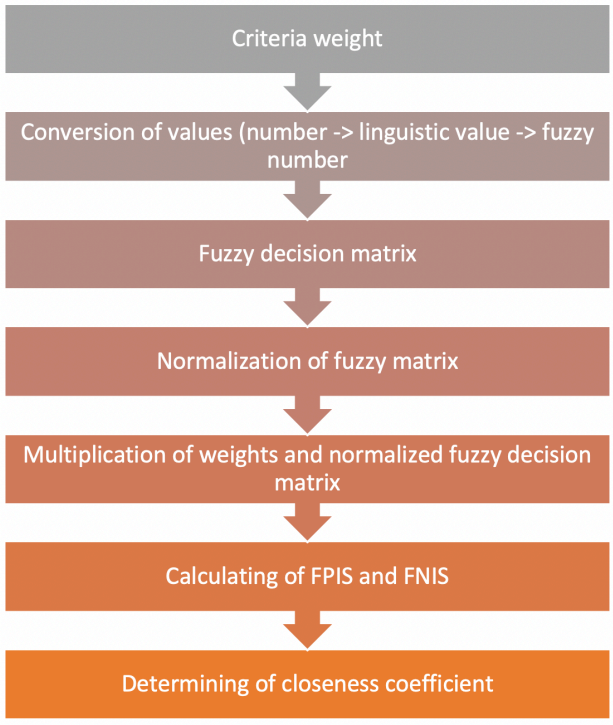


Figure 9:Seven steps of FTOPSIS [1]

Complete calculation and ordering of alternatives according to distance have seven steps (Fig. 9) [1], [16-18].

Table 3: Conversion table [1]

Numerical value	Linguistic value	Fuzzy number
1	Very low	{1, 1, 3}
2	Low	{1, 3, 5}
3	Average	{3, 5, 7}
4	High	{5, 7, 9}
5	Very high	{7, 7, 9}

Usage of linguistic values showed as a not appropriate. Each linguistic value was represented by numerical value which made it easier for our decision-makers. Consequently, we converted it into fuzzy numbers to obtain correct weights for further calculation. (Tab. III.).

Converted fuzzy numbers are used to build a corresponding fuzzy matrix for each decision-maker (three matrices, in our case). Those matrices are merged to the united fuzzy matrix. Every triangular fuzzy numbers is represented by three values ($F=\{a;b;c\}$). Calculation of these values are based on following formulas (1):

$$a_{ij} = min_k\{a_{ij}^k\};\ b_{ij} = \frac{1}{K}\sum_{k=1}^{K} b_{ij}^k; c_{ij} = max_k\{c_{ij}^k\} \quad (1)$$

The uniform matrix is built from a combination of decision matrices and the above formulas. Next step is the normalization of the uniform matrix. Normalization is one of the essential activities for almost all MCDMs. The TOPSIS and also its fuzzy modification work with two types of criteria. The first group brings benefits (eg. increasing efficiency). The second group has a negative influence on the decision maker and this group is called as a cost criteria (eg. salary). Each group has slightly different formula for normalization. Benefits criteria (2) looks for maximal value and cost criteria (3) for minimal value :

$$\tilde{r_{ij}} = \left(\frac{a_{ij}}{c_j^+}, \frac{b_{ij}}{c_j^+}, \frac{c_{ij}}{c_j^+}\right); . c_j^+ = max_i\{c_{ij}\} \quad (2)$$

$$\tilde{r_{ij}} = \left(\frac{a_j^-}{c_{ij}}, \frac{a_j^-}{b_{ij}}, \frac{a_j^-}{a_{ij}}\right);\ a_j^- = min_i\{a_{ij}\} \quad (3)$$

The normalized matrix must be supplemented by criteria weight. Normalized values are simply multiplicated by reciprocal weight values of criterion (4).

$$\tilde{A_1} \otimes \tilde{A_2} = (a_1, b_1, c_1) \otimes (a_2, b_2, c_2) = (a_1 * a_2, b_1 * b_2, c_1 * c_2) \quad (4)$$

Multiplication results are used to find the ideal solution. There are two possible solutions: a positive solution with the "best" values and a negative solution with the worst values. Because this is fuzzy modification than solutions are called as fuzzy positive/negative solutions (FPIS (5)/ FNIS (6)), Calculation of these solution is given by:

$$A^+ = (\tilde{v_1^+}, \tilde{v_2^+}, \ldots \tilde{v_n^+});\ \tilde{v_j^+} = max_i\{v_{ij3}\} \quad (5)$$

$$A^- = (\tilde{v_1^-}, \tilde{v_2^-}, \ldots \tilde{v_n^-});\ \tilde{v_j^-} = min_i\{v_{ij1}\} \quad (6)$$

However, it is the best or the worst situation that probably never occurs. Each involved alternative has its own solution and the distance, which is between FPIS, FNIS and this solution, (7) describes the quality of the involved alternative. The distance is given as:

$$the\ d(\tilde{x}, \tilde{y}) = \sqrt{\frac{1}{3}[(a_1 - a_2)^2 + (b_1 - b_2)^2 + (c_1 - c_2)^2]} \quad (7)$$

The closeness coefficient (8) is the final step to ordering all involved alternatives.

$$CC_i = \frac{d_i^-}{d_i^- + d_i^+} \quad (8)$$

6. Case study

We have focused on applying the TOPSIS method to one cultural event in the case study. The basic information about the event:

- Specialization: music metal concert - group Arakain
- Where: Masters of Rock Café in Zlín, Czech Republic
- Frequency: once a year (various dates)
- Number of visitors: 400 - 500 people

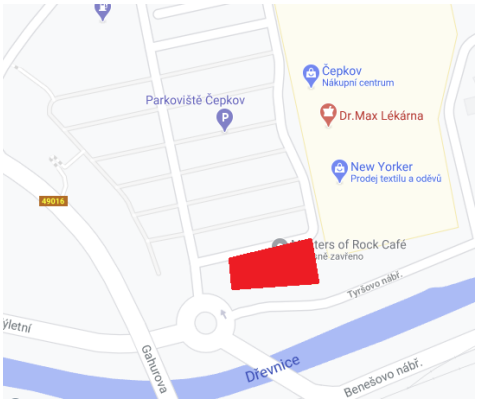


Figure 9: Localization of the Master of Rock Café. [35]



Figure 10: The main entrance of the Mater of Rock Café. [35]

6.1. Risk management

Assets are determined based on a security inspection of the given location, where we analyzed what is crucial for the owner of the building, whether from the point of view of finances, operations, or maintaining continuity. There are assets within the selected object:

- People
- Property of the club, visitors, performers
- Building
- Continuity
- Money
- Etc. [15].

The next step is to identify the threats. Several ways could do this. One possible way is again an on-site security inspection where a security professional has enough experience in the required field and can identify the most problems that may arise. The second way is to start from the threat catalog. A threat catalog is a document or, more accurately, a list of generic and most common threats. This catalog is made by some security experts to provide a tool that may help identify threats and save time.

In our case study, we combined more methods. The threat catalog was thoroughly studied, and some threats were selected as relevant. Check-list had been made from these threats, and this check-list we used for security inspection. Threats identified for our case study:

- **Assault, injury**
- Property theft, damage (vandalism), or destruction
- Pickpocketing
- Fire or flood
- Power failure, delay in the arrival of performers, cancellation of the concert
- Entrance with an invalid ticket [15].

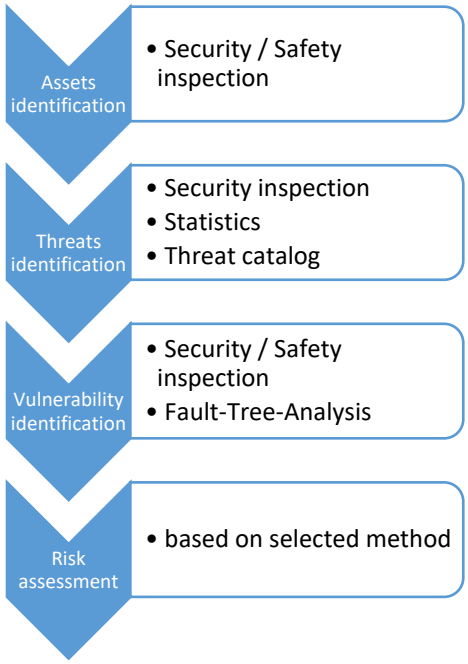


Figure 11: Risk assessment workflow

Vulnerability is always determined at a given location when existing security measures, accessibility to the facility, employee reactions, etc., are checked based on the inspection. Vulnerability identification for our case study:

- Obsolete security systems
- Absence of mechanical restraint systems
- Lack of the Visitor Regulations
- Number of glazed surfaces
- Proximity to the river

- Etc.

The whole process is outlined in the following steps (sub-processes) (fig. 11).

The final risk assessment is made with the help of the TOPSIS method. Inputs are based on the opinion of three "common" security managers. "Common" means that they are not specialized in advanced risk analysis and especially in the field of soft targets protection.

The main objective was to find the most exposed location with the highest probability of attack that may have fatal consequences. Based on experience, the location of attacks is determined by on-site inspection, security assessment, and determining which areas are attractive for the attack. This can be assessed from the point of view of their vulnerability, lack of security, or the number of people we expect at a given location. Thanks to the localization of the attack, when planning security measures, we can directly target security at a given location and thereby strengthen it.

Analyzed localizations are:

- Main entrance
- Audience space
- Stage
- Refreshments space
- Parking lot [15].

Initially, we must define the criteria for TOPSIS, which increase or decrease the risk. The criteria were determined on a similar basis to the location of the attack. Thanks to our many years of experience in planning the security of cultural events, as well as from the methodologies, books, and works of experts that have been published on this topic. Initially, it is always important to say what we want to analyze. For soft targets, one of the most critical locations are those with accumulated large numbers of people. Subsequently, we focus on how easy it is to get into the object, if someone stops the attacker, etc. When I already have security measures, I have to ask if they are effective and if they can detect and respond to an attack.

In many cases, security measures are only passive. It means that security measures serve as a monitoring tool for the occurred security situation but cannot respond. This is entirely unacceptable in the case of an attack on people. Therefore, the effort is to plan measures that will actively detect the attack and trigger some reaction (alarm). Our chosen criteria are defined in the basic document for soft target protection [13]. Those criteria are:

- A number of people – it is labeled as a benefit criterion because more people increase the risk.
- Accessibility – benefit criterion; higher accessibility = lower security -> less effective security measures such as security guards or surveillance system.
- Detectability – because this criterion decreases the risk, it is labeled as a cost criterion. Security guards are able to detect the attack in the early stage.
- Reactivity – another cost criterion. This criterion expresses the time between the start of attack and the first reaction of guards or policemen.

7. Results

There is evaluation from three professionals and security specialists in the following tables (Tab. IV, V, and VI).

Table 4: Decision maker 1 [15]

	Number of people	Accessibility	Detectability	Reactivity
Weight (numerical)	5	3	3	1
Main entrance	{5, 7, 9}	{7, 9, 9}	{1, 3, 5}	{5, 7, 9}
Audience space	{7, 9, 9}	{3, 5, 7}	{3, 5, 7}	{3, 5, 7}
Stage	{1, 3, 5}	{1, 1, 3}	{5, 7, 9}	{5, 7, 9}
Refreshments space	{5, 7, 9}	{3, 5, 7}	{3, 5, 7}	{3, 5, 7}
Parking lot	{1, 3, 5}	{7, 9, 9}	{1, 1, 3}	{1, 3, 5}

Table 5: Decision maker 2 [15]

	Number of people	Accessibility	Detectability	Reactivity
Weight (numerical)	4	4	5	4
Main entrance	{5, 7, 9}	{7, 9, 9}	{5, 7, 9}	{5, 7, 9}
Audience space	{7, 9, 9}	{3, 5, 7}	{1, 3, 5}	{1, 3, 5}
Stage	{1, 3, 5}	{1, 1, 3}	{7, 9, 9}	{5, 7, 9}
Refreshments space	{3, 5, 7}	{3, 5, 7}	{1, 3, 5}	{1, 3, 5}
Parking lot	{1, 3, 5}	{7, 9, 9}	{1, 1, 3}	{1, 1, 3}

Table 6: Decision maker 3 [15]

	Number of people	Accessibility	Detectability	Reactivity
Weight	5	3	4	3
Main entrance	{5, 7, 9}	{7, 9, 9}	{1, 3, 5}	{3, 5, 7}
Audience space	{7, 9, 9}	{3, 5, 7}	{1, 3, 5}	{1, 1, 3}
Stage	{1, 1, 3}	{1, 1, 3}	{7, 9, 9}	{7, 9, 9}
Refreshments space	{3, 5, 7}	{1, 3, 5}	{3, 5, 7}	{3, 5, 7}
Parking lot	{1, 3, 5}	{7, 9, 9}	{3, 5, 7}	{1, 3, 5}

When we have the results from decision makers, we combined the decision fuzzy matrix (1) (Tab. VII).

Table 7: Combined decision fuzzy matrix[15]

		Number of people	Accessibility	Detectability	Reactivity
weight	numerical	5	3	4	3
	linguistic	Very high	Average	High	Average
Main entrance		{5, 7, 9}	{7, 9, 9}	{1, 4.333, 9}	{3, 6.333, 9}
Audience space		{7, 9, 9}	{3, 5, 7}	{1, 3.667, 9}	{1, 3, 7}
Stage		{1, 2.333, 5}	{1, 1, 3}	{5, 8.333, 9}	{5, 7.667, 9}
Refreshments space		{3, 5.667, 9}	{1, 4.333, 7}	{1, 4.333, 7}	{1, 4.333, 7}
Parking lot		{1, 3, 5}	{7, 9, 9}	{1, 2.333, 7}	1, 2.333, 5}

Then we normalized the fuzzy matrix according to (2) and (3) (Tab. VIII).

Table 8: Normalized fuzzy matrix [15]

	Number of people	Accessibility	Detectability	Reactivity
weight	{7, 9, 9}	{3, 5, 7}	{5, 7, 9}	{3, 5, 7}
Main entrance	{0.556, 0.778, 1}	{0.778, 1, 1}	{0.111, 0.231, 1}	{0.111, 0.158, 0.333}
Audience space	{0.778, 1, 1}	{0.333, 0.556, 0.778}	{0.143, 0.273, 1}	{0.143, 0.333, 1}
Stage	{0.111, 0.259, 0.556}	{0.111, 0.111, 0.333}	{0.111, 0.120, 0}	{0.111, 0.130, 0}
Refreshments space	{0.333, 0.630, 1}	{0.111, 0.481, 0.778}	{0.143, 0, 1}	{0.143, 0.032, 1}
Parking lot	{0.111, 0.333, 0.556}	{0.778, 1, 1}	{0.143, 0, 1}	{0.200, 0, 1}

Then we created a weighted normalized fuzzy matrix according to (4) (Tab. IX). The A^+ and A- are according to (5) and (6).

Table 9: Weighted normalized fuzzy matrix [15]

	Number of people	Accessibility	Detectability	Reactivity
Main entrance	{3.889, 7, 9}	{2.333, 5, 7}	{0.556, 1.615, 9}	{0.333, 0789, 2.333}
Audience space	{5.444, 9, 9}	1, 2.778, 5.444}	{0.714, 1.909, 9}	{0.429, 1.667, 7}
Stage	{0.778, 2.333, 5}	{0.333, 0.556, 2.333}	{0.556, 0.840, 1.800}	{0.333, 0.652, 1.400}
Refreshments space	{2.333, 5.667, 9}	{0.333, 2.407, 5.444}	{0.714, 1,615, 9}	{0.429, 0.161, 7}
Parking lot	{0.778, 3, 5}	{2.333, 5, 7}	{0.714, 3, 9}	{0.600, 2.143, 7}
A+	**{5.444, 9, 9}**	**{2.333, 5, 7}**	**{0.714, 3, 9}**	**{0.600, 2.143, 7}**
A-	**0.778, 2.333, 5}**	**{0.333, 0.556, 2.333}**	**{0.556, 0.840, 1.800}**	**{0.333, 0.652, 1.400}**

Then we calculated distance FPIS and FNIS according to (7) (Tab. X).

Table 10: Final FPIS and FNIS distances [15]

	FPIS				FNIS			
Main entrance	1.463	0.000	0.805	3.808	3.977	3.896	4.181	0.650
Audience space	0.000	1.745	0.630	2.217	5.235	2.241	4.203	3.348
Stage	5.235	3.896	4.341	4.253	0.000	0.000	0.000	0.283
Refreshments space	2.632	2.093	0.799	3.002	3.137	2.090	4.182	3.234
Parking lot	4.959	0.000	0.000	1.929	0.385	3.896	4.341	3.433

As the results show, the riskiest area is the audience space. Many people in a small area have a crucial effect on early detection and mainly on possible reactions. A mass of people complicates and prevents the moving of security guards. The second risky area is the main entrance. It reflects practical experience from previous terrorist attacks. Many people standing in the queue are very attractive to attack by vehicles. The parking lot stands in third place. Protection of this area is a little bit complicated because people move in unpredictable directions, and installing anti-collision barriers is almost impossible. The results were calculated by using the formula (8) (Tab. XI).

Table 11: Results [15]

CC_i	**Localizations**	**rank**
0,676	Main entrance	2
0,766	Audience space	1
0,016	Stage	5
0,597	Refreshments space	4
0,636	Parking lot	3

8. Discussion

There are many different methods that should be tested in the risk management of soft targets; however, FTOPSIS shows very promising results (Tab. XI). It can easily involve "standard" security managers and specialists and reflect the opinion of all of them. Also, the results were discussed and consulted with experts for soft targets protection. Their professional view confirmed our results. The exact value of risk is not important. Important is the rank in the case of social or cultural events. It is necessary to test this method and approach for other types of soft targets and validate results with others risk analysis methods.

Used method (FTOPSIS) has provided results that are consistent and comparable with practice. Even if this case study was limited by cooperation with only three experts, the results are representative enough. This case study also validates and confirms the utilization of FTOPSIS as a risk analysis. Many other risk analyses have a problem subjective approach of an evaluator. This could be a serious problem when we have different results from different evaluators. This may bring an underestimation of risk, which means, in general, higher risk. FTOPSIS brings a certain amount of objectivity and it should be used without prominent experience in soft target protection.

This article aimed to examine one of the risk analysis methods and test it with other selected methods. As already mentioned, the intention is to find a method that will be simple for experts, quick to use, and above all, objective. But the work of a security expert does not end there. We now know what and where we are in danger and to what extent. Subsequently, it is necessary to start the process of reducing the risk, where we look for various measures to reduce the resulting value of the threat. There are several ways - security measures, insurance, avoiding a given risk, monitoring, etc. The most commonly used are technical and non-technical security measures, by which we understand various security systems, camera systems, and fire protection, but also the setting of processes for how to behave in the event of a security incident, crisis communication, detection of suspicious behavior, training of security and emergency personnel, etc.

9. Conclusion

Risk analysis is about finding and ordering risks. This process is very similar to other decision-making processes and thus, MCDM should be effectively applied to this field. The biggest challenge, as for every decision-making, is correctly defined criteria. A properly set template may save a lot of time, and usage of FTOPSIS could be faster and more time effective than other risk analysis methods based on subjective opinion. On the other hand, FTOPSIS presumes some elementary knowledge of mathematics.

Next research step is case studies and validation on upcoming events. We would like to use MCDM not only for risk analysis but also for prioritization of proposed security measures and evaluation of incidents.

Acknowledgment

This work was supported by the research project VI20192021163 "Built-up and operation development of security systems at mass events", supported by the Ministry of the Interior of the Czech Republic in the years 2019-2021.

References

[1] D. Kotkova, L. Kralik, L. Kotek, "Multiple Criteria Decision-making: Risk Analyses for Cultural Events as one of the Soft Target Categories," in 2021 IEEE International Carnahan Conference on Security Technology (ICCST) 1-6, 2021, doi: 10.1109/ICCST49569.2021.9717382.

[2] D. Lapkova, L. Kotek, J. Nevrkla, K. Jenckova, "Decision support - Tool for customization of security measures for cultural events," in 2020 11th Iberian Conference on Information Systems and Technologies (CISTI), 1-6, doi: 10.23919/CISTI49556.2020.9140917.

[3] D. Kotkova, M. Hromada, T. Sternova, K. Ljubymenko, "Methodology of identification and protection of soft targets of transport infrastructure – initial study," in 2020 Transport Means - International Conference, 1107-1112. 2020.

[4] L. Kotek, M. Hromada, D. Lapkova, "Protection of Soft Targets from Terrorism," in 2019 IEEE 6th Asian Conference on Defence Technology (ACDT), 71-78. 2019, doi: 10.1109/ACDT47198.2019.9072743.

[5] D. Lapkova, L. Kotek, D. Rozek, "Terrorism Situation and Identification of Soft Target's Attractiveness," in 2019 IEEE 6th Asian Conference on Defence Technology (ACDT), 65-70, 2019, doi: 10.1109/ACDT47198.2019.9072875.

[6] D. Lapkova, L. Kotek, D. Rozek, "Counter-terrorism Measures - Determining of Soft Target's Vulnerability," in 2019 IEEE 6th Asian Conference on Defence Technology (ACDT), 79-84, 2019, doi: 10.1109/ACDT47198.2019.9072861.

[7] L. Kotek, M. Hromada, D. Kotkova, "Educational platform for personal and community protection situations from the perspective of soft targets," in 2020 Iberian Conference on Information Systems and Technologies (CISTI), 1-4, 2020, doi: 10.1109/ICCST52959.2022.9896561.

[8] D. Kotkova, K. Jenckova, "DETERMINING OF SOFT TARGET'S VULNERABILITY," in 2019 DAAAM International Scientific Book, 205-214, 2019, doi: 10.2507/daaam.scibook.2019.16.

[9] D. Kotkova, K. Jenckova, "COMPARISON OF MODUS OPERANDI OF TERRORIST'S ATTACKS - VEHICLE RAMMING ATTACK AND FIREARM ATTACK," DAAAM International Scientific Book 2019. Vienna : DAAAM International Vienna, **2019**, s. 191-198. ISBN 978-3-902734-24-2, doi: 10.2507/daaam.scibook.2019.14.

[10] D. Kotkova, T. Sternova, Z. Kalvach, L. Kotek, "The Process to Plan Security Measures for Cultural Events," in 2021 IEEE International Carnahan Conference on Security Technology (ICCST), 1-6, 2021 doi: 10.1109/ICCST49569.2021.9717380.

[11] D. Kotkova, M. Hromada, M. Malanikova, S. Kovar, "The concept of a software tool with an interactive map for identification and determination of soft targets of transport infrastructure," in 2021 IEEE International Carnahan Conference on Security Technology (ICCST), 1-5, 2021 doi: 10.1109/ICCST49569.2021.9717375.

[12] K. Ljubymenko D. Kotkova, T. Sternova, "Use of detection of suspicious behavior in ensuring the security of persons in transport," in 2021 IEEE International Carnahan Conference on Security Technology (ICCST), 1-4, 2021, doi:10.1109/ICCST49569.2021.9717399.

[13] Z. Kalvach, "Definition of Soft Targets," Prague, 2017

[14] T. Sternova, "Vývoj teroristických útoků v období 2014 - 2017. Rozbor vybraných útoků v roce 2017 (Development of terrorist attacks in 2014-2017. Analysis of selected attacks in 2017.)," in The Conference of Security Technologies, Systems and Management, 2017.

[15] Application for database management of terrorist attacks, *DoVA* [online], Prague, 2019.

[16] L. Hofreiter, V. Berezutskyi, L. Figuli, Z. Zvakova, "Soft Target Protection-NATO Science for Peace and Security Series C: Environmental Security," Springer, ISBN: 978-94-024-1754-8.

[17] S. Nădăban, S. Dzitac, I. Dzitac, "Fuzzy TOPSIS: A General View," Procedia Computer Science, **91**, 823-831, 2016, doi: https://doi.org/10.1016/j.procs.2016.07.088.

[18] I. Mahdavi, A. Heidarzade, B. Sadeghpour-Gildeh, "A general fuzzy TOPSIS model in multiple criteria decision making," Int J Adv Manuf Technol **45**, 406 – 420, 2009. https://doi.org/10.1007/s00170-009-1971-5

[19] N. B. Kore, K. Ravi, S. B. Patil, "A simplified description of fuzzy TOPSIS method for multi criteria decision making," International Research Journal of Engineering and Technology (IRJET), **4**(5), 2047-2050, 2017, e-ISSN: 2395 -0056.

[20] G. Kannan, S. Pokharel, P. S. Kumar, "A hybrid approach using ISM and fuzzy TOPSIS for the selection of reverse logistics provider," Resources, Conservation and Recycling, **54**(1), 28-36, 2009 doi: https://doi.org/10.1016/j.resconrec.2009.06.004.

[21] E. Afful-Dadzie, "Hybridized integrated methods in fuzzy multicriteria decision making (with case studies)," PhD Thesis, Thomas Bata University in Zlin, 2013.
[22] B. Vahdani, S. M. Mousavi, R. Tavakkoli-Moghaddam, "Group decision making based on novel fuzzy modified TOPSIS method," Applied Mathematical Modelling, **35**(9), 4257-4269, 2011, doi: https://doi.org/10.1016/j.apm.2011.02.040.
[23] P. Rani, A. R. Mishra, A. Mardani, F. Cavallaro, M. Alrasheedi, A. Alrashidi, "A novel approach to extended fuzzy TOPSIS based on new divergence measures for renewable energy sources selection," Journal of Cleaner Production, **257**(1), 120352, 2020, doi:10.1016/j.jclepro.2020.120352.
[24] G. K. Koulinas, O. Demesouka, P. Marhavilas, A. P. Vavatsikos, D. Koulouriotis, "Risk assessment using fuzzy TOPSIS and PRAT for sustainable engineering projects," Sustainability, **11**, 615, 2019, doi: 10.3390/su11030615.
[25] M. Nazam, J. Xu, Z. Tao, J. Ahmad, M. Hashim, "A Fuzzy AHP-TOPSIS Framework for the Risk Assessment of Green Supply Chain Implementation in the Textile Industry," International Journal of Supply and Operations Management, **2**(1), 548-568, 2015, doi: 10.22034/2015.1.02
[26] K. Jenckova, "Zajišťování bezpečnosti kulturních akcí za účasti většího množství osob v České republice (Ensuring the safety of cultural events with the participation of a larger number of people in the Czech Republic)," diploma thesis, Police Academy of Czech Republic in Prague, 2019.
[27] D. Kotkova, L. Kralik, "Change Of Soft Target's Resilience In Time – Case Study," in 2019 DAAAM International Scientific Book, 199-204, 2019, doi: 10.2507/daaam.scibook.2019.15.
[28] R. F.S.M. Russo, R. Camanho, "Criteria in AHP: A Systematic Review of Literature, Procedia Computer Science," **55**, 2015, 1123-1132, https://doi.org/10.1016/j.procs.2015.07.081.
[29] M. Behzadian, S. K. Otaghsara, M. Yazdani, J. Ignatius, "A state-of the-art survey of TOPSIS applications, Expert Systems with Applications," Volume **39**(17), 2012, Pages 13051-13069, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2012.05.056.
[30] T. Licek, "Vícekriteriální rozhodování agentů v ekonomických prostředích," University of Hradec Kralove, 2015, Czechia.
[31] A. Daniel, F. E. Onuodu, "A Fuzzy AHP Model for Selection of University Academic Staff," International Journal of Computer Applications **141** (2016), 19-26, 2016, doi: 10.5120/ijca2016908969.
[32] R. R. Venkata, "Decision making in manufacturing environment using graph theory and fuzzy multiple attribute decision making methods," Springer, 2013, doi: https://doi.org/10.1007/978-1-4471-4375-8.
[33] K. Mela, T.Tianen and M. Heinisuo, "Comparative study of multiple criteria decision making methods for building design," Advanced Engineering Informatics, **26**(4), 716-726, 2012, doi: https://doi.org/10.1016/j.aei.2012.03.001.
[34] A. Mardani, E. K. Zavadskas, K. Govindan, A. Amat Senin, A. Jusoh, "VIKOR Technique: A Systematic Review of the State of the Art Literature on Methodologies and Applications," Sustainability. **8**(1), 37, 2016, https://doi.org/10.3390/su8010037
[35] Google Maps Available from: https://www.google.com/maps/place/Zl%C3%ADn/@49.2284554,17.6597917,17z/data=!4m5!3m4!1s0x47130cad4c888e0d:0x418d4c16a8ac214f!8m2!3d49.2244365!4d17.6627635

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 20-30 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Optimization of Query Processing on Multi-tiered Persistent Storage

Nan Noon Noon*, Janusz R. Getta, Tianbing Xia

School of Computing & Information Technology, University of Wollongong, Wollongong, 2500, Australia

ARTICLE INFO

Article history:
Received: 28 September, 2022
Accepted: 11 October, 2022
Online: 13 November, 2022

Keywords:
Multi-tiered persistent storage
Data processing
Scheduling
Storage management

ABSTRACT

The efficient processing of database applications on computing systems with multi-tiered persistent storage devices needs specialized algorithms to create optimal persistent storage management plans. A correct allocation and deallocation of multi-tiered persistent storage may significantly improve the overall performance of data processing. This paper describes the new algorithms that create allocation and deallocation plans for computing systems with multi-tiered persistent storage devices. One of the main contributions of this paper is an extension and application of a notation of Petri nets to describe the data flows in multi-tiered persistent storage. This work assumes a pipelined data processing model and uses a formalism of extended Petri nets to describe the data flows between the tiers of persistent storage. The algorithms presented in the paper perform linearization of the extended Petri nets to generate the optimal persistent storage allocation/deallocation plans. The paper describes the experiments that validate the data allocation/deallocation plans for multi-tiered persistent storage and shows the improvements in performance compared with the random data allocation/deallocation plans.

1. Introduction

In the last decade, we have observed the fast-growing consumption of persistent storage used to implement operational and analytical databases [1]. While the databases become larger, the total number of database applications also continuously increases and the applications themselves, especially the analytical ones, become more sophisticated and advanced than before [2]. Such trends increase the pressure on hardware resources for data processing, particularly on high-capacity and fast persistent storage devices.

Usually, the financial constraints invalidate the *single-step replacements* of all available persistent storage devices with better ones. Instead, a typical strategy is based on the continuous and systematic replacements of persistent storage devices with only a few at a time. Also, from an economic point of view, it is not worth investing significant funds in faster and larger persistent storage devices when only some of the available data is frequently processed. In reality, only some data sets are accessed more frequently than others. Financial and data processing requirements lead to the simultaneous utilization of persistent storage devices of different speeds and capacities. Therefore, data is distributed over many different storage devices with various capacity and speed characteristics [3]. This leads to a logical model of the *multi-tiered organization of persistent storage* [4, 5]. The lower *tiers (levels)* consist of higher capacity and slower persistent storage devices than the lower capacity and faster devices at the higher levels.

Several research works have been already performed on the automatic allocation of storage resources over multi-tiered persistent storage devices and multi-tier caches. For example, in one of the solutions, persistent storage can be allocated over several cache tiers and in different arrangements [6]. Another research work shows how to distribute data on disk storage in the multi-tier hybrid storage system (MTHS) [7].

A number of approaches schedule the allocation of resources over multi-tiered persistent devices based on future predicted workloads. The algorithms for an optimal allocation of persistent storage on multi-tiered devices in environments where future workloads can be predicted have been proposed in [8, 9]. These algorithms arrange data according to an expected workload and contribute to the automated performance tuning of database applications. Most of the existing research outcomes show that performance tuning with the predicted database workloads improves performance during processing time and reduces database administration time [10].

A logical model of multi-tiered persistent storage consists of several *tiers (levels)* of persistent storage visible as a single persistent storage container. The processing speeds at the same tier are alike. Each tier is divided into several *partitions*, where each partition is a logical view of a physical persistent storage device

*Corresponding Author: Nan Noon Noon Email: nnn326@uowmail.edu.au

that implements a particular tier. Such a view is compatible with the organization of persistent storage within cloud systems. Nowadays, data can be stored on several remote devices. It contributes to the changes in the working style where the employees can work from home or distance. Therefore, storage on the cloud is required to provide access to data online anywhere, anytime over an internet connection. In addition, cloud storage for organizations must be shared by many users. Therefore, partitioning of persistent storage is required so that users can access storage simultaneously. Also, the efficient management of storage allocation on multi-tiered persistent storage with partitions is an essential factor for the performance of cloud systems.

The illustration of multi-tiered persistent storage with partitions is shown in Figure 1. Typically, the higher tiers of persistent storage have lower capacities and faster access times than the lower tiers. In a physical model, a sample *multi-tiered persistent storage* consists of the physical persistent storage devices available on the market, such as NVMe, SSD, and HDD [3].

Efficient processing of data stored at *multi-tiered persistent storage* is an interesting problem. When processed, data can "flow" from lower tiers to free space at the higher tiers to speed-up access to such data in the future. Scheduling the data transfers between the tiers require the allocation/deallocation of data based on the speed and capacities of the tiers. Thus, making the correct scheduling decisions automatically and within a short period of time becomes a critical factor for the overall performance of data processing.

Because of its limited capacity, it is impossible to store all data at the topmost and the fastest tier. Therefore, we must prepare for a compromise between the capacity, speed, and price of available *multi-tiered persistent storage*. Such compromise leads to all data being distributed over many tiers of multi-tiered persistent storage.

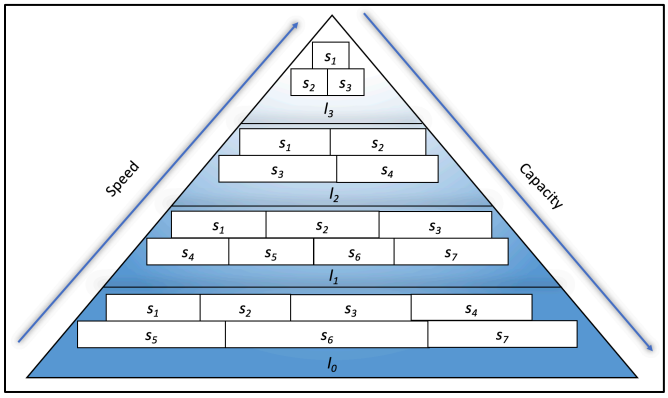


Figure 1: Illustration of multi-tiered persistent storage with four tiers divided into partitions

The main research objective of this work is to speed up data processing in environments where all data is located at *multi-tiered persistent storage*. We assume that data processing is organized as a collection of pipelines where the outputs from one operation are the inputs to the next operation in a pipeline. A strategy to speed up the data processing in a pipeline is to write the outputs of each operation to the highest available persistent tiers. The benefits of such a strategy are twofold. First, data is written faster at the higher tiers than at the lower ones. Second, the following operation in a pipeline reads input data more quickly from the higher tiers. An essential factor in such a strategy is an optimal release of persistent storage allocated at the higher tiers to store temporary data.

Another problem addressed in the paper is the lower-level optimization of query processing on *multi-tiered persistent storage*. The present query optimizers do not consider an efficient implementation of higher-level operations on data located at *multi-tiered persistent storage*. However, it is possible to significantly improve performance by implementing higher-level operations in ways consistent with the characteristics of available persistent storage. Our second research objective is to generate more efficient data processing plans for predicted and unpredicted database workloads through the optimal implementation of elementary operations on data located at *multi-tiered persistent storage*.

The contributions of the paper are the following. First, the paper shows how to extend and apply a notation of Petri nets to describe the data flows in multi-tiered persistent storage. Second, the paper presents the algorithms that find optimal query processing plans through the linearization of Petri nets. Third, the paper analyses the results of experiments to show that query processing plans generated by the new algorithms efficiently use the properties of *multi-tiered persistent storage*.

This paper is an extension of work originally presented in [11]. The paper is organized in the following way. Section 2 extends a notation of Petri nets to describe data flow in query processing. Sections 3 and 4 present the algorithms for the automatic allocation of storage available on multi-tiered persistent storage devices. Section 5 describes an experiment, and Section 6 summarizes the paper.

2. Basic Concepts

In this work, we consider a scenario where a database application submits an SQL query to a relational database server. Then, a typical query optimizer finds an optimal query processing plan. Such a plan includes the list of operations organized as a directed bipartite graph. Then, to discover the data flows between the elementary operations, a plan is transformed into an extended Petri Net [12]. Finally, a graphical notation of Petri Nets represents the flow of data when processing a query.

The original Petri net model includes two types of elements: places and transactions. Places are denoted as circles, and transactions are denoted as vertical rectangles. Zero or more tokens, denoted as small black circles, can be located in the places. Arcs are connected between places and transactions—the illustration of Petri Net is shown in Figure 2.

An *Extended Petri Net* is quadruple $<B, E, A, W>$ where B and E are disjoint sets of *places* and *transitions*. *Places* are visualized as circles, and *transitions* are visualized as rectangles or bars. *Arcs* $A \subseteq (B \times E) \cup (E \times B)$ connect the *places* and *transitions*. A *place* may contain a finite number of tokens visualized as black circles, or it can be empty. A *transition fires* depending on the number of tokens connected at the input place. When an input place has no token, a *transition* is at a waiting state.

In our case, sets of places, B, are interpreted as input/output data sets, and the sets of transitions, E, are interpreted as operations on data sets (see Figure3 below).

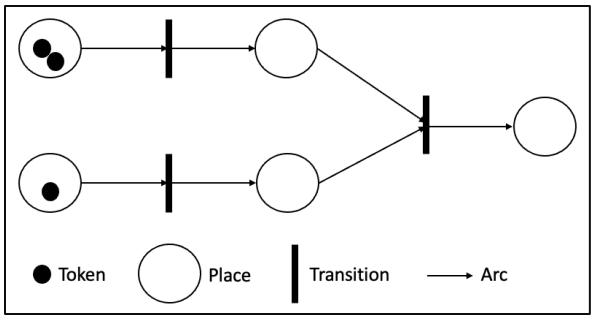


Figure 2: An original sample structure of a Petri net

Arcs, *A,* determine the input and output data sets of the operations. A weight function $W: A \rightarrow N^+$ determines the estimated total number of data blocks read and written by each operation. See the numbers attached to the *arcs* in Figure3. The *root nodes* are the *places* (data sets) in *B* that do not have the arcs from the *transitions* (operations) *E*. The root nodes represent the input data sets located in a database. An operation can have more than one input and output data set. Likewise, more than one operation can share input and output data sets.

There are two types of persistent storage data sets *B*: *permanent* or *temporary*. Both types of data sets are stored in *multi-tiered storage. A temporary* data set can be removed from the *storage* when it is no longer needed.

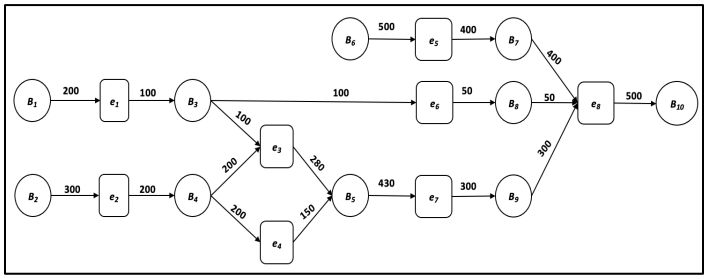


Figure 3: Data processing and data flows represented as an extended Petri net

Let $E = \{e_i, ..., e_j\}$ be a set of operations obtained for a query *Q* processing plan created by a database query optimizer. Each operation e_i is represented by a pair $e_i = (\{B_i, ..., B_m\}, \{B_{(m+1)}, ..., B_n\})$, where $\{B_i, ..., B_m\}$ are the input data sets and $\{B_{(m+1)}, ..., B_n\}$ are the output data sets of an operation e_i in *E*. Some of the input and output data sets processed by an operation e_i are permanent, and some of them are temporary.

This paper focuses on managing the multi-tiered persistent storage efficiently for each input/output data sets of each operation and generates the best allocation plan, called a processing plan, for each query. A list of persistent storage tiers is denoted as a sequence $L = <l_0, ..., l_n>$. We have $n+1$ tiers where l_0 denotes the lowest tier, called a base tier or a backup tier. The highest tier is denoted by l_n. Each tier l_i for $i = 0 \ldots n$ is described as a triple $(r_i, w_i, <s_{i_1}, ..., s_{i_n}>)$ where r_i is a read speed per data block at tier l_i, w_i is a write speed per data block at tier l_i, and $<s_{i_1}, ..., s_{i_n}>$ is a sequence of partitions arranged from smallest available storage to largest available storage, where s_{i_j} is an amount of free space available in partition *j* at tier l_i. The read and write speed per data block is measured in the standardized time units that can be converted into real-time units when a specific hardware implementation of a persistent storage tier is considered.

The processing plan is denoted as $P = <D_{i_j}, ..., D_{n_k}>$, where D_{j_i} is a set of plans for an operation e_i, where e_i is in E_{s_j}. Each plan is a pair and is denoted as (B_i, s_{n_i}) and B_i is the total number of input/output data blocks located in partition *i* in tier l_n (index of s_{n_i}). If $l_j = l_0$, then the data blocks are located at the lowest tier in the *multi-tiered storage.*

After the output result of the operation e_i is allocated, unnecessary temporary input storage of operation e_i needs to be removed from multi-tiered persistent storage according to the persistent storage release plan. The persistent storage release plan is denoted as a sequence of pairs $\delta = <(e_i, (B_k, ..., B_h)), \ldots, (e_j, (B_x, ..., B_y))>$, where each pair $(e_i, (B_k, ..., B_h)$ is a group of data sets $(B_k, ..., B_h)$ that can be released after the results of an operation e_i are saved.

Then, the total processing time to read (B_i, s_{i_j}) data blocks located at partition *j* in tier l_i is $T_{B_i} = B_i * r_i$. Formula (1) given below determines the total processing time T_r to read the data blocks for each input data set $D_i = \{(B_1, s_{n_j}), \ldots, (B_m, s_{k_n})\}$.

$$T_r = \text{Max}\left(T_{B_1} + \cdots + T_{B_m} \right) \quad (1)$$

Similarly, formula (2) determines the total processing time T_w to write the data blocks for each output data set $D_j = \{(B_{(m+1)}, s_{i_j}), \ldots, (B_n, s_{k_h})\}$.

$$T_w = \text{Max}\left(T_{B_{(m+1)}} + \cdots + T_{B_n} \right) \quad (2)$$

Let an *Extended Petri Net* <*B, E, A, W*> represent a query processing plan. Then, whenever it is possible, the output data sets of the operations $E = \{e_i, ..., e_k\}$ are written to the tiers located above the tiers of their input data sets. In other words, whenever possible, each operation tries to *push up* the results of its processing towards the partition at the higher tiers of persistent storage.

The benefits of such a strategy are as follows. First, it is faster to write the output data sets at a higher tier than at a lower tier. Second, the data sets written at a higher tier can be read faster by the next operation in a pipeline of a query processing plan. Therefore, the benefits include the time we gain from writing output data sets to the higher tiers and from reading the same data sets by the other operations.

3. Resource Allocation for a Single Query

This section describes the algorithms that convert a query processing plan obtained from the execution of `EXPLAIN PLAN` statement for a single query into an Extended Petri Net and serialize the operations of the Net in order to minimize the amounts of persistent storage required for query processing.

3.1. Creation of Extended Petri Net

We consider a query *Q* expressed as a `SELECT` statement and submitted for processing by a relational database server operating on *multi-tiered persistent storage.* A query *Q* is *ad-hoc* processed by the system in the following way. First, a query optimizer transforms the query into the best query processing plan. SQL statement `EXPLAIN PLAN` can be used to list a plan found by a

query optimizer. Next, the implementations of extended relational algebra operations such as selection, projection, join, anti-join, sorting, grouping, and aggregate functions are embedded into the query processing plan. Then, the plan is converted into an *Extended Petri Net*.

Operation $e_j \in E$ where e_j is member of query Q. Each operation has the estimated amounts of input data to be read from persistent storage and the estimated output amounts of data to be saved in persistent storage. Algorithm 1 below transforms a query Q into an *Extended Petri Net*.

The algorithm obtains a query processing plan from the results of the `EXPLAIN PLAN` statement. From there, we get a set of operations $E = \{e_i, ..., e_m\}$. A query processing plan provides information about the input data sets read by the operations and the output data sets written by the operations. The permanent and temporary data sets read and written by the operations in E form a set of places B in *Extended Petri Net*. The arcs leading from the input data sets to the operations and the arcs leading from the operations to the output data sets create a set of arcs A. A query processing plan provides information about the amounts of storage read and written by the operations. Such values contribute to the weight values w_j attached to each arc and are represented by a pair (a_j, w_j) in W. A weight value w_j attached to an arc between the input data set and an operation represents the total number of data blocks read by the operation. A weight value w_j attached to an arc between the operation and output data set represents the total number of data blocks written by the operation.

Algorithm 1: Create Extended Petri Net according to input a query q

Input: a query Q
Result: an extended Petri Net $<B, E, A, W>$

(1) Get a query processing plan through the application of `EXPLAIN PLAN` statement.
(2) Get operations from a query processing plan with the estimated amounts of input and output storage required by each operation. Then, create a set of operations E where e_i is an operation.
(3) **while** *Iterate over E* **do**
- Let the current operation be e_i.
- Let $B_i, ..., B_m$ be the input data sets and let $B_{(m+1)}, ..., B_n$ be the output data sets of an operation e_i.
- Append to A the arcs $a_1, ..., a_m$ linking $B_{i-1}, ..., B_{i-m}$ and operation node.
- Next, label the arcs with the values $w_{i_1}, ..., w_{i_m}$ representing the total number of input data blocks read by an operation e_i and append the pairs *to a weight function* W.
- Append to A the arcs $a_{m+1}, ..., a_n$ linking an operation node e_i and the nodes $\{B_{(m+1)}, ..., B_n\}$. Then, label the arcs with values representing the total number of output data blocks written by an operation e_i and append the pairs $(a_1, w_{i_1}), ..., (a_m, w_{i_m})$ to a weight function W.

end

3.2. Serialization of Extended Petri Net

Preparation of an Extended Petri Net $<B, E, A, W>$ for processing requires serialization of its operations in a set E. Serialization arranges a set of operations E into a sequence of operations E_s. There exist many ways that the operations in E can be serialized in the implementations of more complex queries. If the operations in E can be serialized in more than one way, then we try to find a serialization that requires smaller amounts of persistent storage for its processing. Algorithm 2 finds a serialization of operations in E that tries to minimize the total amounts of persistent storage allocated during query processing. As a simple example, consider an Extended Petri Net given in Figure 3. The operations e_1 and e_2 write 100 and 200 units of persistent storage. Then, 100 units of persistent storage are released after the processing of operations e_3 and e_6, and 200 units of persistent storage are released after the processing of operations e_3 and e_4. Since the processing of e_3 and e_4 releases more storage, we assign e_3 and e_4 before e_6 in the serialization.

A *root operation* in E is an operation such that at least one of its input arguments is a permanent data set that belongs to the contents of a database. A *top persistent storage data* set in B is a data set that contains the final results of query processing and such that is not read by any other operation in E. It is possible that a query may output more than one top persistent storage data set. A *top operation* is an operation that writes only to the top persistent storage data sets.

Algorithm 2 traverses the arrows in A backward from the *top operations* to the *root operations* and generates a sequence of operations E_s from a set of operations E.

At the very beginning, a sequence of operations E_s is empty, and none of the persistent storage data sets in B is marked as released. In the first step, the algorithm finds the *top operations* and inserts such operations into a set of candidate operations E_c.

For each operation in E_c, the algorithm computes profit using formula (4). To compute profit, the algorithm deducts the total amounts of persistent storage allocated by an operation from the total amounts of persistent storage released by an operation. For example, assume that the current operation is e_i. To find the amounts of persistent storage released by an operation e_i, we perform the summation of the amounts of storage in all input data sets of operation e_i not marked as released yet. It is denoted by $\sum_{j=1}^{n} B_j$. To find the profits from the processing of an operation e_i, we deduct from the released amounts the amounts of persistent storage included in the output data sets of the operation and denoted as $\sum_{k=(n+1)}^{m} B_k$ in formula (3) below.

$$p(e_i) = \sum_{j=1}^{n} B_j - \sum_{k=(n+1)}^{m} B_k \quad (3)$$

The profits are computed for each operation in E_c. The algorithm selects an operation with the smallest profit. If more than one operation has the smallest profit, then the algorithm randomly chooses one of them. Let a selected operation be e_i. The algorithm removes e_i from E_c, appends it in front of E_s, and marks all data sets read by e_i as released. Next, the algorithm finds the operations that write only to data sets marked as released by e_i, such that data sets released by e_i are read only by the operation already in E_s. An

objective is to find the operations that can be appended to E_c in a correct order of processing determined by a set of arcs A. Then, the algorithm appends those selected operations to a set of candidate operations E_c and the profits for each operation in E_c are computed again. An operation with the lowest profits is appended in front of E_c. The algorithm repeats the procedure until all operations in E are appended in front of E_s.

Algorithm 2: Generating Sequence of Operations E_s

Input: An *Extended Petri Net* $<B, E, A, W>$ from Algorithm 1.
Result: A sequences of operations E_s for query Q.

(1) Create empty sequence $E_s = <>$ for Q, create empty candidate operation sets $E_c = \varnothing$,
(2) Use a *Petri Net* $<B, E, A, W>$ and get the top operation(s) from *Petri Net* that is connected to the top container/container that stored the final result and put those operation(s) into a set of operations $E_c = \{e_i, \ldots, e_k\}$.
(3) **while** E_c *not empty* **do**
  if E_c *contains more than one operation* **then**
    - Use formula (3) to compute the estimated profit for each operation in E_c.
    - Get all the operations with the smallest profit and randomly select one of the operations with the smallest profit.
  end
  - Let selected operation be e_i.
  - Append e_i in front of E_s and remove it from E_c.
  - Mark all data sets read by e_i as released.
  - Find all operations that write only to the data sets marked as released and save them in a set E_w.
  - Remove from E_w the operation that writes to data sets read by the other operations not in E_s.
  - Append the operations from E_w to E_c.
end
(4) End algorithm and return the result E_s.

Example1: The example explains a sample trace of Algorithm 2 when applied to an Extended Petri Net given in Figure3. In the first step, E_c contains only one operation e_8 connected to top data set B_{11}. Therefore, e_8 has the lowest profit in E_c. We append it in front of $E_s = <e_8>$ and we remove it from E_c. The data sets B_5, B_6, and B_7 read only by e_8 are marked as released. The operations writing only to the data sets B_8, B_9, and B_{10} are e_5, e_6, e_7. None of the data sets B_8, B_9, and B_{10} are read by an operation that is not in E_s. Hence the operations e_5, e_6, e_7 can be added to $E_c = \{e_5, e_6, e_7\}$. The estimated profits for each operation computed with a formula (3) are the following: $p(e_5) = (0 - 400) = -400$, $p(e_6) = (100 - 50) = 50$ and $p(e_7) = (430 - 300) = 70$. According to those results, e_5 is picked, appended in front of $E_s = <e_5, e_8>$, and removed from $E_c = \{e_6, e_7\}$. A data set B_7 is an input data set, and it cannot be marked as released. As no new data sets are released, E_c remains the same. Next, an operation e_6 which has the second smallest profit, is taken from E_c. It is appended in front of $E_s = <e_6, e_5, e_8>$ and is removed from $E_c = \{e_7\}$. A data set B_3 read by e_6 is marked as released. A data set B_3 is marked as written by operation e_1 and later it is read by e_3. In this case, e_1 cannot be appended to E_c because a data set B_3 is read by e_3 that is not assigned to E_s yet. Next, e_7 is taken from E_c, it is appended in front of $E_s = <e_7, e_6, e_5, e_8>$, and is removed from E_c. A data set B_5 read by e_7 is marked as released. A data set B_5 is written by the operations e_3 and e_4. Both operations are appended to $E_c = \{e_3, e_4\}$. The computations of the profits provide: $p(e_3) = 0 - 280 = -280$ and $p(e_4) = 0 - 150 = -150$. According to the results, e_3 is appended in front of $E_s = <e_3, e_7, e_6, e_5, e_8>$ and it is removed from $E_c = \{e_4\}$. The data sets B_3 and B_4 read by e_3 are marked as released. The operations e_1 and e_2 are written to data sets B_3 and B_4. An operation e_1 can be appended into E_c because e_6 that read B_3 is already assigned in E_s. An operation e_2 cannot be appended into $E_c = \{e_1, e_4\}$ because e_4 that reads a data set B_4 is not in E_s yet. The computations of the profits provide: $p(e_1) = 0 - 100 = -100$ and $p(e_4) = 200 - 150 = 50$. According to the results, e_1 is appended in front of $E_s = <e_1, e_3, e_7, e_6, e_5, e_8>$ and is removed from $E_c = \{e_4\}$. Operation e_1 reads an input data set, therefore, no new operations can be added to E_c. Next, operation e_4 is append to $E_s = <e_4, e_1, e_3, e_7, e_6, e_5, e_8>$ and is removed from E_c. A data set B_4 read by e_4 is marked as released and operation e_2 is appended to E_c. Finally, operation e_2 is appended in front of $E_s = <e_2, e_4, e_1, e_3, e_7, e_6, e_5, e_8>$. Operation e_2 reads from an input data set and because of that, no more operations can be assigned to E_c. The final sequence of operations $E_s = <e_2, e_4, e_1, e_3, e_7, e_6, e_5, e_8>$ is generated at the end.

3.3 Estimation of Storage Requirements and Generation of Persistent Storage Release Plan

A sequence of operation E_s obtained from the linearization of an Extended Petri Net with Algorithm 2 is used to estimate the requirements on the amounts of persistent storage needed for the processing of the sequence. Algorithm 3 estimates the persistent storage requirements while processing the operations in E_s. The output of Algorithm 3 is the estimated processing time T for E_s using the read/write speed of single-tier, the maximum required storage V to process E_s, and a sequence of the persistent storage release plans $\delta = <(e_i, (B_k, \ldots, B_h)), \ldots, (e_j, (B_x, \ldots, B_y))>$.

Algorithm 3 uses a sequence of operations E_s and information about reading speed per data block r_n and writing speed per data block w_n parameters unchanged and extracted from a single tier in multi-tiered persistent storage. Typically, the algorithm is processed with the read/write parameters of the highest tier. First, the algorithm sets $V = 0$, which stores the maximum storage required to be allocated in the multi-tiered persistent storage, and $T = 0$, which stores the estimated processing time for E_s. Next, the algorithm iterates over E_s in step (2) in Algorithm 3 and lets the current operation be e_i. First, the total read storage V_r is computed by summing all input storage for e_i and the total write storage V_w is computed by summing all output storage for e_i. After that, the current temporary storage $V_{temp} = V_{temp} + V_w$ is updated, where V_{temp} stores the temporary storage required to process until the current operation. Next, the algorithm compares V_{temp} with V. If V_{temp} is larger than V, then $V = V_{temp}$ is updated. Next, the algorithm computes the estimated total processing time t_i for current operation e_i using the following equation.

$$T(e_i) = (V_r * r_n) + (V_w * w_n) \quad (4)$$

Then, the algorithm updates the total processing time until operation e_i by summing $T = T + T(e_i)$.

Next, the algorithm gets the storage released by operation e_i, denoted as a pair $(e_i, (B_x, \ldots, B_y))$. The detailed procedure is shown

from step (2)(vi) to step (2)(viii) in the algorithm. After this, the storage is appended into the persistent storage release plan δ.

If e_i is the member of δ then the algorithm needs to release data blocks after the output data blocks of e_i are allocated. Next, the algorithm needs to update $V_{temp} = V_{temp}$ - *the total number of release storage by* e_i.

Algorithm 3 iterates those procedures above for each operation from the input sequence E_{s_i}. Finally, Algorithm 3 will generate V, T, and $\delta = <(e_i, (B_k, \ldots, B_h)), \ldots, (e_j, (B_x, \ldots, B_y))>$.

Algorithm 3: Estimate the total storage required, estimate the execution time, and create a persistent storage release plan.

Input: A sequence of operations $E_s = <e_i, \ldots, e_k>$ obtained from Algorithm 2 and information about the reading speed per data block r_n and the writing speed per data block w_n of a single tier in multi-tiered persistent storage.
Result: The maximum storage V and the estimated processing time T required to process E_s and a sequence of the persistent storage release plan $\delta = <(e_i, (B_k, \ldots, B_h)), \ldots, (e_j, (B_x, \ldots, B_y))>$.

(1) Set $V = V_{temp} = T = 0$, $E_{temp} = \{\}$ and $\delta = <>$
(2) **while** *iterate over* E_{s_i} **do**
  (i) Let current operation be $e_i = (\{B_j, \ldots, B_k\}, \{B_m, \ldots, B_n\})$ where $\{B_j, \ldots, B_k\}$ is the input set of the number of data blocks and $\{B_m, \ldots, B_n\}$ is the output set of the number of data blocks.
  (ii) Next, get the total reading (V_r) and writing (V_w) data blocks for e_i.
  (iii) Next, use the values of V_r, V_w, r_n, w_n to compute $T(e_i)$ by using a formula (4). Update $T = T + T(e_i)$.
  (iv) Update $V_{temp} = V_{temp} + V_w$.
  (v) **if** $V_{temp} > V$ **then** $V = V_{temp}$.
  (vi) Get all the input storage read only by e_i and get all the input storage connected to both e_i and other operations in E_{temp} and create a group of data sets $(B_x, \ldots, B_y)$.
  (vii) Next, remove a data set from $(B_x, \ldots, B_y)$ if the data set is the input argument of root operations in E_s.
  (viii) Create a pair $(e_i, (B_x, \ldots, B_y))$ and append it into δ.
  if $e_i \in \delta$ **then**
    - Add the storage requirements of $(B_x, \ldots, B_y)$ to get the total number of data blocks R_i released by e_i.
    - Update $V_{temp} = V_{temp} - R_i$.
  end
end
(3) Return V, T and $\delta = <(e_i, (B_k, \ldots, B_h)), \ldots, (e_j, (B_x, \ldots, B_y))>$.

Example 2: A sample trace of Algorithm 3 is performed on a sequence of operations $E_s = <e_2, e_4, e_1, e_3, e_7, e_6, e_5, e_8>$ from Example 1 together with the reading and writing speed per data block from the highest tier $l_3 = (0.05, 0.055, <500, 300, 100>)$. As it has been mentioned before, the reading and writing speed per data block is measured in the standardized time units.

At the very beginning, we set $V = V_{temp} = T = 0$, and $\delta = <>$. Next, we iterate over E_s, and let the current operation be e_2. It reads $V_r = 300$ data blocks and writes $V_w = 200$ data blocks. Next, we compute $t_2 = (300*0.05) + (200*0.055) = 26$ and update $T = T + t_2 = 26$. Then, we set $V_{temp} = V = 200$. A persistent storage release plan δ remains empty because e_2 is the first operation in the sequence E_s.

In the next iteration, the current operation is e_4. We compute $t_4 = (200*0.05) + (150*0.055) = 18.25$ and update $T = 26 + 18.25 = 44.25$ and $V_{temp} = 200 + 150 = 350$. Due to V_{temp} being larger than V, we increase $V = 350$. The operation does not release storage and there is no need to update δ.

In the next iteration e_1 is the current operation. $T = 59.75$, $V_{temp} = 350 + 100 = 450$ and $V = 450$.

In the next iteration e_3 is the current operation. $T = 90.15$, $V_{temp} = 450 + 280 = 730$ and $V = 730$. This time, e_3 released 200 data blocks, therefore, we need to update $\delta = <(e_3, (B_4))>$. Total number of data blocks released by e_3 is 200. Therefore, we update $V_{temp} = 730 - 200 = 530$.

In the next iteration, e_7 is the current operation. $T = 128.15$ and $V_{temp} = 530 + 300 = 830$ and $V = 830$. This time we release storage B_5 and update $\delta = <(e_3, (B_4)), (e_7, (B_5))>$ and $V_{temp} = 830 - 430 = 400$.

In the next iteration, e_6 is the current operation. $T = 135.9$ and $V_{temp} = 400 + 50 = 450$ and $V = 830$. The storage released by e_6 is B_3 and update $\delta = <(e_3, (B_4)), (e_7, (B_5)), (e_6, (B_3))>$ and $V_{temp} = 450 - 100 = 350$.

In the next iteration, e_5 is the current operation. The algorithm computes $T = 182.9$ and $V_{temp} = 350 + 400 = 750$ and $V = 830$. An operation e_5 does not release any storage.

The last operation from the sequence E_s is e_8. The algorithm computes $T = 247.9$ and $V_{temp} = 750 + 500 = 1250$ and $V = 1250$. This operation releases three storages, B_8, B_9, and B_{10}. The updated storage released plan is $\delta = <(e_3, (B_4)), (e_7, (B_5)), (e_6, (B_3)), (e_8, (B_7, B_8, B_9))>$.

Finally, Algorithm 3 returns $T = 247.9$ of time units, $V = 1250$ data blocks and $\delta = <(e_3, (B_{i4})), (e_7, (B_5)), (e_6, (B_3)), (e_8, (B_7, B_8, B_9))>$ for a sequence of operations E_s.

4. Resource Allocation for a Group of Queries

Algorithm 4 takes on input a set of query processing plans $\mathcal{E} = \{E_{s_1}, \ldots, E_{s_n}\}$ created by Algorithm 2, a set of estimated processing times $\mathcal{T} = \{T_1, \ldots, T_n\}$, a set of maximum storage requirements for each processing plan $\mathcal{V} = \{V_1, \ldots, V_n\}$, a set of deallocation plans $\mathcal{D} = \{\delta_1, \ldots, \delta_n\}$ from the Algorithm 3, and a sequence of persistent storage tiers $L = <l_0, \ldots, l_n>$. The output of the algorithm is a sequence of storage allocation plans P.

First, the algorithm gets the estimated processing time T and the estimated highest allocation storage V for processing plans from $\mathcal{E}$. Next, the algorithm collects the candidate operations from $\mathcal{E}$, where a candidate operation is the first operation from each sequence, and puts them into a set of candidate operations E_c. Next, the algorithm picks one operation from E_c in the following way. First, the algorithm picks the operations from the sequences with the smallest volume, V. If more than one operation is found, then from the operations found so far, the algorithm picks the operations from the sequences with the shortest time, T. If again more than one operation is found, then the algorithm uses formula (3) to compute a profit for each operation and picks the operations with

the highest profit. Again, if more than one operation is still found, then the algorithm randomly picks one operation. Let selected in this way, the candidate operation be $e_i = (\{B_i, ..., B_j\}, \{B_k, ..., B_h\})$, and the total number of data blocks in output data sets of the operation be V_w.

Algorithm 4 passes the information about L, V_w, and P to Algorithm 4.1 to find the best partition and generate the updated storage allocation plan P. Next, Algorithm 4 must update L with information about temporary storage to be released after the processing of e_i. To do so, Algorithm 4 passes a deallocation plan δ_i for E_{s_i}, a sequence of tiers L, and the operation e_i to Algorithm 4.2 to update information about available storage in L.

Finally, an operation e_i is removed from E_{s_i}. The algorithm repeats the steps above for each operation from each sequence until all sequences are empty. The algorithm returns the final storage allocation plan P.

Algorithm 4: Generating storage allocation plans

Input: A set of processing plans $\mathcal{E} = \{E_{s_1}, \ldots, E_{s_n}\}$ from algorithm 2, deallocation plan δ, the maximum storage V, and the estimated processing time T for each sequence from Algorithm 3. A sequence of tiers $L = <l_0, \ldots, l_n>$ where each tier l_i is described as triple $(r_i, w_i, <s_{i_j}, \ldots, s_{i_i}>)$.
Result: A sequence of allocation plans $P = <D_{i_j}, \ldots, D_{n_k}>$.

(1) Create empty sequence of allocation plans $P = <>$.
(2) **While** *until all sequences in* $\mathcal{E}$ *are empty* **do**
- Create an empty set of candidate operations $E_c = \{\}$.
- Get the first operations from the sequences with the smallest volume V and put them into E_c.
if *more than one operation is found in* E_c **then**
- Keep the operations from the sequence $E_{s_1}, \ldots, E_{s_n}$ with the smallest estimated processing time T and remove the rest of the operations from E_c.
else if *more than one operation is found in* E_c **then**
- Keep the operations with the highest profit using formula (3) and remove the rest of the operations from E_c.
else if *more than one operation is found in* E_c **then**
- Select one operation randomly and make E_c empty.
end
- Let the selected operation be $e_i = \{B_i, \ldots, B_j\}, \{B_k, \ldots, B_h\}$ from E_{s_j} and its input storage be $\{B_i, \ldots, B_j\}$ and its output storage be $\{B_k, \ldots, B_h\}$.
- Let the number of total output storage be V_w.
- Use the information of L, V_w, and P to find the best partition using Algorithm 4.1 and get the result of updated plan P from algorithm 4.1.
- Pass the information about a deallocation plan δ_j, a state of multi-tiered persistent storage L, and the operation e_i to Algorithm 4.2 to update L with information about temporary storage released by e_i.
- Remove e_i from E_{s_j}.
- Update estimated processing time T and V for E_{s_j}.
end
(3) Return the sequence of multi-tiered storage allocation plans $P = <D_{i_j}, \ldots, D_{n_k}>$.

Algorithm 4.1 finds the best partition using the information in L, V_w, P and e_i provided by Algorithm 4.

Let e_i be a member of E_{s_j}, and a set of allocation plans for e_i is denoted D_{j_i}. First, the algorithm needs to find the best partition located at the highest possible tier in order to achieve the best processing time. Next, the algorithm finds a partition that can accommodate the storage V_w where V_w is the total output storage required to store the temporary/permanent result of e_i. The input storage $\{B_i, \ldots, B_j\}$ is already assigned to one of the partitions on L. Therefore, the algorithm needs to find the best storage allocation for V_w only.

Next, the algorithm iterates over the tiers in L. Let the current tier be $(r_n, w_n, <s_{n_j}, \ldots, s_{n_i}>)$. The algorithm needs to check whether the required storage V_w can be allocated entirely in one partition or over more partitions. When the amount of storage available at one of the partitions in the highest tier is larger or equal to V_w, the algorithm creates an allocation plan (B_i, s_{n_j}) where $B_i = V_w$ and appends that plan to D_{j_i}. But sometimes, the amounts of storage available at all partitions in the highest tier are smaller than the required storage V_w. In that case, the algorithm splits storage V_w into multiple storages V_w', …, V_w''. Some storage is to be allocated at the faster tier, and some may be at the lower tier. For that case, the algorithm splits V_w to allocate more than one partition and creates allocation plans (B_i, s_{n_j}), …, (B_j, s_{n_k}). Next, the algorithm appends all those plans into D_{j_i}. Finally, the allocation plan D_{j_i} is appended to P and returned to Algorithm 4.

Algorithm 4.1: Finding the best partition and generating a plan for storage allocations

Input: A multi-tiered persistent storage $L = <l_0, \ldots, l_n>$, storage requirements V_w of an operation e_i, a sequence of storage allocation plans P.
Result: The updated sequence of storage allocation plans P.

(1) Let the amounts of temporary storage $V_{temp} = 0$ and let an initial storage allocation plan D_{j_i} for the operation e_i be empty.
(2) **while** *iterate over L in reverse order* **do**
(i) Let current tier be $l_i = (r_i, w_i, <s_{i_i}, \ldots, s_{i_n}>)$ and let $<s_{i_i}, \ldots, s_{i_n}>$ be a sequence of partitions in l_i arranged from the smallest to the largest partition.
(ii) Iterate over partitions $<s_{i_i}, \ldots, s_{i_n}>$ and choose a partition s_{i_k} such that its size is equal or larger than V_w.
(iii) **if** *size of* s_{i_k} *is larger or equal to* V_w **then**
- Create a pair (B_i, s_{n_i}) where $B_i = V_w$ and append it into D_{j_i}.
- Update $s_{i_k} = s_{i_k} - V_w$ and $V_w = V_{temp}$.
- Sort a sequence of partitions arranged from smallest available storage to largest available storage.
else if *all storage in current set is smaller than* V_w **then**
- Update $V_{temp} = V_w$.
while *iterate over partition in reverse order and until* $V_{temp} = 0$ *or all available storage become zero* **do**
- Let the current storage in a set be s_{i_k}.
- Split storage into two parts: V_w, where $V_{temp} = V_{temp} - s_{n_i}$ and $V_w = s_{i_k}$

- Create a pair (B_i, s_{i_k}) where $B_i = V_w$ and append it into D_{j_i}.
- Update $s_{i_i} = 0$, and $V_w = V_{temp}$.
- Sort a sequence of partitions arranged from smallest available storage to largest available storage.

end

end

end

(3) Append $D_{j_i} = \{(B_i, s_{n_i}), \ldots, (B_j, s_{m_k})\}$ into P.

(4) Return P.

Algorithm 4.2 releases persistent storage no longer needed by an operation e_i. An input to the algorithm is a deallocation plan δ_i, a sequence of tier $L = <(r_n, w_i, <s_{n_i}, \ldots, s_{n_i}>), \ldots, (r_0, w_0, <s_{n_0}, \ldots, s_{n_j}>)>$, and the operation e_i passed from Algorithm 4.

Algorithm 4.2 checks whether e_i releases any persistent storage. If e_i is a member of the deallocation plan δ_i, then the algorithm needs to update L according to the deallocation plan, such as removing the storage released by e_i from the occupied storage. If e_i is not a member of the deallocation plan, then the algorithm does not need to update L.

Algorithm 4.2: Deallocation the storage

Input: A deallocation plan δ_i, a multi-tiered persistent storage $L = <l_0, \ldots, l_n>$, and the operation e_i.
Result: The updated multi-tiered persistent storage $L = <l_0, \ldots, l_n>$.

(1) **if** $e_i \in \delta_i$ **then**
- Get the storage released by e_i like $(B_x, \ldots, B_y)$.
while *iterate over* $(B_x, \ldots, B_y)$ **do**
- Let current storage be B_i and the location of B_i be (B_i, s_{h_i}) where storage B_i is located at partition i from tier l_h.
- Remove a storage B_i from partition i in level l_h and update $s_{h_i} = s_{h_i} + B_i$.
- Let s_{h_i} is belong to the sequence $<s_{h_x}, \ldots, s_{h_y}>$.
- Sort a sequence of partitions $<s_{h_x}, \ldots, s_{h_y}>$ arranged from smallest available storage to largest available storage.

end

end

(2) Return the updated $L = <l_0, \ldots, l_n>$.

Algorithm 5 computes the total processing time T_f for the allocation plan P created by Algorithm 4. First, the algorithm iterates over plan P. Let the current set of plans be $D_{j_i} = \{(B_i, s_{n_i}), \ldots, (B_j, s_{i_k})\}$ where D_{j_i} is a set of plans for the operation e_i in sequence E_{s_j}. Next, the algorithm iterates over D_{j_i}. Let the current pair be (B_i, s_{n_i}). Then, the algorithm computes the processing time for that pair. If B_i is a member of input data blocks for e_i, then compute $T_f = T_f + (B_i * r_n)$ where r_n is a reading speed per data block. If B_i is the member of output data blocks for e_i, then compute $T_f = T_f + (B_i * w_n)$ where w_n is a writing speed per data block. Next, the algorithm checks whether e_i is the last operation in the sequence E_{s_j} or not. If e_i is the last operation, then the algorithm needs to compute the reading time for final storage such as $T_f = T_f + (V_w * r_n)$, where V_w is the total data blocks written by operation e_i. Finally, the algorithm returns the total execution time for plan P.

Algorithm 5: Final estimated processing time for allocation plan

Input: A sequence of allocation plans $P = <D_{i_j}, \ldots, D_{n_k}>$.
Result: Total estimated processing time T_f for allocation plan P.

(1) Let total processing time for sequences be $T_f = 0$.
(2) **while** *iterate over* P **do**
- Set total writing data block be $V_w = 0$.
- Let the current plan be $D_{j_i} = \{(B_i, s_{n_i}), \ldots, (B_j, s_{i_k})\}$ for operation $e_i = \{B_i, \ldots, B_j\}, \{B_k, \ldots, B_h\}$ from E_{s_j}.
while *iterate over* D_{j_i} **then**
- Let current pair be (B_i, s_{n_i}) where B_i is a total number of data blocks allocated at a tier n in a partition i (s_{n_i}).
if B_i *is member of input storage* $\{B_i, \ldots, B_j\}$ **then**
- $T_f = T_f + (B_i * r_n)$
else
- $T_f = T_f + (B_i * w_n)$
end
end
if e_i *is the last operation from* E_{s_j} **then**
- Read the final result and release the storage.
- Update $T_f = T_f + (V_w * r_n)$.
end
end
(3) Return total processing time for allocation plan T_f.

Example 3: In this example, we use a multi-tiered persistent storage with 3 tiers. We assume that the highest tier l_2 has 3 partitions, the lower one l_1 has 2 partitions, and the bottom tier l_0 has 3 partitions. The parameters of the tiers are listed below.

$L = <l_0, l_1, l_2>$
- $l_0 = (0.2, 0.21, <200, 500, 1000>)$
- $l_1 = (0.1, 0.105, <100, 200>)$
- $l_2 = (0.05, 0.055, <40, 50>)$

Next, we use a set of sequences $\mathcal{E} = \{E_{s_1}, E_{s_2}, E_{s_3}\}$. A sequence E_{s_1} consists of the following 3 operations $E_{s_1} = <e_1, e_2, e_3>$ where
- $e_1 = (\{50, 50\}, \{100\})$
- $e_2 = (\{100\}, \{50\})$
- $e_3 = (\{50\}, \{20\})$

The total estimated processing time for E_{s_1} is $T_2 = 37.85$ and $V_1 = 150$.

The sequence E_{s_2} consists of one operation, $E_{s_2} = <e_1>$ where
- $e_1 = (\{150\}, \{100\})$

The total estimated processing time for E_{s_2} is $T_2 = 40.50$ and $V_2 = 100$.

The last sequence E_{s_3} consists of 4 operations $E_{s_3} = <e_1, e_3, e_2, e_4>$ where

- $e_1 = (\{50\}, \{40, 40\})$
- $e_2 = (\{40\}, \{20\})$
- $e_3 = (\{40\}, \{10\})$
- $e_4 = (\{30\}, \{10\})$

The total estimated processing time for E_{s_3} is $T_3 = 22.60$ and $V_3 = 110$.

Following step (2) of Algorithm 4, we picked an operation e_1 from sequence E_{s_2}, because V_2 is the smallest volume, and put it into $E_c = \{e_1\}$. Since we have only one candidate operation, it is passed to Algorithm 4.1 to find the storage allocations for the outputs of operation e_1. According to Algorithm 4.1, the best storage is s_{2_3} and therefore, we created a plan like (100, s_{2_3}). We then appended the plan into $P = <(100, s_{2_3})>$. Next, Algorithm 4 released storage if e_1 needs to release some storage. According to Algorithm 4.2, e_1 does not need to release any storage. We repeated the above procedure and finally get the plan P for $\mathcal{E}$ where $P = <\{(40, s_{2_1}), (50, s_{2_2}), (10, s_{1_1})\}, \{(40, s_{2_1}), (40, s_{2_2})\}, \{(10, s_{2_2}), (10, s_{1_1})\}, \{(10, s_{2_1})\}, \{(10, s_{2_1})\}, \{(40, s_{2_1}), (50, s_{2_2}), (10, s_{1_1})\}, \{(50, s_{1_1})\}, \{(20, s_{2_1})\}>$.

Next, we compute the processing time for a sequence of plan P by using the algorithm 5. The total processing time T_f for $\mathcal{E}$ is 112.55 time units. With random allocation, the total processing time for $\mathcal{E}$ become 143.45 time units. Without multi-tiered persistent storage with partitions, the total execution time for $\mathcal{E}$ is 219.90 time units.

5. Example/Experiment

Different types of persistent storage devices such as SSD, HDD, and NVMe can be used to create a multi-tiered persistent storage system. In the experiment, we picked a sample multi-tiered persistent storage L that consists of 4 tiers $<l_0, l_1, l_2, l_3>$, where l_3 is the fastest tier such as NVMe and l_0 is the slowest tier, such as HDD and l_2 and l_1 are faster and slower SSDs. Each tier is divided into two partitions of different sizes. Table 1 shows the read and write speed per data block expressed in standardized time units and the total number of data blocks available at each partition.

Table 1: Read/Write Speed with Available Size for Multi-tiered Storage Devices

Level of Devices	Reading Speed	Writing Speed	Partition 1	Partition 2
l_0	$0.15 * 10^3$	$0.155 * 10^3$	1000	2000
l_1	$0.1 * 10^3$	$0.105 * 10^3$	500	600
l_2	$0.08 * 10^3$	$0.085 * 10^3$	150	200
l_3	$0.05 * 10^3$	$0.055 * 10^3$	50	100

In the experiment, we applied Algorithm 1 to convert eight query processing plans into the Extended Petri Nets. Next, we used Algorithm 2 to find the optimal sequences of operations for each Petri Net. Next, Algorithm 3 was used to get the deallocation plans δ, the maximum volumes V, and the estimated execution times T required for each sequence of operations found by Algorithm 2. The values for each sequence are the following.

$E_{s_1} = <e_1, e_3, e_2, e_4, e_5, e_6, e_7, e_8, e_9>$

$V_1 = 150$, $T_1 = 110.70$

Table 2: Operations With Input and Output Data sets For E_{s_1}

Operation	Input data set	Output data set
e_1	300	100
e_2	200	50
e_3	100	50
e_4	50	20
e_5	50, 20	50
e_6	50	20, 20
e_7	20	10
e_8	20	10
e_9	10, 10	10

$E_{s_2} = <e_2, e_5, e_1, e_4, e_7, e_9, e_3, e_6, e_8, e_{11}, e_{10}, e_{12}>$

$V_2 = 300$, $T_2 = 108.35$

Table 3: Operations with Input and Output Data sets For E_{s_2}

Operation	Input data set	Output data set
e_1	100	50
e_2	100	70
e_3	100	60
e_4	50	30
e_5	70	30
e_6	60	40
e_7	30, 30	50
e_8	40	20
e_9	50	30
e_{10}	20	10
e_{11}	30	20
e_{12}	20, 10	20

$E_{s_3} = <e_1, e_3, e_2, e_4, e_7, e_5, e_8, e_6, e_{11}, e_{10}, e_9, e_{12}, e_{13}, e_{14}, e_{15}>$

$V_3 = 300$, $T_3 = 147.40$

Table 4: Operations with Input and Output Data sets For E_{s_3}

Operation	Input data set	Output data set
e_1	500	400
e_2	600	300
e_3	400	100, 50
e_4	300	200, 50
e_5	100	30
e_6	50	30
e_7	200	100
e_8	50	10
e_9	30	20
e_{10}	30	10
e_{11}	100, 10	60
e_{12}	20	10
e_{13}	60	50
e_{14}	50	40
e_{15}	10, 10, 40	50

$E_{s_4} = <e_1, e_2, e_4, e_3, e_5>$

$V_4 = 250$, $T_4 = 100.70$

Table 5: Operations with Input and Output Data sets For E_{s_4}

Operation	Input data set	Output data set
e_1	80	70
e_2	70	30, 30
e_3	30	20
e_4	30	10
e_5	20, 10	20

$E_{s_5} = <e_1, e_3, e_5, e_2, e_4, e_6, e_7, e_9, e_8, e_{10}>$

$V_5 = 200$, $T_5 = 93.50$

Table 6: Operations with Input and Output Data sets For E_{s_5}

Operation	Input data set	Output data set
e_1	300	100

e_2	100	50
e_3	200	100
e_4	50	30
e_5	100	50
e_6	30	20
e_7	100	50
e_8	50	30
e_9	20, 50	40
e_{10}	30, 40	50

$E_{s_6} = <e_1, e_2, e_3>$

$V_6 = 150, \quad T_6 = 37.85$

Table 7: Operations with Input and Output Data sets For E_{s_6}

Operation	Input data set	Output data set
e_1	50, 50	100
e_2	100	50
e_3	50	20

$E_{s_7} = < e_1>$

$V_7 = 100, T_7 = 40.50$

Table 8: Operations with Input and Output Data sets For E_{s_7}

Operation	Input data set	Output data set
e_1	150	100

$E_{s_8} = < e_1, e_3, e_2, e_4, >$

$V_8 = 110, T_8 = 22.60$

Table 9: Operations with Input and Output Data sets For E_{s_8}

Operation	Input data set	Output data set
e_1	50	40, 40
e_2	40	20
e_3	40	10
e_4	20, 10	10

Next, we used Algorithm 4 to decide on an order of concurrent processing of operations from the sequences in $\mathcal{E}$. Algorithms 4.1 and 4.2 were used within Algorithm 4 to find the best allocation of partitions and to generate a persistent storage allocation plan. The details of the plan are listed in the Appendix. After that, we used Algorithm 5 to simulate the processing of the plan to get the total estimated execution time T_f, see Table X.

In the second experiment, we used the same set of sequences of operations $\mathcal{E}$, and whenever more than one operation could be selected for processing, we randomly picked an operation. The total execution time for the second experiment is given in Table X.

In the third experiment, we used only a single tier of persistent storage, and like before, whenever more than one operation could be selected for processing, we randomly picked an operation. The final results of all experiments are summarized in a Table X.

Table 10: Summary of Experimental Results

Experiment	Method	Total execution time-units
Experiment 1	Using allocation plan over multi-tiered persistend storage that proposed in this work.	857
Experiment 2	Using random allocation plan over multi-tiered persistend storage.	1,068.35
Experiment 3	Using allocation plan without multi-tiered persistend storage.	1,466.95

By comparing those three results, one can find that the execution plan for experiment 1 achieves better performance and faster execution time than experiment 2 and experiment 3.

6. Summary and Future Work

This paper presents the algorithms that optimize the allocations of persistent storage over a multi-tiered persistent storage device when concurrently processing a number of database queries. The first algorithm converts a single query processing plan obtained from a database system into an Extended Petri Net. An Extended Petri Net represents many different sequences of database operations that can be used for the implementation of a query processing plan. The second algorithm finds in an Extended Petri Net a sequence of operations that optimizes storage allocation in multi-tiered persistent storage when a query is processed. The third algorithm estimates the maximum amount of persistent storage and processing time needed when a sequence of operations found by Algorithm 2 is processed.

In the second part of the paper, we considered the optimal allocation of multi-tiered persistent storage when concurrently processing a set of queries. We assumed that the first three algorithms are used for individual optimization of storage allocation plans for each query in a set. The fourth algorithm optimizes the allocation of multi-tiered persistent storage when a set of sequences of operations obtained from the first three algorithms is concurrently processed. The algorithm creates a persistent storage allocation and releases a plan according to the available size and speed of the devices implementing multi-tiered persistent storage.

The last algorithm processes a storage allocation plan created by the previous algorithm and returns the estimated processing time. To validate the proposed algorithms, we conducted several experiments that compared the efficiency of processing plans created by the algorithms with the random execution plans and execution plans without multi-tiered persistent storage. According to the outcomes of experiments, the storage allocation plans obtained from our algorithms consistently achieved better processing time than the other allocation plans.

Several interesting problems remain to be solved. An optimal allocation of persistent storage in the partitions of multi-tiered storage contributes to a dilemma of spreading a large allocation over smaller allocations at many higher-level partitions versus a single allocation at a lower partition.

Another interesting question is related to a correct choice of a level at which storage is allocated depending on the stage of query processing. It is almost always such that the initial stages of query processing operate on the large amounts of storage later reduced to the smaller results. It indicates that the early stages of query processing should be prioritized through storage allocations at higher levels of multi-tiered storage. It means that the parameters of multi-tiered storage allocation may depend on the phases of query processing with faster storage available at early stages.

The next interesting problem are the alternative multi-tiered storage allocation strategies. In one of the alternative approaches, after the serialization of Extended Petri Nets representing individual queries, it is possible to combine the sequence of operations into one large Extended Petri Net and apply serialization again. Yet another idea is to combine the Extended

Petri Nets of individual queries into one Net and try to eliminate multiple accesses to common data containers.

The next stimulating problem is what to do when the predictions on the amounts of data read and/or written to multi-tiered persistent storage do not match the reality. A solution for these cases may require ad-hoc resource allocations and dynamic modifications of existing plans.

It is also possible that a set of queries can be dynamically changed during the processing. For example, a database application can be aborted, or it can fail. Then, the management of persistent storage also needs to be dynamically changed. A solution to such a problem would require the generation of a plan with the options where certain tasks are likely to increase or decrease their processing time.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

I would also like to extend my thanks to the anonymous reviewers for their valuable time and feedback. Finally, I wish to thank my parents for their support and encouragement throughout my studies.

References

[1] Data Storage Trends in 2020 and Beyond, https://www.spiceworks.com/marketing/reports/storage-trends-in-2020-and-beyond/ (accessed on 30 April 2021)

[2] U. Sivarajah, M. M. Kamal, Z. Irani, V. Weerakkody, "Critical Analysis of Big Data Challenges and Analytical methods". In Journal of Business Research, **70**, 263--286, 2017.

[3] S. wang, Z. Lu, Q. Cao, H. Jiang, J. Yao, Y. Dong, P. Yang, C. Xie, "Exploration and Exploitation for Buffer-controlled HDD-Writes for SSD-HDD Hybrid Storage Server", ACM Trans. Storage 18, 1, Article 6, February 2022.

[4] Apache Ignite Multi-tier Storage, https://ignite.apache.org/arch/multi-tier-storage.html (accessed on 30 April 2021)Trans. Roy. Soc. London, A247, 529–551, April 1955.

[5] Tiered Storage, https://searchstorage.techtarget.com/definition/tiered-storage (accessed on 30 April, 2021).

[6] E. Tyler, B. Pranav, W. Avani, Z. Erez, "Desperately Seeking Optimal Multi-Tier Cach Configurations", In 12th USENIX Workshop on Hot Topics in Storage and File System (HotStorage 20), 2020.

[7] H. Shi, R. V. Arumugam, C. H. Foh, K. K. Khaing, "Optimal disk storage allocation for multi-tier storage system", Digest APMRC, Singapore, 2012, 1-7.

[8] N.N. Noon, "Automated performance tuning of database systems", Master of Philosophy in Computer Science thesis, School of Computer Science and Software Engineering, University of Wollongong, 2017.

[9] N.N. Noon, J.R. Getta, "Optimisation of query processing with multilevel storage", Lecture Notes in Computer Science, 9622 691-700. Da Nang, Vietnam Proceedings of the 8th Asian Conference, ACIIDS 2016.

[10] B. Raza, A. Sher, S. Afzal, A. Malik, A. Anjum, Adeel, Y. Jaya Kumar, "Autonomic workload performance tuning in large-scale data repositories", Knowledge and Information Systems. 61. DOI: 10.1007/s10115-018-1272-0.

[11] N.N. Noon, J.R. Getta, T. Xia, "Optimization Query Processing for Multi-tiered Persistent Storage", IEEE 4th International Conference on Computer and Communication Engineering Technology (CCET), 2021, 131-135, doi: 10.1109/CCET52649.2021.9544285.

[12] R. Wolfgang, "Understanding Petri Nets: Modeling Techniques, Analysis Methods, Case Studies", Springer Publishing Company, Incorporated, 2013.

APPENDIX

A processing plan for the experiments described in a section 6 of the paper is the following sequence P of the individual plans.

$P = < D_{7_1}, D_{8_1}, D_{8_3}, D_{8_2}, D_{8_4}, D_{6_1}, D_{6_2}, D_{6_3}, D_{1_1}, D_{1_3}, D_{1_2}, D_{1_4}, D_{1_5}, D_{1_6}, D_{1_7}, D_{1_8}, D_{1_9}, D_{5_1}, D_{5_3}, D_{5_5}, D_{5_2}, D_{5_4}, D_{5_6}, D_{5_9}, D_{5_8}, D_{5_{10}}, D_{4_1}, D_{4_2}, D_{4_4}, D_{4_3}, D_{4_5}, D_{2_2}, D_{2_5}, D_{2_1}, D_{2_4}, D_{2_7}, D_{2_9}, D_{2_3}, D_{2_6}, D_{2_8}, D_{2_{11}}, D_{2_{10}}, D_{2_{12}}, D_{3_1}, D_{3_3}, D_{3_2}, D_{3_4}, D_{3_7}, D_{3_5}, D_{3_8}, D_{3_6}, D_{3_{11}}, D_{3_{10}}, D_{3_9}, D_{3_{12}}, D_{3_{13}}, D_{3_{14}}, D_{3_{15}} >$

Each plan D_{i_j} is the following set of pairs:

$D_{7_1} = \{(50, s_{3_1}), (100, s_{3_2})\}$, $D_{8_1} = \{(40, s_{3_1}), (10, s_{3_1}), (30, s_{3_2})\}$

$D_{8_3} = \{(20, s_{3_2})\}$, $D_{8_2} = \{(10, s_{3_1})\}$, $D_{8_4} = \{(10, s_{3_1})\}$

$D_{6_1} = \{(100, s_{3_2})\}$, $D_{6_2} = \{(50, s_{3_1})\}$, $D_{6_3} = \{(20, s_{3_2})\}$

$D_{1_1} = \{(100, s_{3_2})\}$, $D_{1_3} = \{(50, s_{3_1})\}$, $D_{1_2} = \{(50, s_{3_2})\}$, $D_{1_4} = \{(20, s_{3_2})\}$, $D_{1_5} = \{(50, s_{3_2})\}$, $D_{1_6} = \{(20, s_{3_1}), (20, s_{3_1})\}$, $D_{1_7} = \{(10, s_{3_1})\}$, $D_{1_8} = \{(10, s_{3_1})\}$, $D_{1_9} = \{(10, s_{3_1})\}$

$D_{5_1} = \{(50, s_{3_1}), (100, s_{3_2}), (50, s_{2_1})\}$, $D_{5_3} = \{(100, s_{2_1})\}$, $D_{5_5} = \{(50, s_{3_1})\}$, $D_{5_2} = \{(50, s_{3_2})\}$, $D_{5_4} = \{(30, s_{3_2})\}$, $D_{5_6} = \{(20, s_{3_2})\}$, $D_{5_7} = \{(50, s_{3_2})\}$, $D_{5_9} = \{(30, s_{3_2}), (10, s_{2_1})\}$, $D_{5_8} = \{(30, s_{3_2})\}$, $D_{5_{10}} = \{(50, s_{3_1})\}$

$D_{4_1} = \{(70, s_{3_2})\}$, $D_{4_2} = \{(30, s_{3_1}), (30, s_{3_2})\}$, $D_{4_4} = \{(10, s_{3_1})\}$, $D_{4_3} = \{(20, s_{3_1})\}$, $D_{4_5} = \{(20, s_{3_1})\}$

$D_{2_2} = \{(70, s_{3_2})\}$, $D_{2_5} = \{(30, s_{3_2})\}$, $D_{2_1} = \{(50, s_{3_1})\}$, $D_{2_4} = \{(30, s_{3_2})\}$, $D_{2_7} = \{(50, s_{3_1})\}$, $D_{2_9} = \{(30, s_{3_2})\}$, $D_{2_3} = \{(60, s_{3_2})\}$, $D_{2_6} = \{(40, s_{3_1})\}$, $D_{2_8} = \{(20, s_{3_2})\}$, $D_{2_{11}} = \{(20, s_{3_1})\}$, $D_{2_{10}} = \{(10, s_{3_1})\}$, $D_{2_{12}} = \{(20, s_{3_1})\}$

$D_{3_1} = \{(150, s_{2_1}), (100, s_{2_2}), (50, s_{3_1}), (100, s_{3_2})\}$, $D_{3_3} = \{(50, s_{1_1}), (100, s_{2_2})\}$, $D_{3_2} = \{(150, s_{2_1}), (50, s_{3_1}), (100, s_{3_2})\}$, $D_{3_4} = \{(150, s_{1_1}), (50, s_{2_2}), (50, s_{2_2})\}$, $D_{3_7} = \{(100, s_{3_2})\}$, $D_{3_5} = \{(30, s_{3_1})\}$, $D_{3_8} = \{(10, s_{3_1})\}$, $D_{3_6} = \{(20, s_{2_1}), (10, s_{3_1})\}$, $D_{3_{11}} = \{(60, s_{2_1})\}$, $D_{3_{10}} = \{(10, s_{3_1})\}$, $D_{3_9} = \{(20, s_{3_2})\}$, $D_{3_{12}} = \{(10, s_{3_1})\}$, $D_{3_{13}} = \{(50, s_{3_2})\}$, $D_{3_{14}} = \{(40, s_{3_1})\}$, $D_{3_{15}} = \{(50, s_{3_2})\}$

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 5, 27-34 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

An Improved Model to Analyze the Impact of Cyber-Attacks on Power Systems

Muhammad Musleh Uddin, Kazi Rafiqul Islam*, Md. Monirul Kabir

Department of Electrical and Electronic Engineering, Dhaka University of Engineering and Technology (DUET), Gazipur-1707, Bangladesh

ARTICLE INFO

Article history:
Received: 21 June, 2022
Accepted: 17 August, 2022
Online: 09 September, 2022

Keywords:
Load Frequency Control (LFC)
Automatic Generation Control (AGC)
Automatic voltage regulator (AVR)
Proportional-Integral-Derivative (PID) Controller
Cyber-attack and cyber-security

ABSTRACT

In this paper, an improved model has been proposed for investigating the impact of cyber-attacks on power systems regarding frequency disturbances and voltage disruption while changing the load called ICAPS. The proposed ICAPS model is formulated by five different controllers, such as LFC, AGC, AGC-PID, AVR, and AVR-PID, implemented in two sets of the system model. Specifically, a stable limit of the speed regulation of LFC, integral controller gain of AGC, and amplifier gain of AVR are determined from their characteristic equations derived from the Routh-Hurwitz array. In contrast, the Proportional-Integral-Derivative (PID) controller gains for AGC-PID and AVR-PID are determined using the tuning process. The key aspect of this paper is to obtain the impact of cyber-attacks on the power system in terms of frequency disturbances and voltage disruptions while changing the load. According to our knowledge, no one has considered these issues at a time. In order to evaluate the proposed ICAPS model and how it works, a series of experiments was conducted using the MATLAB Simulink tool. The simulation results are presented in this paper in terms of frequency deviation and voltage disruption (i.e., positive and negative biased cyber-attack) and system oscillations. Finally, the simulation results successfully identified the most severe attack in this model to prevent the power system from unstable conditions.

1. Introduction

The modern power system industry uses supreme technological innovation that enables more adaptable and efficient access to the system structure. In addition, the reliability and security of information communication technology (ICT) and digital computer techniques are crucial for the operation of power systems, which increases the risk of cyber-attacks. Due to cyber-attack, the cyber-vulnerable components of the power system and the effects of cyber-attack should be adequately analyzed [1]-[3].

The advanced control loop, sensor and communication networks are the cyber component of the power plant [4]. The sensors collect data for various parameters, which are sent to the controller to ensure stability of the power system [5]. Data is transferred using the communication network, which is one of the cyber-vulnerable components of the power system. An attacker can get it while sending data from the sensing department to the control department at the power station. The attacker can gain control over the automatic controller of a power plant after accessing confidential information. Ultimately, any change to the controller actions may create malicious data accessing in the controller, which disrupts the stability of the system.

However, in order to investigate the impact of cyber-attack on the power system, a few works have been done in the literature (e.g., [6]-[9]). Specifically, in [6], the authors have discussed about the infrastructure of power system for identifying vulnerabilities. After that, a framework has been proposed to overcome the unstable condition of power system. Furthermore, the effect of cyber-attack on wind farms and power system networks that is also investigated in [7]. On the other hand, an anomaly detection technique has been proposed in [8] as a countermeasure against the effect of faulty data injection on AGC. The authors have investigated the impact of cyber-attack to AGC loop in the power system in [9]. It is noted that the reliability of the power system is an important factor in assessing cyber-security threats. That's why, the impact of such threats need to be estimated upon the cost of the whole power system [10]. Finally, it can be concluded that, although the existing techniques (e.g., LFC, AGC) tried to investigate the impact of cyber-attacks properly, they failed to achieve significant improvement in such cases. Therefore, the

*Corresponding Author Kazi Rafiqul Islam, Email: rafiqul@duet.ac.bd

effects of cyber-attacks should be properly analyzed to maintain the system's stability.

However, to analyze the proper deviation to be provided the system stability, the impact of cyber-attacks on power system has been presented in this paper. The proposed ICAPS is combined by five different controllers, such as, LFC, AGC, AGC-PID, AVR and AVR-PID that are implemented in two different sets of the system models, among which the first three controllers are included for one set of power system models, whereas, the rest of two controllers are for different set of system models. This work focuses on the impact of cyber-attacks on system in terms of frequency and voltage disruptions during changing the load. It should be noted that, our proposed ICAPS can perform well to investigate the impact of cyber-attacks on the power system in terms of frequency disturbances and voltage disruption during changing the load. The most serious cyber-attack on power plant is also identified successfully.

The rest of the paper has been organized into three sections, among which proposed system investigated, in Section 2. Finally, Section 3 discusses the conclusion of our paper.

2. Proposed System Modelling

In this paper, the impact of cyber-attacks on power system has been proposed to ensure the system stability. To understand the proposed model ICAPS clearly, the following five subsections need to be discussed: (i) cyber-attack impacts on LFC, (ii) cyber-attack impacts on AGC, (iii) cyber-attack impacts on AGC-PID, (iv) cyber-attack impacts on AVR, and (v) cyber-attack impacts on AVR-PID**.**

2.1 Cyber-Attack Impacts on LFC

A load frequency controller (LFC) is an essential power plant component. The proper operation of the LFC is important for the safe operation of the power system [11]. However, the issue is that LFC is a powerful cyber-attack tool. The load on power plant is constantly changing. It is well known that the frequency of an alternator varies as the load of power station changes. The system can become unstable as the frequency changes. Once again, it is impossible to repair the load, and the only option is to regulate the prime mover speed. In this case, LFC is a crucial part. . A simple LFC is composed by a turbine, a generator, and a governor presented in Figure 1.

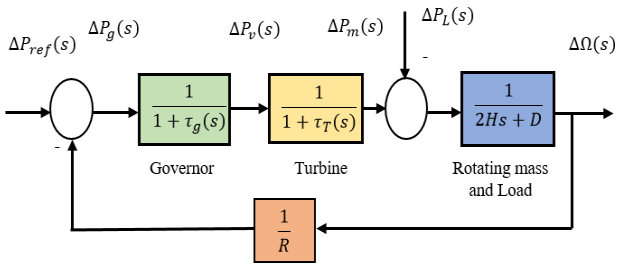


Figure 1: Isolated power system control block diagram based on LFC adapted from [12]

Whether the control system is in stable condition (i.e., not attacked by an unauthorized person), the system frequency falls than the normal value for a short time due to sudden load changes. The frequency sensor detects a decrease in the system's frequency and transmits the signal to the LFC for proper speed control. The LFC prime mover adjusts the governor speed control based on the signal from the frequency sensor to compensate the speed and frequency of the system [12]. On the other hand, in case of an unauthorized access (i.e., cyber-attack) to the LFC loop, if any change to the speed regulation (R) occurs then that can create frequency fluctuation, which disrupts the system stability. The aim of LFC is to balance the actual (i.e., active) power in the power system by controlling the frequency of the system. In a system when the imbalance between load and generation occurs, it must be corrected within seconds to avoid frequency deviation. Depending on speed regulation (R), the governor adjusts generation with varying load demand, maintaining a stable frequency. Its value can be determined by applying Routh–Hurwitz array in its characteristics equation. The closed-loop transfer function of this system is,

$$\frac{\Delta\Omega(s)}{-\Delta P_L(s)} = \frac{(1+\tau_g s)(1+\tau_T s)}{(2H_s+D)(1+\tau_g s)(1+\tau_T s)+\frac{1}{R}} \quad (1)$$

On the other hand, the open loop transfer function of this system is,

$$KG(s)H(s) = \frac{1}{R(1+\tau_g s)(1+\tau_T s)(2Hs+D)} \quad (2)$$

here, R refers to speed regulation, τ_T refers to turbine time constant, $\tau_g$refers to governor time constant, H refers to generator inertia constant, D refers to load coefficient (1% change in frequency) f, refers to nominal frequency, ΔP_Lrefers to Change in load of the power system.

In LFC system, increasing the load demand (ΔP_L) causes a fall in generator frequency, and vice versa. The reason is that the response of the rotating mass inertia is insufficient to bring the generator frequency to the nominal value. A detailed discussion can be found in [1], whereas the block diagram of LFC system with cyber-attack point specifications shown in Figure 2.

Table 1: Parameters of Proposed LFC, AGC, and AGC-PID Power System Models, Adapted from [12]

Symbols	Parameters	Values
τ_T	Turbine time constant	0.5 s
τ_g	Governor time constant	0.25 s
H	Generator inertia constant	8s
D	Load coefficient (1% change in frequency)	1.16
R	Speed Regulation	0.04
f	Nominal frequency	50 Hz
ΔP_L	Change in load	0.25pu

The parameters used in LFC system are presented in Table I. The power station is considered an isolated power station with a turbine-rated output is 200 MW at nominal frequency of 50Hz. Considering the sudden load changes 50 MW, according to the Routh-Hurwitz array to LFC loop with considering the values of parameters mentioned in Table I and Eq. 1, the obtained value of speed regulation (R) is R> 0.009678. As per theoretical calculation,

it is proved that if R is equal to 0.04 then the system will be marginal stable. The frequency deviation during a cyber-attack for various R values presented in Figure 3. It is seen that the system frequency is stable for the value of R=0.04. It means that for stable operation of this power station, R should be equal to 0.04. It is clear that, depending on the affected value of R, cyber-attacks can be categorized into two ways: positively biased attack and negatively biased attack.

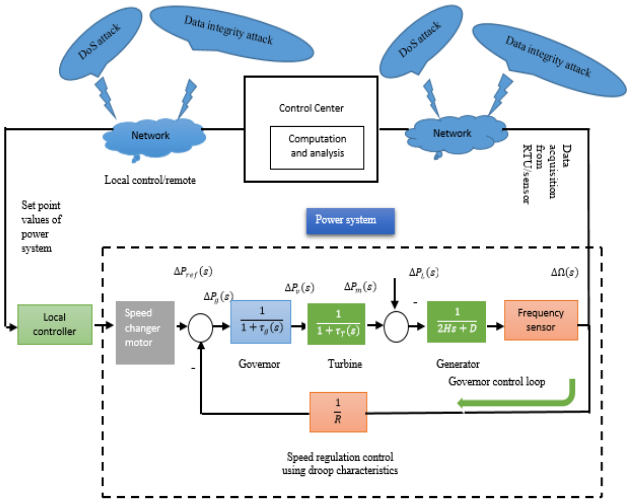


Figure 2: Attack points and general block diagram of a single-area LFC system, adapted from [1].

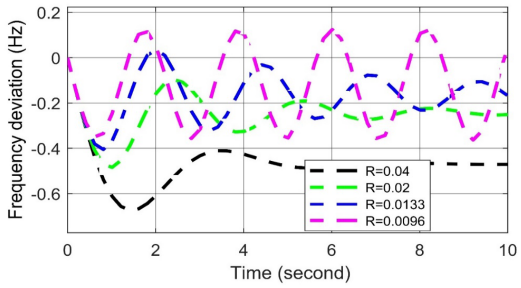


Figure 3: Frequency deviation response to different R values in LFC in case of cyber-attack.

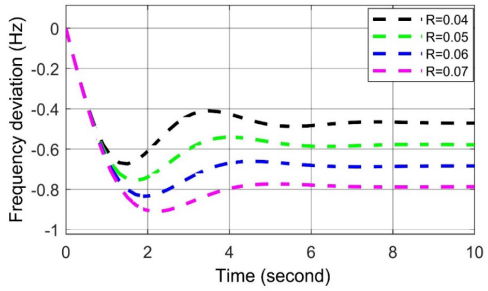


Figure 4: Frequency deviation response for positive biased cyber-attack in LFC.

2.1.1 *Positively biased cyber-attack on LFC*

According to the result of the cyber-attack on LFC system, the value of R increases from the set value, which is called positively biased cyber-attack. It is observed from Figure 4 that the frequency oscillation curve becomes stable quickly when the value of R increases, although the frequency decreases. According to positive bias cyber-attack, the LFC loop decreases the frequency of the system, but it quickly returns to a stable position. However, an excessive deviation in system frequency is not permitted here since the governor's excessive deviation cannot restore the system's frequency.

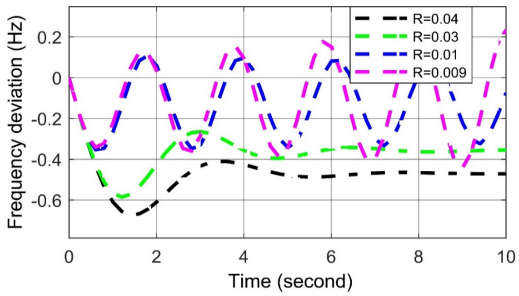


Figure 5: Frequency deviation response for negative biased cyber-attack in LFC.

2.1.2 *Negatively biased cyber-attack on LFC*

According to the result of the cyber-attack on LFC system, the value of R decreases from the set value, which is called negatively biased cyber-attack. It is observed from Figure 5 that, the governor can restore the system frequency when R = 0.03 but it takes more time than the set value R = 0.04. In contrast, other frequency curves never reach to a steady state in both cases (i.e., R = 0.01 and R = 0.009). As a result of the cyber-attack, speed regulation value falls down the steady state position (i.e., R> 0.009678), so frequency deviation is being fluctuated in nature. It increases the oscillations since the governor cannot restore the system's frequency. It can be said that the negative biased cyber-attack is very difficult to maintain a stable system condition, and such a cyber-attack can disrupt the whole system.

2.2 *Cyber-Attack Impacts on AGC*

It is known that changes in system loads at primary LFC result in fluctuations in the steady state frequency depending on the governor speed regulation [12]. It takes more time to restore the system frequency of the desired value (i.e., LFC primary loop). A change in LFC is required to mitigate the frequency fluctuation to zero. Adding an integral controller to the LFC system is referred to as an automatic generation control (AGC). The necessary controller gain (K_I) improves the nature of the system in such a way that the activity finally forces the frequency fluctuation to zero [2]. The KI must be adjusted with a proper set value for a suitable transient response. Thus, it is essential to set the value of this parameter to an appropriate level for proper operation. Because of the failure to select K_I, the governor cannot restore the system's frequency. As a result of the cyber-attack, K_I is considered as a vulnerable quantity. According to unauthorized access to the AGC loop, any change to K_I may create frequency fluctuation, disrupting the system's stability. Here, the severe effect of cyber-attack on automatic generation control is analyzed. The closed-loop transfer function is shown in Eq. 3, and the block diagram of a simple AGC presented in Figure 6, consisted by some parameters mentioned in Table I. The closed loop transfer function for AGC loop as,

$$\frac{\Delta\Omega(s)}{-\Delta P_L(s)} = \frac{(1+\tau_g s)(1+\tau_T s)}{(2H_s+D)(1+\tau_g s)(1+\tau_T s)+K_I+\frac{s}{R}} \quad (3)$$

It is known that an attacker can change the values of vulnerable parameters by increasing (positive biased cyber-attack) or

decreasing (negatively biased cyber-attack) them during data manipulation [8], [13]-[14] . An attacker can operate them simultaneously or, individually. According to Equation (3) and Figure 6, the various curves for frequency deviation depending different K_I values are presented in Figure7. The appropriate K_I value equals to 9 for the zero frequency deviation is determined here.

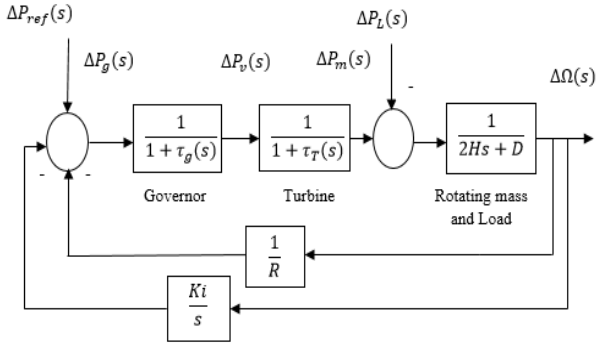


Figure 6: Isolated power system control block diagram of typical AGC

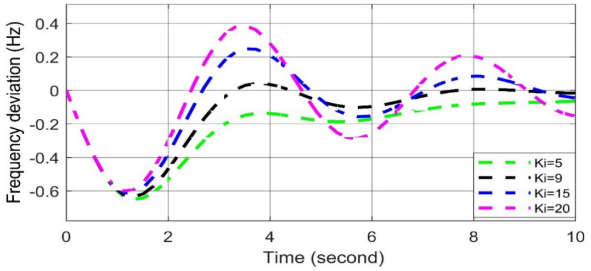


Figure 7: Frequency deviation response to different *Ki* values in AGC in case of cyber-attack.

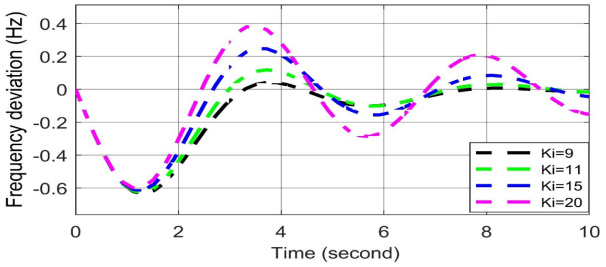


Figure 8: Frequency deviation response for positive biased cyber-attack in AGC.

2.2.1 *Positively biased cyber-attack on AGC*

According to the result of the cyber-attack on AGC system, the value of K_I increases from the set value, which is called positively biased cyber-attack. AGC system can lead to frequency deviation due to malicious data access. Hence, the frequency deviation of the system oscillated and became unstable. A positive biased cyber-attack's impact on the AGC is presented in Figure 8. It is observed that the frequency oscillation curve can return quickly to the stable state when the set value of $K_I = 9$. On the other hand, the frequency deviation curve oscillates slightly when $K_I = 11$. The frequency deviation curves oscillate more than the previous ones while $K_I = 15$ and 20, respectively. The reason is that the governor cannot restore the system's frequency, which ultimately disrupts the system's stability. Thus, it can be concluded that the positive biased cyber-attack severely affects the system's frequency.

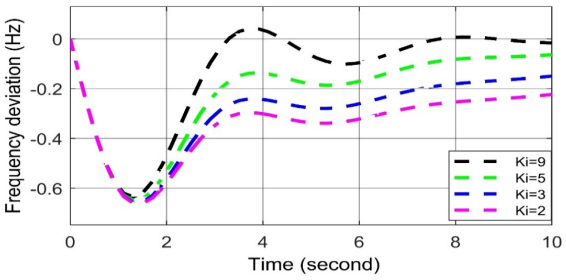


Figure 9: Frequency deviation response for negative biased cyber-attack in AGC.

2.2.2 *Positively biased cyber-attack on AGC*

According to the result of the cyber-attack on AGC system, the value of K_I increases from the set value, which is called positively biased cyber-attack. AGC system can lead to the frequency deviation due to malicious data access. Hence, the frequency deviation of the system is oscillated and become unstable. The impact of positive biased cyber-attack on the AGC presented in Figure 8. It is observed that the frequency oscillation curve can return quickly to the stable state when the set value of $K_I = 9$. On the other hand, the frequency deviation curve oscillates slightly when $K_I = 11$. The frequency deviation curves oscillate more than the previous ones while $K_I = 15$ and 20, respectively. The reason is that the governor cannot restore the system's frequency, which ultimately disrupts the system's stability. Thus, it can be concluded that the positive biased cyber-attack severely affects the system's frequency.

2.2.3 *Negative biased cyber-attack on AGC*

According to the result of the cyber-attack on AGC system, the value of K_I decreases from the set value, which is called negatively biased cyber-attack presented in Figure 9. It is observed that because of the negative biased attack, frequency is deviated from the set value (i.e., $K_i = 9$) and an unwanted delay in restoring system frequency. As a result, the negative biased attack disables the integral controller and opposes the purpose of using the K_I. Negatively biased cyber-attack less serious than positively biased cyber-attack for frequency disturbance because it does not oscillate the nature of the frequency fluctuation.

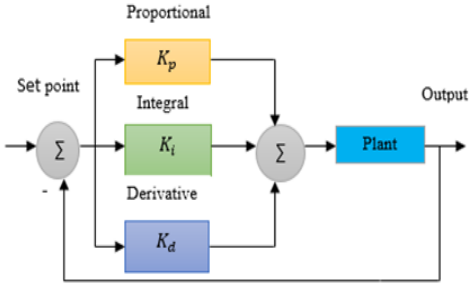


Figure 10: Block diagram of PID controller

2.3 *Modelling of PID controller*

In industrial applications, PID controller is the most popular approached feed-back controller. It is used to stabilize power systems, provide better frequencies, and reduce errors in the steady state. In order to evaluate the PID controller output, three different gain parameters are used in this controller, such as, the

proportional gain, integral gain, and derivative gains are used [15]-[16],exhibited in Eq. (4). The conventional PID controller structure and parameter values presented in Figure 10 and Table 2.

$$C(S) = K_p + \frac{K_i}{S} + K_d S \quad (4)$$

2.3.1 Cyber-Attack Impacts on AGC-PID

The stability of the proposed typical AGC model can be updated using a PID controller for better response. The isolated power system block diagram based on AGC-PID shown in Fig 11 and the parameters used in this model mentioned in Tables I and II. It is optimistic that cyber-attacks on AGC-PID may disrupt the stability of the frequency operation. For a suitable transient response, the PID controller gain (i.e., K_p, K_i and K_d) must be adjusted with proper tuning [2], [15]-[17]. Because of the failure of selecting the proper PID controller gain, the governor is unable to restore the frequency of the system. Thus, it is important to determine the values of these parameters optimally for proper operation. That's why, PID controller gain (i.e., K_p, K_i, and K_d) considers as the vulnerable quantity during cyber-attacks. An attacker can change the values of vulnerable parameters by increasing (i.e., positive biased cyber-attack) or, decreasing (i.e., negatively biased cyber-attack) them during data manipulation [8], [13]-[14].

Table 2: Tuning Parameters of AGC-PID controller

Symbols	Parameters	Values
K_p	Proportional gain	90.58
K_i	Integral gain	70.89
K_d	Derivative Gain	57.80

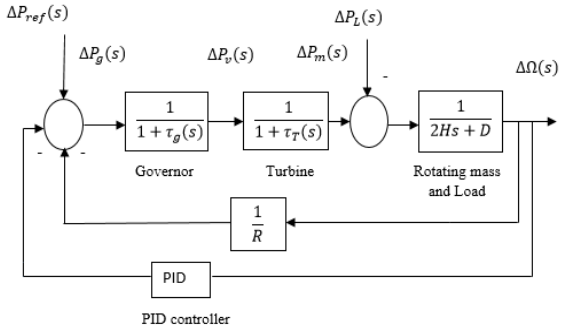


Figure 11: Block diagram of Power system control based on AGC-PID

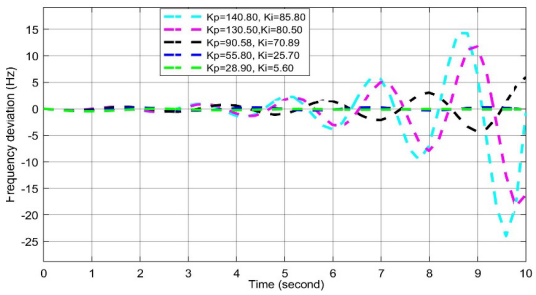


Figure 12: Frequency deviation response to different K_p and K_i values in AGC-PID in case of cyber-attack

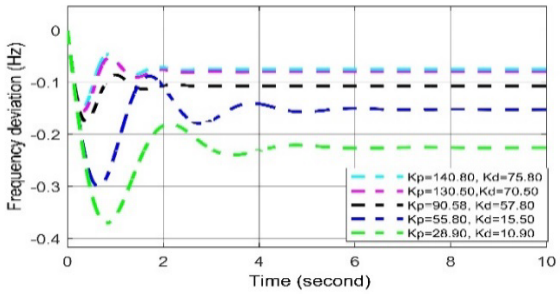


Figure 13: Frequency deviation response to different K_p and K_d values in AGC-PID in case of cyber-attack

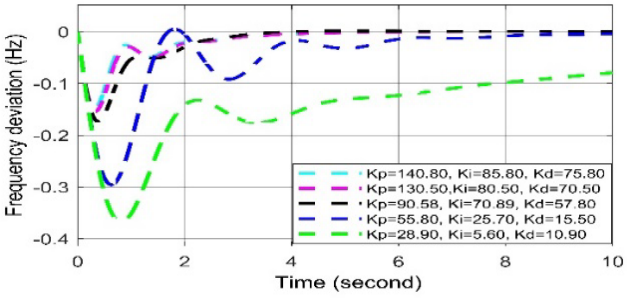


Figure 14: Frequency deviation response to different K_p, K_i and K_d *(i.e., PID)* values in AGC-PID in case of cyber-attack.

For positive biased cyber-attack, in Figs. (12-14), it is observed that the frequency oscillation curve can return quickly to the stable state in all cases when the set values of K_p= 90.58, K_i = 70.89, K_d = 57.80. When any two values among the three are being increased, the overshoot and settling time response increase. However, an excessive deviation in system frequency is not permitted here since the governor's excessive deviation cannot restore the system's frequency.

In case of negative biased cyber-attack, it is observed in Figs. (12-14) In all cases, the frequency oscillation curves are being oscillated more than the curves of set values (i.e., K_p= 90.58, K_i = 70.89, K_d = 57.80). Thus, it can be said that the negatively biased cyber-attack more serious than positively biased cyber-attack.

2.4. Cyber-Attack Impacts on AVR

The generator uses AVR as the primary means of controlling reactive power. The AVR control system considers the amplifier, exciter, generator, and sensor dynamics. The error signal is amplified and used to change the exciter terminal voltage by controlling the exciter field. The generated EMF changes by changing the current of the voltage generator at the exciter terminal. Thus, the reactive power is modified to keep the voltage stable [12]. It is known that cyber-attacks on the AVR is able to disrupt the stability of the voltage operation that may damage the conductor insulations and home appliances. Therefore, it is essential to consider the impact of a cyber-attack on AVR system in order to avoid such difficulty. The block diagram of AVR presented in Figure 15, where the amplifier gain (K_A) is determined using Routh-Hurwitz array. The integrated system parameters for AVR are mentioned in Table III. The closed loop transfer function for AVR loop is as,

$$\frac{V_t(s)}{V_{ref}(s)} = \frac{K_A K_E K_G K_R (1+\tau_R S)}{(1+\tau_A S)(1+\tau_E S)(1+\tau_G S)(1+\tau_R S)+K_A K_E K_G K_R} \quad (5)$$

Figure 16 shows the terminal voltage disruptions during cyber-attack for various K_A values. It is observed that because of the cyber-attack on AVR, any changes in K_A, the voltage fluctuations are obtained that disrupt the stability of the system. It is clear that, depending on the affected value of K_A, cyber-attacks can be categorized into two ways: positively biased cyber-attack (i.e., increasing K_A) and negatively biased cyber-attack (i.e., decreasing K_A).

Table 3: Parameters of Proposed AVR System model

Parameters	Symbol and value	Time constant
Amplifier	K_A=10	τ_A=0.1s
Exciter	K_E=1	τ_E=0.4s
Generator	K_G=1	$\tau_G = 1s$
Sensor	K_R=1	$\tau_R = 0.05s$

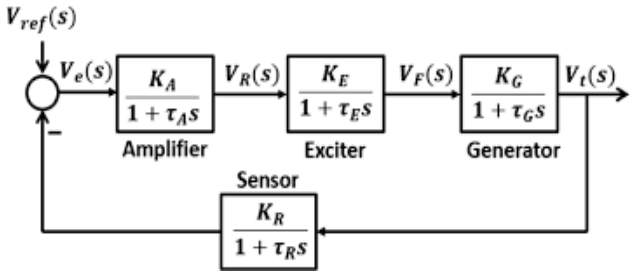


Figure 15: Power system control block diagram based on AVR, adapted from [12]

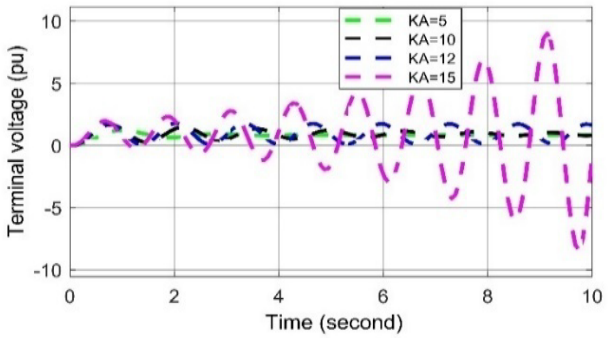


Figure 16: Voltage deviation response to different K_A values in AVR in case of cyber-attack

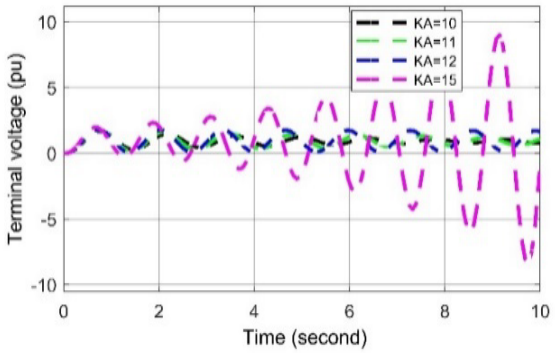


Figure 17: Voltage deviation response for positive biased cyber-attack in AVR

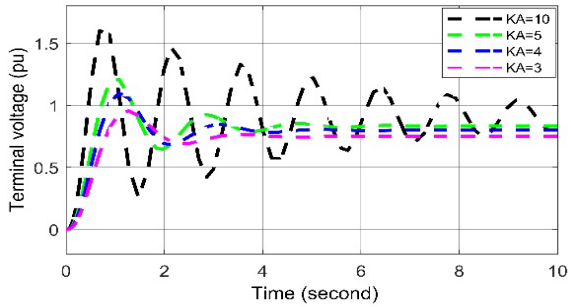


Figure 18: Voltage deviation response for negative biased cyber-attack in AVR

2.4.1 Positively biased cyber-attack on AVR

According to the result of the cyber-attack on AVR system, the value of K_A increases from the set value, which is called positively biased cyber-attack. When the cyber-attack is performed on AVR, the system's terminal voltage oscillates and becomes unstable. In the impact of positive biased cyber-attack on AVR presented in Figure 17. It is observed that the voltage fluctuation curve is oscillated in nature when the set value of K_A = 10. On the other hand, voltage fluctuation curves are slightly oscillated while K_A = 11 and 12, respectively. Furthermore, serious voltage fluctuation is occurred in case of K_A = 15. The reason is that because of not selecting the proper values of K_A, the exciter is unable to restore the voltage of the system. Thus, it can be concluded that the positive biased cyber-attack on AVR severely affects voltage disturbances as oscillating in nature.

2.4.2. Negatively biased cyber-attack on AVR

According to the result of the cyber-attack on AVR system, the value of K_A decreases from the set value, which is called negatively biased cyber-attack presented in Figure 18. It is observed that because of the negative biased attack, voltage is deviated than nominal value (i.e., K_A = 10) and an unwanted delay in the terminal voltage. As a result, the negative biased attack disables the amplifier gain and opposes the purpose of using the K_A. Negatively biased cyber-attack less serious than positively biased cyber-attack for voltage disruptions but both are oscillate in nature.

2.5. Cyber-Attack Impacts on AVR-PID

The stability of the proposed AVR model can be updated using a PID controller for better response. The reason is that the normal AVR is suffered by two limitations: (i) availability of long-term oscillation response and (ii) larger steady-state error. PID controller has three advantages: (i) minimizing the steady-state error, (ii) decreasing the settling time, and (iii) reducing oscillation and overshoot [12]. However, during the cyber-attack on AVR-PID, for a suitable transient response, the PID controller gain (i.e., K_p, K_i and K_d) must be adjusted with proper tuning [17]-[18]. Because of the failure to select the proper PID controller gain, the exciter cannot restore the voltage of the system. Thus, it is important to determine the values of these parameters optimally for proper operation [17]. That's why PID controller gain (i.e., K_p, K_i, and K_d) considers as a vulnerable quantity during cyber-attacks. An attacker can change the values of vulnerable parameters by increasing (i.e., positive biased cyber-attack) or, decreasing (i.e., negatively biased cyber-attack) them during data manipulation [8],

[13]-[14]. Figure 19 shows the block diagram of a simple AVR-PID, where some parameters used presented in Tables II and III.

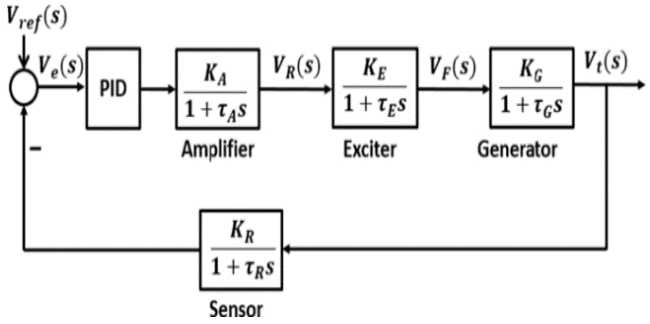


Figure 19: Power system control block diagram based on AVR-PID, adapted from [12]

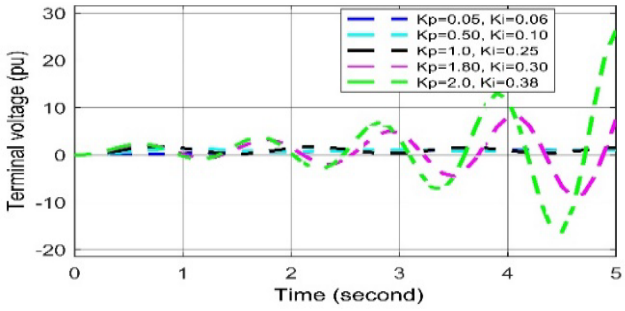


Figure 20: Voltage deviation response to different K_p and K_i values in AVR-PID in case of cyber-attack.

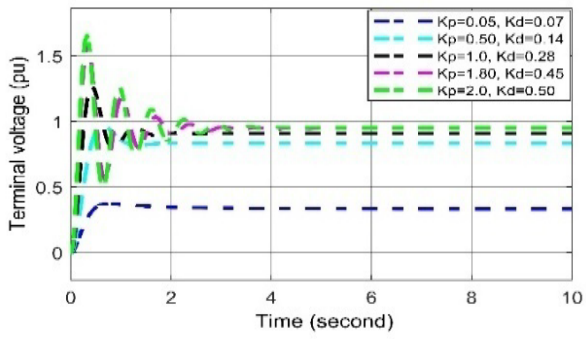


Figure 21: Voltage deviation response to different K_p and $K_d$values in AVR-PID in case of cyber-attack

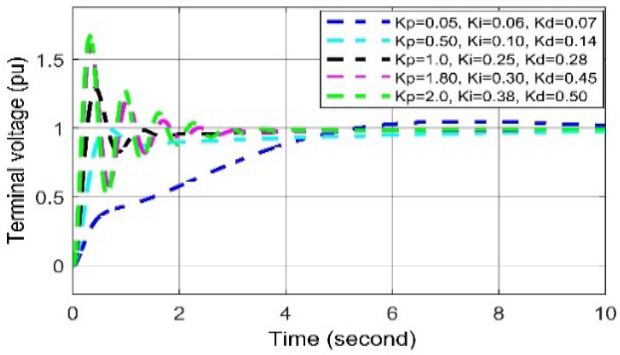


Figure 22: Voltage deviation response to different K_p, K_i and K_d values in AVR-PID in case of cyber-attack

In case of positive biased cyber-attack, in Figs. (20-22), it is observed that the voltage oscillation curve can return quickly to the stable state in all cases when the set values of K_p= 1.0, K_i = 0.25, K_d = 0.28. When any two values among the three are being increased the overshoot and settling time response increasing and in order to, an unwanted delay in restoring system voltage.

For negative biased cyber-attack, it is observed in Figs. (20-22), In all cases, the voltage oscillation curves are being oscillated more than the curves of set values (i.e., Kp= 1.0, Ki = 0.25, Kd = 0.28.), which can create voltage disturbance, which disrupts the system stability. Thus, it can be said that the negatively biased cyber-attack more serious than positively biased cyber-attack, where both forms of attacks are oscillated in nature.

3. Conclusion

The impact of cyber-attack on power systems in terms of frequency disturbances and voltage disruptions during changing loads has been investigated in this paper. In this regard, five individual controllers (i.e., LFC, AGC, AGC-PID, AVR, and AVR-PID) are introduced and incorporate two sets of system models.

In order to investigate the impact analysis properly, a series of experiments were conducted using the MATLAB-Simulink tool. The simulation results were obtained depending on the nature of the cyber-attacks shown according to the positive and negative biased cyber-attacks. It was observed that in LFC, a negative biased cyber-attack is more severe than a positive biased attack for frequency disturbance (see Figures 3-5), whereas, in AGC, a positive biased cyber-attack is more severe than a negative biased attack for frequency disturbances (see Figures 7-9). Furthermore, in AGC-PID, negative-biased cyber-attacks are more severe than positive-biased attacks for frequency disturbances (see Figures 12-14). In contrast, regarding AVR, negative-biased cyber-attack has less impact than positive-biased cyber-attack for stability in voltage disruptions, but both are oscillating in nature (see Figures. 16-18). In addition, a negative biased cyber-attack is more severe than a positive biased cyber-attack for AVR-PID, although both forms of cyber-attack oscillate in nature (see Figures 20-22).

The simulation results also confirm that, along with LFC and AGC, AVR is significant for ensuring the stability of the power system while changing the load. Finally, it can be concluded that our proposed ICAPS model is very effective in identifying the most severe attacks on the isolated power system.

In the future, we desire to achieve a reliable control system of cyber protection against unwanted access and reduce the severity of cyber-attack by considering the self-healing effectiveness process in the proposed ICAPS.

Conflict of Interest

The authors declare no conflict of interest

References

[1] M. Mohan, N. Meskin, and H. Mehrjerdi, "A comprehensive review of the cyber-attacks and cyber-security on load frequency control of power systems," Energies, **13**(15), 1–33, 2020, doi: 10.3390/en13153860.

[2] M. M. Uddin and M. M. Kabir, "Reduction of Frequency Disruption during Cyber-Attack in the Power System," 2nd International Conference on Sustainable Technologies for Industry 4.0, STI, 19–20, 2020, doi: 10.1109/STI50764.2020.9350518.

[3] C. Chen, M. Cui, X. Wang, K. Zhang, and S. Yin, "An investigation of coordinated attack on load frequency control," IEEE Access, **6**, 30414–30423, 2018, doi: 10.1109/ACCESS.2018.2845300.

[4] A. Farraj, E. Hammad and D. Kundur, "On using distributed control schemes to mitigate switching attacks in smart grids," IEEE 28th Canadian Conference on Electrical and Computer Engineering (CCECE), Halifax, Canada, 1578-1582, 2015.

[5] A. K. Farraj, E. M. Hammad, A. Al Daoud, and D. Kundur, "A game-theoretic control approach to mitigate cyber switching attacks in Smart Grid systems," IEEE International Conference on Smart Grid Communications, 958-963, 2014 , doi:10.1109/SmartGridComm.2014.7007772.

[6] D. Kirschen and F. Bouffard, "Keep the Lights On and the Information Flowing," IEEE Power and Energy Magazine, **7**(1), 55–60, 2009, doi: 10.1109/MPE.2008.930656.

[7] J. Yan, C. C. Liu, and M. Govindarasu, "Cyber intrusion of wind farm SCADA system and its impact analysis," IEEE/PES Power Systems Conference and Exposition, PSCE, 1–6, 2011, doi: 10.1109/PSCE.2011.5772593.

[8] S. Sridhar and M. Govindarasu, "Model-based attack detection and mitigation for automatic generation control," IEEE Transactions on Smart Grid, **5**(2), 580–591, 2014, doi: 10.1109/TSG.2014.2298195.

[9] M. Vrakopoulou, P. M. Esfahani, K. Margellos, J. Lygeros, and G. Andersson, "Cyber-attacks in the automatic generation control," Power Systems,**79**, 303–328, 2015, doi: 10.1007/978-3-662-45928-7_11.

[10] A. Dagoumas, "Assessing the impact of cybersecurity attacks on power systems," Energies, **12**(4), 2019, doi: 10.3390/en12040725.

[11] O. I. Elgerd and C. E. Fosha, "Optimum Megawatt-Frequency Control of Multiarea Electric Energy Systems," IEEE Transactions On Power Apparatus and Systems, **89**, 556-563, 1970.

[12] H. Saadat, Power system analysis, Mcgraw-Hill, 1999.

[13] S. Biswas and A. Sarwat, "Vulnerabilities in two-area automatic generation control systems under cyberattack," Proceedings - 2016 Resilience Week, RWS, 40–45, 2016, doi: 10.1109/RWEEK.2016.7573304.

[14] S. Adepu, N. K. Kandasamy, J. Zhou, and A. Mathur, "Attacks on smart grid: power supply interruption and malicious power generation," International Journal of Information Security, **19**(2), 189–211, 2020, doi: 10.1007/s10207-019-00452-z.

[15] M. M. Uddin, M. K. Saifullah, and M. M. Kabir, "PID Controller Based Automatic Generation Control for Three Area Interconnected Power System," International Conference on Information and Communication Technology for Sustainable Development, (ICICT4SD), 300–305, 2021, doi: 10.1109/ICICT4SD50815.2021.9396971.

[16] R. P. Borase, D. K. Maghade, S. Y. Sondkar, and S. N. Pawar, "A review of PID control, tuning methods and applications," International Journal of Dynamics and Control, 2020, doi: 10.1007/s40435-020-00665-4.

[17] G. Singh and J. K. Dhami, "Load Frequency and Voltage Control of Two Area Interconnected Power," International Journal of Engineering Research & Technology (IJERT), **8** (15), 1–4, 2016.

[18] M. Nagendra and M. Krishnarayalu, "PID Controller Tuning using Simulink for Multi Area Power Systems," International Journal of Engineering Research & Technology (IJERT), **1**(7), 1–9, 2012.

Appendix A: List Of Abbreviations

Abbreviation	Complete Meaning
ICA	Impact of Cyber-Attacks
PS	Power System
ICAPS	Impact of Cyber-Attacks on Power System
LFC	Load frequency control
AGC	Automatic generation control
PID	Proportional, integral and derivative controller
AGC-PID	Automatic generation control-PID
AVR	Automatic voltage regulator
AVR-PID	Automatic voltage regulator-PID

Appendix B: All Simulink Models

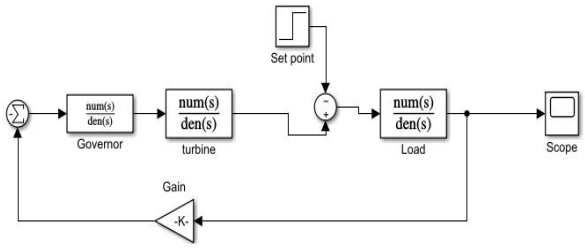


Figure 23: Simulink block diagram of LFC

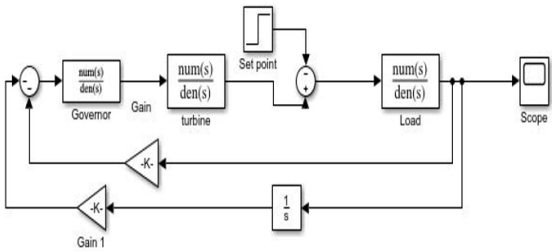


Figure 24: Simulink block diagram of AGC

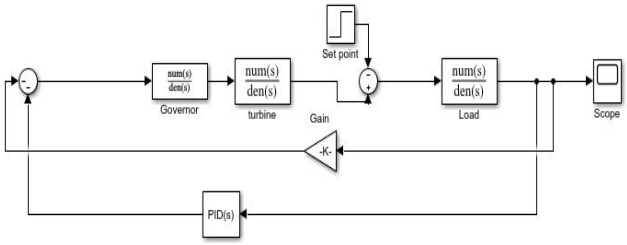


Figure 25: Simulink block diagram of AGC-PID

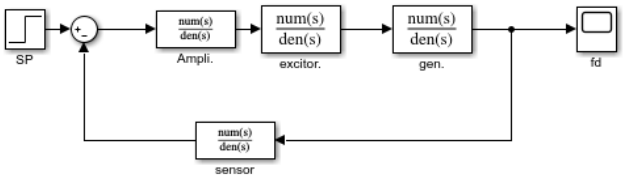


Figure 26: Simulink block diagram of AVR

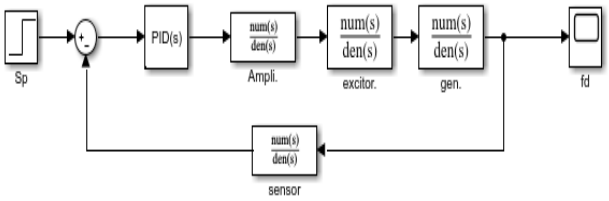


Figure 27: Simulink block diagram of AVR-PID

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7 No. 4, 28-33 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Assessment of Electromagnetic-Based Sensing Modalities for Red Palm Weevil Detection in Palm Trees

Mohammed M. Bait-Suwailam[*,1,2], Nassr Al-Nassri[1], and Fahd Al-Khanbashi[1]

[1]*Department of ECE, Sultan Qaboos University, Muscat, P.C. 123, Oman*

[2]*Remote Sensing and GIS Research Center, Sultan Qaboos University, Muscat, P.C. 123, Oman*

ARTICLE INFO

Article history:
Received: 02 June, 2022
Accepted: 12 July, 2022
Online: 23 July, 2022

Keywords:
Date palm
Microwaves
Red palm weevil

ABSTRACT

In this paper, we investigate the utilization of three effective detection methods to identify potential threats of insects in date palm trees. The detection techniques presented here are the application of infrared radiation, microwave antennas and metamaterials based sensors. Experimental trials using IR radiation took place in a local farm. Moreover, the second sensing system is based on microwave antennas that are designed and numerically simulated at the 2.45 GHz-band. Lastly, the third detection method focuses on the design and development of low-power microwave sensor based on metamaterials concept. Based on the processed and analyzed results, the aforementioned sensing techniques are able to predict existence of red palm weevils within date palms.

1 Introduction

Food security and production are considered very essential infrastructures that have great impact on the way we consume food and dietary products towards human well-being and health. In principle, food systems include all stages of food production, including harvesting, processing, distribution and storage. It is understood that the quality of food products can be compromised or even degraded when farmers/supply chains are not taken safety measures and food quality standards into consideration. For instance, the use of an improper storage spaces for food or the use of unhealthy measures for food harvesting and production [1]-[3].

One of food quality degradation can be attributed to the surrounding environment. In real-life, growth and safety of crops is directly related to the environment. For instance, date palm trees are attacked heavily by many insects.

Oman relies heavily on the production of dates, which helps quite well in the development of the country's economy. The overall area of dates trees in Oman is over 30,000 hectares [4]-[6]. Currently, Oman has carried out numerous initiatives and projects over the Sultanate area in order to enhance and maximize the yield from such an important agricultural crop in Oman. Unfortunately, numerous palm pests deeply affect the health and safety of such trees, among which red palm weevil is a highly risk factor [7]-[8].

The life cycle of RPW is summarized in the diagram shown in Fig. 1, where RPW adults could last between 14 to 20 days. It is not an easy task to visualize by human eyes occurrence of an early stage infections within date palms, which is due to the hidden larvae.

Potential signs of infection, include very slow production rate of date palms, appearance of yellow sticky paste in palms (see Fig. 2), and appearance of apertures on the trunk [9]. It can be observed that infections in palm trees extend up to almost 2 meters high from soil level.

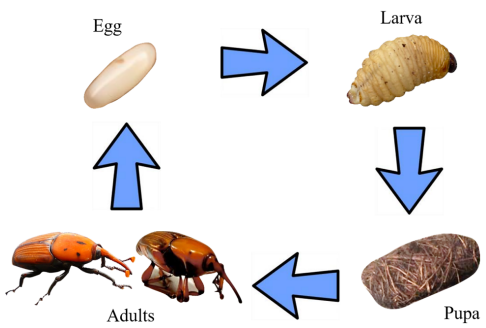


Figure 1: A conceptual chart showing red palm weevil growth cycle.

[*]Corresponding Author: Mohammed M. Bait-Suwailam (email: msuwailem@squ.edu.om)



Figure 2: An image from a local date palm farm, where waste and watery liquid can be seen as a clear symptom of RPW existence in palm trees.

Microwave-based sensing and detection technique is mature and well-known for its non-destructive and hygienic testing modality for food sensing, evaluation and analysis, with more emphasis on the medical and industrial sectors [10], [11]. In principle, microwave sensing technology takes advantage of electromagnetic radiation within microwaves band, ranging from 300 MHz - 30 GHz. There have been extensive research studies focusing on the development of microwave sensing and imaging systems for a large number of applications, including but not limited to healthcare and biomedical treatments, industrial, scientific and security needs.

There are a number of techniques in the literature aiming to mitigation effects of infestation in date palms, with the objective of providing an early detection mechanism for possible infestations. The detection methods include but not limited to the use of trained dogs [12], sound-based sensors for monitoring palm pests activities [13], utilization of microwaves and high power systems treatment [14], [15], deployment of remote sensing and geographic information systems [16]–[18] and development of transmission lines based resonators [19]. The research work in [13] presented an experimental trials using acoustic signals to capture RPWs' activities through pizoelectric sensing device, where sound signals were recorded in-situ and then digitally processed. The application of microwave heating treatment was introduced in many studies, for instance the research work in [14], [15] and references therein, with the objective of treating the infected trees using high-power microwaves. Although earlier detection techniques based on acoustics, fiber optics, or even high-power microwave illumination are effective, they suffer from the complexity in the design and setup needed in monitoring a large number of RPW infestation or the need of deploying high-power radiators that are harmful and costly, thus limiting their practical use by farmers. The utilization of low-power sensing alternatives is expected to be more attractive and cost effective, which this research work aims to explore a number of alternatives based on low-power microwave sensing modalities.

In [20], a comparison between two sensing modalities was presented, namely: microwave-based sensory antennas and infrared-based sensing. However, the work presented here develops a more comprehensive study covering a number of attractive sensing modalities for the detection of RPW within date palm trees with more emphasis on the application of microwaves using antennas and highly-sensitive sensors based on metamaterials concept. In the remaining parts of the article, additional numerical and experimental results are presented and discussed.

2 Potential sensing and detection techniques for RPW detection

This section covers a comprehensive study of potential yet effective measures for sensing RPW in date palms, namely: 1) deployment of infrared-based sensing systems, 2) integrating microwave sensors (antennas) with a processing unit, and 3) deployment of microwave-based metamaterial sensors.

2.1 RPW sensing using Infrared-based imagery data

An infrared imaging is a very attractive technique, which utilizes radiation to identify distinctive features of temperature distribution. Nowadays, IR-based sensors are embedded in many electronic devices. This makes it easy to deploy IR in many applications, including detection of undesirable insects in food products.

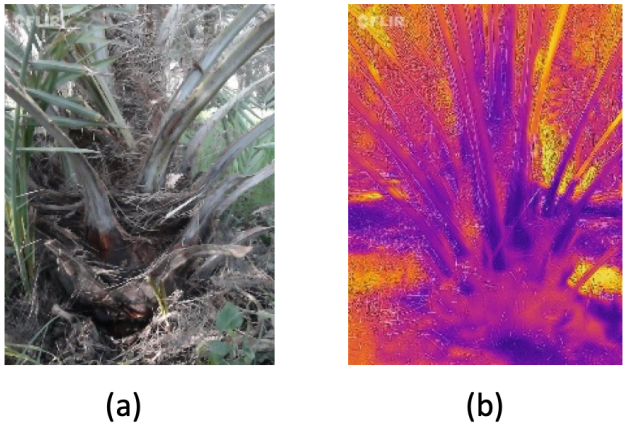


Figure 3: (a) Photo of a normal date tree and (b) the thermal heat distribution of the same palm tree by the IR sensor.

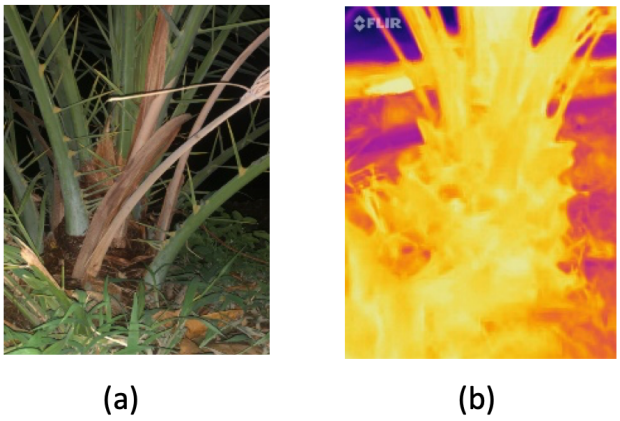


Figure 4: (a) Photo of RPW infestation case at night time and (b) the thermal heat distribution of the same palm tree by the IR sensor.

Since IR radiation has longer wavelength than visible band, thus it makes it difficult for humans to visualize IR thermal distribution. We can then utilize infrared sensors to inspect thermal distribution in palm trunk to detect any activities from the red palm weevil. IR

images from healthy palm trunk were also recorded and processed for comparison purposes. Fig. 3 shows the temperature distribution within the trunk. As shown, the thermal heat distribution from the trunk was not significant enough, from which we could conclude that the palm is healthy. The heat distribution for the case of unhealthy trunk can be seen in Fig. 4 during night time, where more heat is expected from the trunk body itself.

Finally, the larva of the RPW was extracted. Interestingly, it was gluey. Additionally, it was observed that sufficient high radiated thermal energy was generated by the RPW and identified as yellow color from the sensor, as in Fig. 5(b).

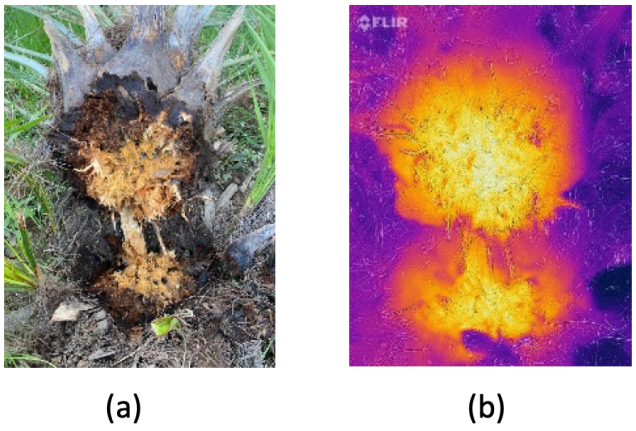


Figure 5: (a) Photo illustrating waste from infected palm trunk and (b) the thermal heat distribution of the same trunk by the IR sensor.

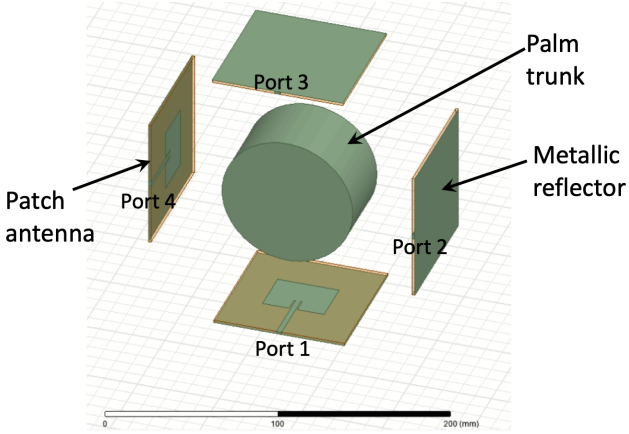


Figure 6: The developed model of the trunk surrounded with four microstrip antennas.

2.2 RPW sensing using microwave antennas

We investigate next the application of microwave-based sensors for red palm weevils detection. We have numerically modeled and simulated a small-scaled model comprising an unhealthy date palm and within its body a modeled pest. For ease of microwave sensing development, microstrip patch antennas are used here as the elementary sensors in order to predict any abnormalities within the coupled energy between the antenna elements. The spatial distance separating the radiating elements is maintained at 80mm. The modeled trunk has a height of 44.7mm and a diameter of 80mm (its dielectric constant, ϵ_r = 31.5 and loss factor of 11.5). Since the solution to such 3D numerical model is mainly dependent on finite-element method of HFSS, storing the solution in each iterative process until convergence is reached would require huge amount of memory. To minimize the demands to such memory and disk space in the computer, a small-scale numerical model was constructed. Furthermore, the dimension of the RPW insect was modeled as a cylindrical object, with height = 2mm and radius = 3mm, with electric permittivity of 9.3 and bulk conductivity of 0.38 S/m. A single palm pest was placed in two different scenarios: within trunk's center and edge side of the palm, as illustrated in Figs. 7 (a)-(b). Note that all aforementioned constitutive parameters and associated losses were embedded into the models [15].

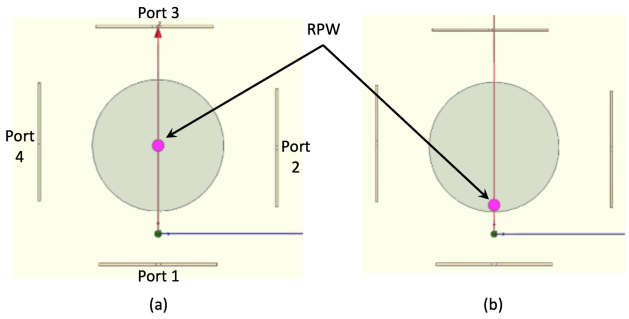


Figure 7: Top view of the trunk with a homogenized model mimicking RPW placed at two different scenarios, located at: (a) center, and (b) near to inner surface of the trunk.

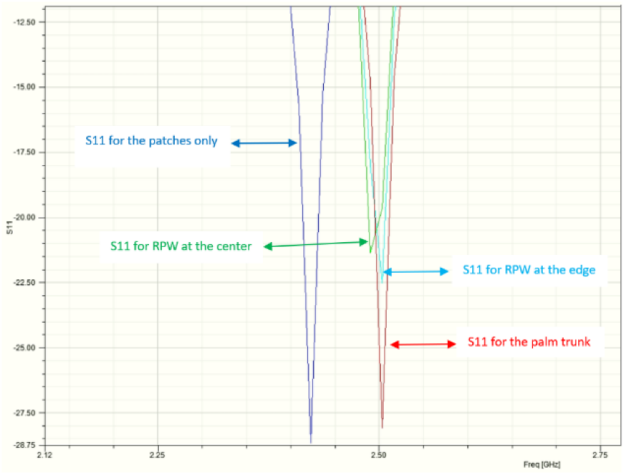


Figure 8: Simulated $|S\,11|$ for the healthy and infected palm trunk.

For comparison, a finite size structure consisting of the four patch antennas was considered. The case of the four patch antennas alone, i.e. no palm trunk, is expected to provide good impedance matching and low coupling coefficients between the individual antenna elements. The case of a finite-sized healthy trunk (modeled here as cylindrical lossy object) was numerically simulated, with the four antenna elements surrounding the trunk, as illustrated in Fig. 7. Good matching for the microwave antenna elements is still maintained as shown in Fig. 8. Furthermore, an appreciable shift

in the transmission coefficient between ports 1 and 3 ($|S\,13|$) was achieved, which is referred to the presence of RPW inside the palm trunk. It is worth mentioning here that no major frequency shift in the peak of $|S\,13|$ was observed when comparing the cases of RPW at center and at the edge, as shown in Fig. 9.

The transmitted energy strength from port 1 to 4 was also considered. The sudden peak movement of $|S\,14|$ for the unhealthy palm is expected, as depicted in Fig. 10. Furthermore, we note that transmission coefficient, $|S\,13|$, is stronger than that from ports 1 and 4, which is due to the strong mutual interaction from port 1 to 3 in comparison to the case of coupled energy from ports 1 and 4.

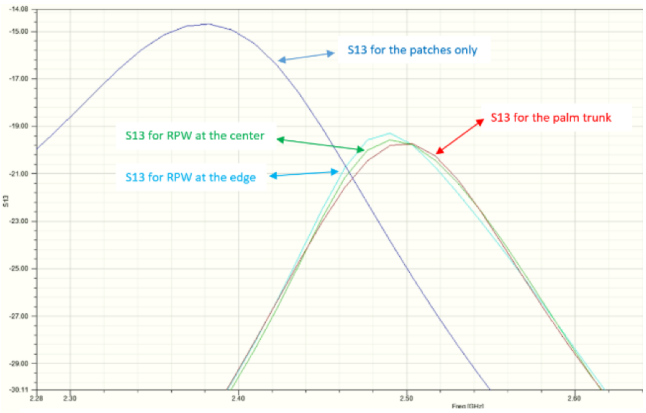


Figure 9: Simulated $|S\,13|$ for the healthy and infected palm trunk.

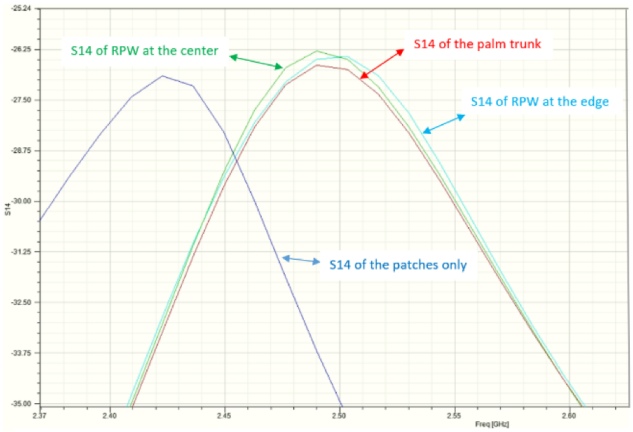


Figure 10: Simulated $|S\,14|$ for the healthy and infected palm trunk.

2.3 *RPW sensing using microwave TL-based metamaterial sensors*

In this section, we present a numerical study of an interesting sensing modality using microwave transmission line-based metamaterial sensors. Such sensors are engineered resonant particles or unit cells that have found lots of applications from engineering and science. Amongst the popular structures is the artificial magnetic conductor, also known as split-ring resonator (SRR). The dual of the SRR is the complementary split-ring resonator (CSRR), i.e., artificial electric conductor [21]. A schematic representation of the physical structure of the two inclusions can be seen in Fig. 11, where square rings are depicted for convenience.

In this study, a single unit cell is presented for ease of demonstration and compactness of the sensor. However, it is also permissible that more concentric or cascaded resonant units as a 1 dimensional array of rings could be explored. Details of the dimensions of the SRR/CSRR inclusions include length of the ring represented by L, a is the width of the SRR ring (or slot for CSRR), g is the cut gap and S_p is the spacing between the two rings. From the physical operational point of view, the SRR resonates well once excited through a normal external magnetic field component, while the CSRR resonates very well through an external normal electric field. Interestingly, the dual-operation mechanism of SRR/CSRR is based on the Babinet's principle, as presented in [21].

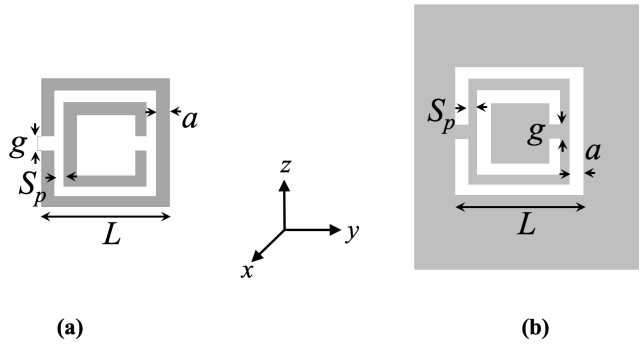


Figure 11: (a) the artificial magnetic conductor unit inclusion, and (b) its counterpart, the artificial electric conductor unit inclusion.

Fig. 12 depicts the developed microwave TL-based metamaterial structure. Two ports are deployed here in order to characterize the performance of the sensor through the computed transmission coefficient between the two ports. Once the sensing element is brought in close proximity to the date palm trunk, the resonance of the sensing element is expected to shift, and even further change is expected whenever RPW insects are within the inside of the trunk. The optimized dimensions of the developed CSRR microwave sensor at 2.45 GHz-band are available in Table. 2, along with the dimensions of the modeled 3D palm trunk.

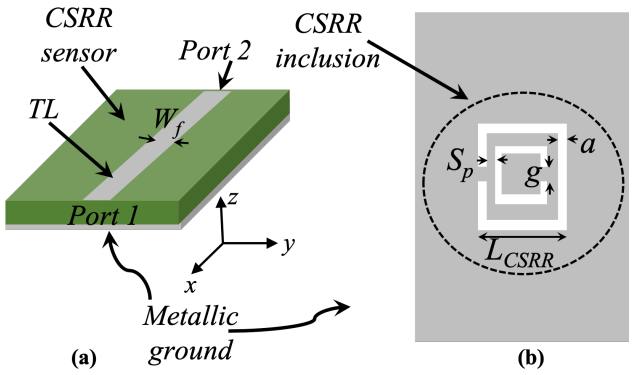


Figure 12: The developed TL-based metamaterial two-port sensor: (a) lateral view, and (b) bottom view of the metallic ground showing the adopted CSRR unit inclusion. The grey area here represents metallization.

Table 1: A comparison between several existing techniques suitable for RPW infestation in palm trees.

Techniques	sample required?	special set	Trained labor required?	Cost	Suitability
Sniffing dogs	required	No	Yes	Low	small (uncontrolled) farm
Acoustic	required	Yes	Yes	Moderate	small (controlled) farm
IR imaging	No	Yes	Yes	Moderate	large scale environment
microwave heating	No	Yes	Yes	High	High power radiation exposure
remote sensing	Yes	Yes	Yes	Moderate	small and large scale environment
microwave detection	Yes	Yes	Yes	Moderate	small (uncontrolled) farm

Table 2: The optimized parameters of the microwave detection structure.

Model parameters	Dimension(in mm)
Feedline width, W_f	2.52
Sensor thickness	0.8
Substrate length	40
Substrate width	30
Palm trunk diameter	50
Palm trunk height	40
Stand-off distance, S_z	2
L_{CSRR}	10.5
a	0.45
S_p	1.575
g	0.5

Fig. 13 shows a 3D view of the developed TL-metamaterial sensor that is placed on top of a date palm sample. For convenience, a stand-off distance, S_d, of 2mm was considered. Coupled energy via the sensor's ports is then computed and used as a metric to quantify the performance of the sensor in detecting RPW in date trees. Comparison was also made for the case of healthy palm trunk. Note that the resonance frequency of the reference sensor was 2.45 GHz. According to full-wave simulations, the close proximity of the sensor to a dry palm (representing dead one) resulted in a shift to the resonance frequency by 2.53% to lower frequencies due to the loss nature of the palm, while further increased shift to resonance frequency by 5.84% was incurred for the case of unhealthy (infected) palm tree.

Table. 1 provides a a comparison between several techniques that are available for the detection of infested date trees. It can be seen that each method has its own advantage and limitation, depending on various metrics.

3 Conclusion

In this paper, we presented a comparison study concerning the effectiveness of three detection techniques for red palm weevils in date palm trees. In the first detection method, detection of thermal variations within the date palm trunk was permissible using IR sensor. Experimental studies were performed in a small farm, containing a number of healthy and unhealthy palm trees. The experiments were carried out at different timings. Based on IR imagery data, quite noticeable heat was generated from the infested palm trees. A numerical model for the detection of RPW using four microstrip antenna elements working at 2.45 GHz was developed. Throughout the presented detection method using microwave antenna system and depending on the adopted number of ports, any sudden abnormalities in the response of either S11 or S21 magnitude and/or phase can be taken as an indicator for existence of RPWs in palm trunk.

Lastly, a 3D numerical model was developed for the case of deploying metamaterials-based microwave sensors. Based on the findings from this study, the two microwave-based detection models are considered as low-cost and effective tools for sensing the existence of RPWs.

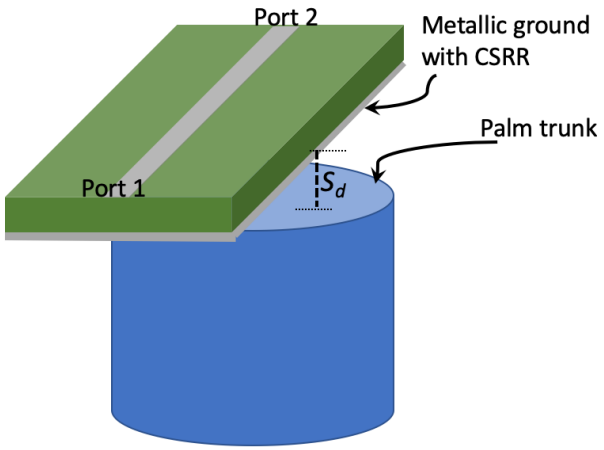


Figure 13: A 3D schematic showing the TL-based CSRR sensor in close proximity to a small-scale date palm trunk.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The authors would like to acknowledge the support of Sultan Qaboos University, Oman, and Omantel company for providing the financial resources.

References

[1] T. Bosona, G. Gebresenbet"Food traceability as an integral part of logistics management in food and agricultural supply chain," Food Control, **33**, 32–48, 2013, doi: 10.1016/j.foodcont.2013.02.004.

[2] K. G. Grunert, "Food quality and safety: consumer perception and demand," European Review of Agricultural Economics, **32**(3), 369–391, 2005, doi: 10.1093/eurrag/jbi011.

[3] J. Zhang, L. Liu, W. Mu, L. M. Moga, X. Zhang, "Development of temperature-managed traceability system for frozen and chilled food during storage and transportation," Journal of Food, Agriculture and Environment, **7**(3), 28–31, 2009.

[4] A.S. Al-Marshudi, "Oman traditional date palms: production and improvement of date palms in Oman," Tropiculture, **20**(4), 203–209, 2002.

[5] R. Al-Yahyai, M. Khan, Date Palm Status and Perspective in Oman, in: Al-Khayri, J., Jain, S., Johnson, D. (eds) Date Palm Genetic Resources and Utilization, 1st Edition, Springer, doi: 10.1007/978-94-017-9707-8-6.

[6] R. Al-Yahyai, "Improvement of date palm production in the Sultanate of Oman," Acta Hort, **736**, 337–343, 2007, doi: 10.17660/ActaHortic.2007.736.32.

[7] K. Azam, S. Razvi, I. Al-Mahmuli, "Survey of red palm weevil, (Rhynchophorus Ferrugineus Oliver) infestation in date palm in Oman," Iraqi date Palms Network, http://www.iraqi-datepalms.net [accessed 4th October 2021].

[8] K. Al-Kindi, P. Kwan, N. Andrew, M. Welch, "Impacts of human-related practices on Ommatissus lybicus infestations of date palm in Oman," PLoS ONE, **12**(2), 1-17, 2017, doi: 10.17605/OSF.IO/HXPYP.

[9] D. Kontodimas,V. Soroker,C. Pontikakos,P. Suma,L. Beaudoin-Ollivier,F. Karamaouna,P. Riolo,"Visual identification and characterization of Rhynchophorus Ferrugineus and Paysandisia archon infestation," in: Handbook of Major Palm Pests Biol. Management, 187-208, 2016, John Wiley and Sons, doi: 10.1002/9781119057468.ch9.

[10] E. Fear, S. Hagness, P. Meaney, M. Okoniewski, M. Stuchly, "Enhancing breast tumor detection with near-field imaging," IEEE Microwave Magazine, **3**(1), 48-56, 2002., doi: 10.1109/6668.990683.

[11] E. Fear, P. Meaney, M. Stuchly, "Microwaves for breast cancer detection?" IEEE Potentials, **22**(1), 12-18, 2003, doi: 10.1109/MP.2003.1180933.

[12] S. Salem, "Accuracy of trained dogs for early detection of red palm weevil, Rhynchophorus ferrugineus Oliv. infestations in date palm plantations," Swift Journal of Agric. Res., **1**(1), 1-4, 2015.

[13] A. Hertzroni, V. Soroker, Y. Cohen, "Toward practical acoustic red palm weevil detection," Computers and Electronics in Agriculture, **124**, 100-106, 2016, doi: 10.1016/j.compag.2016.03.018.

[14] S. Nelson, "Review and assessment of radio-frequency and microwave energy for stored-grain insect control," Transactions of the ASAE, **39**(4), 1475–1484, 1996, doi: 10.13031/2013.27641.

[15] R. Massa, G. Panariello, D. Pinchera, F. Schettino, E. Caprio, R. Griffo, M. Migliore "Experimental and numerical evaluations on palm microwave heating for Red Palm Weevil pest control," Scientific Reports, **7**, 45299, 2017, doi: 10.1038/srep45299.

[16] F. Marzukhi, M. Said, A. Ahmad, "Coconut Tree Stress Detection as an Indicator of Red Palm Weevil (RPW) Attack Using Sentinel Data," International Journal of Built Environment and Sustainability, **7**(3), 1-9, 2020, doi: 10.11113/ijbes.v7.n3.459.

[17] H. Kurdi, A. Al-Aldawsari, I. Al-Turaiki, A. Aldawood, "Early detection of red palm weevil rhynchophorus ferrugineus (olivier), infestation using data mining," Plants, **10**(1), 95, 2021, doi: 10.3390/plants10010095.

[18] D. Kagan, G. F.Alpert, M. Fire, "Automatic large scale detection of red palm weevil infestation using street view images," ISPRS Journal of Photogrammetry and Remote Sensing, **182**, 122-133, December 2021, doi: 10.1016/j.isprsjprs.2021.10.004.

[19] M.M. Bait-Suwailam, "Numerical Assessment of Red Palm Weevil Detection Mechanism in Palm Trees Using CSRR Microwave Sensors," Progress in Electromagnetic Research Letters, **100**, 63-71, 2021, doi: 10.2528/PIERL21080303.

[20] F. Al Khanbashi, Nassr Al Nassri, M. Bait-Suwailam, "A Comparative Study of Electromagnetic-Based Sensing Modalities for Red Palm Weevil Detection in Palm Trees," in: 3rd IEEE Middle East and North Africa COMMunications Conference (MENACOMM), 69-73, 2021, doi: 10.1109/MENACOMM50742.2021.9678240.

[21] F. Falcone, T. Lopetegi, M. A. G. Laso, J. D. Baena, J. Bonache, M. Beruete, R. Marqués, F. Martín, M. Sorolla, "Babinet Principle Applied to the Design of Metasurfaces and Metamaterials," Physical Review Letters , **93**(19), 197401, November 2004, doi: 10.1103/PhysRevLett.93.197401.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 31-38 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Regular Tessellation-Based Collective Movement for a Robot Swarm with Varying Densities, Scales, and Shapes

Kohei Yamagishi*, Tsuyoshi Suzuki

Tokyo Denki University, Department of Information and communication Engineering, Adachi-ku, 120-8551, Japan

ARTICLE INFO

Article history:
Received: 25 August, 2022
Accepted: 02 November, 2022
Online: 13 November, 2022

Keywords:
Multi-robot systems
Collective movement
Distributed control
Regular tessellations

ABSTRACT

In complex swarm robotic applications that perform different tasks such as transportation and observation, robot swarms should construct and maintain a formation to adapt and move as a single large-scale robot. For example, transportation and observation tasks require unique robot swarms with either high densities to support the weight of the transported objects or low densities to avoid overlapping field of views and avoid obstructions. Previous literature has not focused on structure optimization because swarming provides a large-collective capability. This paper proposes a leader-follower-controlled collective movement method by calculating direction and distance potentials between robots based on geometric constraints, constricting robot positioning along radial gradients around the leader robot according to these potentials. This paper demonstrates a robot swarm applying the proposed method while maintaining formations with different densities while moving and evaluates the robot swarm structure-maintaining performance.

1. Introduction

Cooperation between multiple robots is one method to complete large-scale and parallel tasks that a single robot cannot solve alone. Swarm robotic systems achieve this cooperative capability without centralized control by implementing swarm intelligence. In this system, a swarm of robots behaves in an organized manner through robot interactions [1]. When robots must complete a complex task that requires object transportation and tracking, they must move in a group or swarm. Various methods have been studied to determine swarm behaviors suitable for the task. Heavy object transportation requires a high-density robot swarm to support the weight. In contrast, the transport of light objects requires a low-density robot swarm to reduce the number of operating robots. In addition, monitoring dynamic targets, such as a school of fish, requires a low-density robot swarm to expand and optimize the observation range. Swarm robotics requires adaptive swarm behaviors that control robots to achieve different densities and shapes.

Swarm behavior control for object transportation robots includes collective movements in which the robot swarm moves while maintaining either a pre-organized shape in a two-dimensional [2] or three-dimensional space [3], or a flock, in which robots move while changing their arrangement to adapt to the velocities and positions of other robots [4, 5]. These approaches typically control the robots using local interactions based on inter-robot relative distances and orientations [6], global interactions based on a virtual leader robot [7], environmental geometry, and the positions of the robots [8]. Local interactions distribute control based on the states of neighboring robots; formations determined in this way must place robots at intervals in which they recognize their neighbors. Robots can build a dense hexagonal close-packed structure to ensure that they do not deform, split, or break formation, while at the same time prioritizing swarming [9] or area cover [10]. Global interactions, on the other hand, require centralized control to share arbitrary information. Formations constructed using this method can freely position the robots individually. The number of controllable robots in this global-control scheme then depends on the communication performance of the robots and the computational performance of the centralized control unit. Hence, distributed control based on local information is advantageous when controlling a large-scale robot swarm. However, the formations constructed under this control method cannot provide density and area coverage suitable for many cooperative tasks with different densities. Therefore, we consider a method for constructing and moving a large-scale robot swarm with different densities, scales, and shapes to perform cooperative tasks. Cooperation between multiple robots is one method to complete large-scale and parallel tasks that a single robot cannot solve alone. Swarm robotic systems achieve this cooperative

*Corresponding Author: Kohei Yamagishi, 5 Senju Asahi-cho, Adachi-ku, Tokyo, Japan 120-8551, k.yamagishi@nrl.c.dendai.ac.jp

capability without centralized control by implementing swarm intelligence. In this system, a swarm of robots behaves in an organized manner through robot interactions [1]. When robots must complete a complex task that requires object transportation and tracking, they must move in a group or swarm. Various methods have been studied to determine swarm behaviors suitable for the task. Heavy object transportation requires a high-density robot swarm to support the weight. In contrast, the transport of light objects requires a low-density robot swarm to reduce the number of operating robots. In addition, monitoring dynamic targets, such as a school of fish, requires a low-density robot swarm to expand and optimize the observation range. Swarm robotics requires adaptive swarm behaviors that control robots to achieve different densities and shapes.

Figure 1 shows two types of structures based on connections between neighboring robots: crystalline structures, which arrange robots in a regular tessellation [11], and amorphous structures, which place robots in an irregular pattern suited to a given condition [12]. Comparing the density of each crystalline structure when packed by circular robots, the hexagonal lattice is the highest, and the honeycomb lattice is the lowest, depending on the number of neighbors. In crystalline structures, robots can maintain their construction with only the relative angles and distances between them, making the formations match the constructed lattice. In contrast, an amorphous structure has irregular angles between neighboring robots. Therefore, the swarm shape can meet evolving environmental requirements. Amorphous swarms also prevent splitting by reinforcing high-load sections during transportation. However, owing to these properties, amorphous structures cannot build a shape with high density. This study focuses on crystalline robotic formations to stably control swarms with different densities.

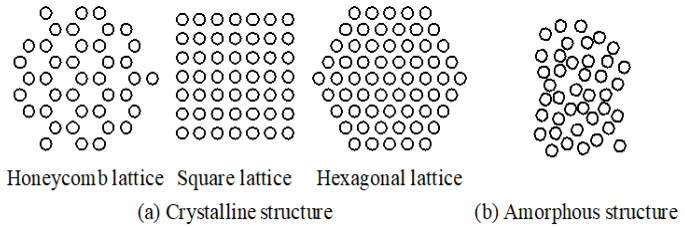


Figure 1: Examples of crystalline and amorphous structures

Regular tessellations provide latticed structures with high density by arranging the robots according to geometric constraints. The study of modular robots in lattice structures has facilitated the development of self-organizing robots with angular joints between neighbors in honeycomb, square, and hexagonal lattices [13, 14, 15]. These robots can self-organize various swarm shapes by recursively communicating with their neighbors, and can behave rigidly as they are physically interconnected with neighboring robots. However, these modular robots cannot organize other lattice formations owing to the physical constraints of their joint assemblies. In this study, we design a virtual interaction to construct lattice-based swarm shapes without the physical restriction of joints, such as the joint limits of modular robots. In our previous study, we proposed a collective movement using a distance-based local interaction based on the Lennard-Jones potential, an intermolecular potential [16]. This potential converges into a hexagonal close-packed structure, as described above. This study develops a group-forming model that adds an angular potential to this distance potential. For collective movement, the method introduces a leader-follower control, in which the robots orient with respect to a single leader robot in the group. Host systems and operators can then control the robot swarm by managing only the path of the leader. Therefore, the system does not require a large-capacity network to connect all robots. The above global-information-based approach can achieve collective movement without this physical leader by introducing a goal location and a virtual leader. For the host system to broadcast such information to the swarm, the system must either deploy an area network to communicate the information to all robots or equip them with a sensor device that can detect the direction of a physical goal, such as a light [17] or sound [18]. The robots then move autonomously toward the position indicated by the global information. Unfortunately, it is difficult for systems to manage the moving path of the whole swarm because the robots move autonomously. Therefore, this study considers the collective movement of a large-scale robot swarm using leader-follower control, controlling the swarm path in real-time by communicating with a single robot. We developed a collective movement method for a robot swarm with a density, scale, and shape suitable for cooperative tasks solved by robot swarms via teleoperation of a leader robot. This paper shows that a robot swarm applying the proposed method can move while maintaining various swarm shapes constructed with different densities.

The rest of this paper is structured as follows: Section 2 describes the configuration of the robot swarm and mathematical variables. Section 3 proposes a group-forming potential and collective movement method using leader-follower control. In Section 4, we discuss changing the robot swarm structures, and apply the changes during movements with different densities, sizes, and shapes. Finally, Section 5 summarizes the study.

2. Preliminaries

The robot swarm used in this study consists of two or more homogeneous mobile robots that may move within a linearly controllable speed v_{max} in all directions in a two-dimensional plane without any obstacles. The robot is equipped with both a distance sensor, such as light detection and ranging (LiDAR), to obtain the relative positions of surrounding neighbor robots and a local communication device, such as an infrared communicator, to exchange unique information with neighbors detected by the sensor. The position information is input to the proposed controller. As shown in Figure 2, the controller recognizes the neighbors to construct the lattice connected by the red lines among the communicable neighbors in all directions. The outer shape of the robots is circular with a diameter of σ [m] to prevent the occurrence of distance error due to robots measuring each other from all directions. All robots require a shared common reference direction to determine their neighboring angles and construct lattice structures; to provide this orientation, we initialize the robots with the same heading or synchronize directional information by averaging the data through local communication [19]. The i-th robot obtains the physical information of a neighbor robot with a distance-vector r_{ij} , composed of the distance $\|r_{ij}\|$ and the direction angle θ_{ij}, to the j-th neighboring robot.

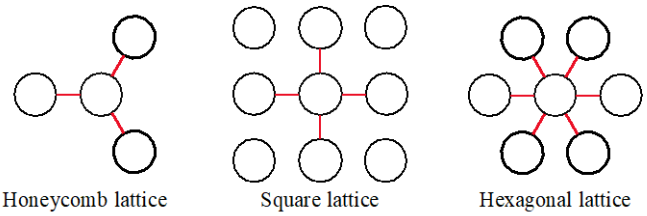


Figure 2: Neighbors that can communicate with the central robot in each latticed structure. The red line represents the connection with the neighboring robots equidistant from the central robot.

3. Multi-Robot Movement Control Method

3.1. Group-forming potential with latticed structures

Geometric swarm tessellations depend on the distance and angle between neighboring robots, treated as points. To maintain the structure of these points, distance and angle are used to compute and converge on a target structure. We propose a virtual potential-based interaction that depends on the relative position of neighboring robots to maintain cohesion.

Regarding the distance potential, we took inspiration from the interaction of molecules that organize molecular lattice structures. Among the various potential models that describe molecular bonds, the Lennard-Jones potential [20] is calculated with simple operations. With this potential, the repulsive and attractive forces diverge to infinity as the distance between molecules shrinks with an arbitrary maximum value determined by a coefficient past a certain distance depending on the molecular diameter. Considering the robot as a molecule, this study modifies the potential using the normalized coefficient adjusted to ensure that the maximum attractive force is one; the force on the relative distance, which is the first-order derivative of the Lennard-Jones potential, is expressed as follows:

$$f_i^D(r_{ij}) = \frac{169\sqrt[6]{\frac{13}{7}}}{63 \cdot 2^{\frac{5}{6}}} \frac{1}{\|r_{ij}\|} \left(12 \left(\frac{\sigma}{\|r_{ij}\|} \right)^{12} - 6 \left(\frac{\sigma}{\|r_{ij}\|} \right)^{6} \right) \hat{r}_{ij}. \quad (1)$$

The robot moves according to the calculated force, converging to a distance of $\|r_0\|$, where $f_i^{D\prime}(r_0) = 0$, depending on the diameter. The robots are controlled only by this potential move toward the balance distance between themselves and their neighbors, moving freely about a circle of radius $\|r_0\|$. Consequently, the robots organize into a close-packed hexagonal lattice structure. This study introduces an angular potential that converges in equally spaced directions, suitable for lattices with different densities. Robots constructing the lattice structure must place their neighbors in the angular intervals of $2\pi/3$, $\pi/2$, and $\pi/3$ [rad], divided by the number of connections in honeycomb, square, and hexagonal lattices, respectively, as shown in Figure 2. This study applies a simple angular potential model to converge robots into these lattices, as follows:

$$f_i^A(r_{ij}) = \sin(2L\,\theta_{ij}) R \left(\frac{\pi}{2} \right) \hat{r}_{ij} \quad (2)$$

where L is three, four, and six for honeycomb, square, and hexagonal lattices, respectively, matching the number of neighbors constructing the lattice structure. This model works normal to the relative distance vector between each pair of robots and can modify the angular error between robots within $\pm\pi/4L$ [rad]. Thus, the robot moves to arrange its neighbors into angles matching the target lattice structure. By combining the distance and angular potentials, the robots converge into a lattice with their neighbors, as shown in Figure 3. The swarm organized by these robots maintains a latticed structure.

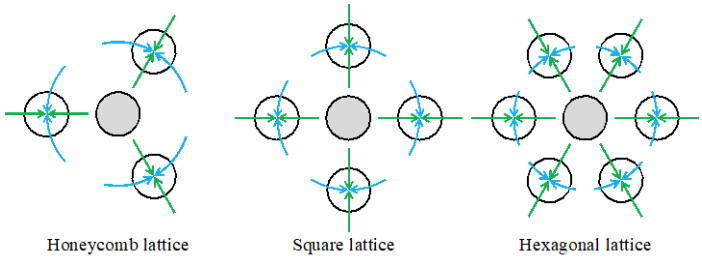


Figure 3: Positions of neighbors converged by the proposed angular and distance potentials

3.2. Collective movement using potential weights

The collective movement method in this study uses leader-follower control to ensure that the host system can manage the robot swarm similarly to the individual robots. Because the host system controls only the leader, as described above, we consider autonomous decentralized control for followers maintaining positional relationships with their neighbors. The proposed potential models rectify the positional errors caused by an organized latticed structure; if a robot swarm includes a leader, it can move while maintaining formation by adapting to structural changes caused by the leader's movement. Followers will react to structural changes caused by both the leader and each other, because this interaction affects all neighbors uniformly. Consequently, the approach using the proposed potential models creates a slow and inefficient collective movement for the robots', limiting formation speed.

We facilitate follower behavior during structural change caused by the leader's movement by weighting robot interactions. Interactions in the direction of the leader are weighted higher than interactions away from the leader. In a previous study [16], this approach was applied only to the distance potential. This study formulates a collective movement model for followers that introduces additional weights to the model potentials as follows:

$$v_i = \sum_j^{N_i} \begin{cases} P_r v_{max}\left(f_i^D(r_{ij}) + f_i^A(r_{ij}) \right) & ; l_i > l_j \\ P_w v_{max}\left(f_i^D(r_{ij}) + f_i^A(r_{ij}) \right) & ; l_i < l_j \\ v_{max}\left(f_i^D(r_{ij}) + f_i^A(r_{ij}) \right) & ; l_i = l_j \end{cases} \quad (3)$$

where N_i is the set of neighbors around the i-th robot. The robot applies a weight-adjusted potential to each neighbor, where P_r and P_w are the weights for neighbors closer to or farther from the leader, respectively. These weights are constrained by $P_r + P_w = 2$; $(P_r > P_w > 0)$ to ensure that the adjusted interactions provide the same interaction magnitude as the interactions without the new collective movement approach. Followers weigh their neighbors based on the distance between themselves and the leader. We introduce an ordering layer based on the lattice construction, which represents the routing metric to the leader as shown in Figure 4.

The leader is given a layer reference value of zero, and followers set their layer value, l_i, as $l_i = \min(l_j) + 1$ based on the layer value of their neighbors, l_j . The followers update this metric by repeating the calculation by each follower, and quickly produce a radial gradient around the leader. Consequently, each follower adjusts their potential weights according to this gradient to ensure that they more strongly perceive the structural change caused by the leader.

The movement of the leader is teleoperated through a host system to control the destination of the swarm. The leader's moving vector, related to a given moving command, is as follows:

$$v_i = P_l v_{max} f_c \tag{4}$$

where P_l is an upper-speed limit coefficient for the leader designed as $P_r - 1$ to ensure that the leader's movement does not exceed the convergence force for the neighbors around the leader, as shown in (3). f_c is a velocity vector given via teleoperation that is less than one.

With these control methods, this study realizes the collective movement of a leader-follower robotic swarm while maintaining a lattice swarm shape. The proposed method can be implemented into any swarm shape with lattice construction.

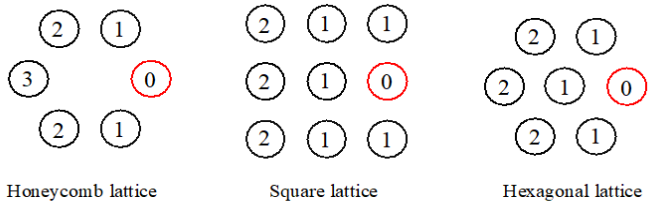


Figure 4: Calculated layer metric between the leader and followers according to the lattice structures

4. Verification of Collective Movement

Collective movement is an essential technology to support swarm robotic applications. Robot swarms can contribute to cost reduction and efficiency in their applications. In Section 1, we focus on transport and observation tasks using multiple robots, explaining the necessity for unique swarm densities, scales, and shapes. This section confirms that the proposed control method satisfies these goals. The proposed method can control an arbitrary number of robots via distributed control based on local measurements and information exchanged with neighbor robots. This study demonstrates that a robot swarm applying the proposed method can move collectively in all directions, and quantitatively evaluates the scalability of the method for any swarm size [21].

We developed numerical simulations for these evaluations using Java. The robot in this simulation is a moving object with a diameter of 0.25 m, moving with a speed limit of 0.5 m/s. The robot computes a velocity output via the proposed method at 2 ms intervals and simultaneously updates the communicated information. First, this paper evaluate the design and effectiveness (collective movement efficiency) of the parameters for the proposed method, P_r and P_w.

4.1. Collective movable speed relative to P_r–P_w ratio

The proposed collective movement performance depends on the design of P_r, P_w, and P_l. Therefore, we evaluate the maximum value of the parameter P_l that can be set under the parameters P_r and P_w. Because collective movement must be able to move in all directions, the experiment uses an environment in which formations of each lattice structure with ten-layers filled moves on a circle with a diameter of 10 m.

Figure 5 plots the maximum P_l that did not change in the structure before and after movement, as P_l is increased by 0.01 from zero under each setting P_r and P_w condition. This property also shows the collective movement efficiency for each P_r–P_w ratio. Compared to the condition with $P_w = 1$, where P_r and P_w were not introduced, the introduction of the P_r–P_w ratio resulted in higher collective movement efficiency. This result also shows that the smaller P_w is, the higher movement efficiency.

We focus on the properties of the proposed method for hexagonal and honeycomb lattices. By introducing an angular potential, the proposed method can detect and adjust the structural changes caused by moving faster than the distance-potential-only collective movement method in the previous study [16]. As a result, the proposed method achieves high collective movement efficiency even for large values of P_w. On the other hand, this property of the square lattice did not improve significantly. This is because the square lattice formation moving in the vertical and horizontal directions has no angular potential action of the neighbors in that direction and less distance potential action for robots closer to the leader. Therefore, the followers could not maintain the connection with the neighbors in the vertical and horizontal movement directions, and the collective movement efficiency decreased.

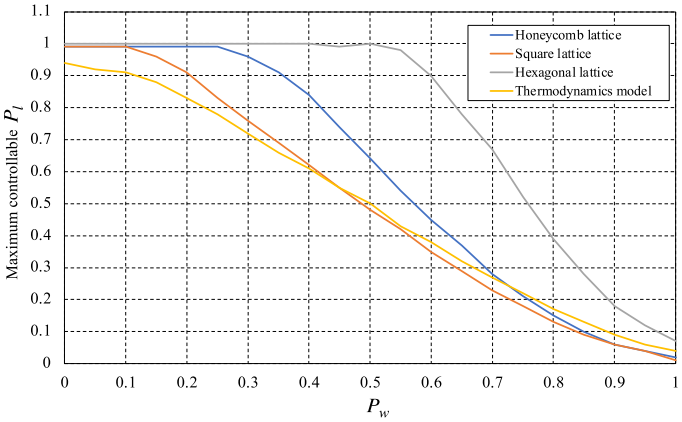


Figure 5: Maximum P_l that can be set according to the P_l–P_w ratio

The proposed collective migration requires the design of parameters, P_r, P_w, and P_l based on Figure 5. In subsequent experiments, P_l, P_r, and P_w are evaluated as 0.5, 1.55, and 0.45, respectively, to ensure that the collective movement efficiency is $P_l = 0.5$ for all lattices. These parameters are the balance values for the performance of swarm cohesion and collective movement speed, which have a trade-off relationship.

4.2. Demonstration of collective movement using various swarm shapes

This study focuses on cooperative transportation and sweep coverage using swarm robotics. These tasks require collective movement while maintaining a cohesive robot swarm in the shape

of the transported object around its center of gravity [22], or a robot swarm assembled in a row [23].

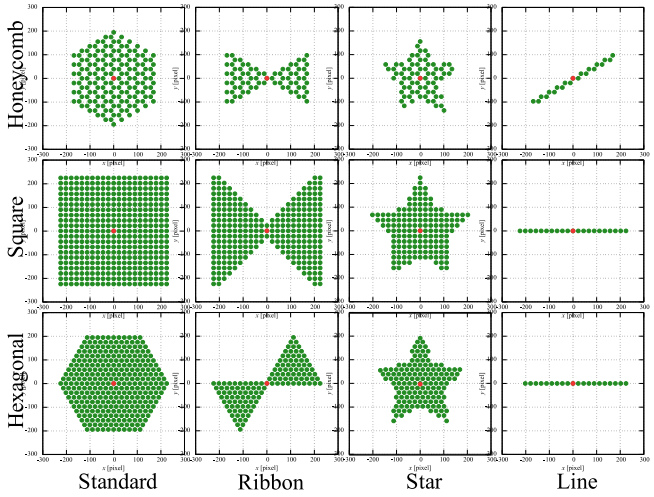


Figure 6: Formations with a scale of ten layers represented according to latticed structures

Therefore, we consider the characteristic formations in a ten-layer formation, as shown in Figure 6. The swarm shapes are as follows: standard formation, which builds a large-scale robot by enlarging the lattice structure around the leader; ribbon formation, which connects two groups at the leader position to transport multiple objects simultaneously; star formation, which is an example of an arbitrary swarm shape represented by a latticed structure, and line formation, which provides an efficient sweep. We confirm that robots implementing the proposed method can move in all directions without breaking formation by demonstrating circular movement.

Figure 7 shows the trajectories of all robots during the movement using these formations. The leader of each formation moves counterclockwise. In all conditions, followers move alongside the remotely controlled leader. This means that while the robots move according to the proposed method, they maintain their positional relationships with each other in a latticed structure. Followers move while modifying the structural changes caused by the movement of robots in the inner layer. Thus, the standard and star formations, arranging the robots to fill the inner layer, transmit their changes to the outer followers from both directions. In contrast, the line and ribbon formations cannot react as strongly, because formations with reduced interconnectivity cannot transmit the information required for the proposed method fast enough to achieve perfect collective movement; this suggests that the proposed method works best when the followers and the leader are adjacent.

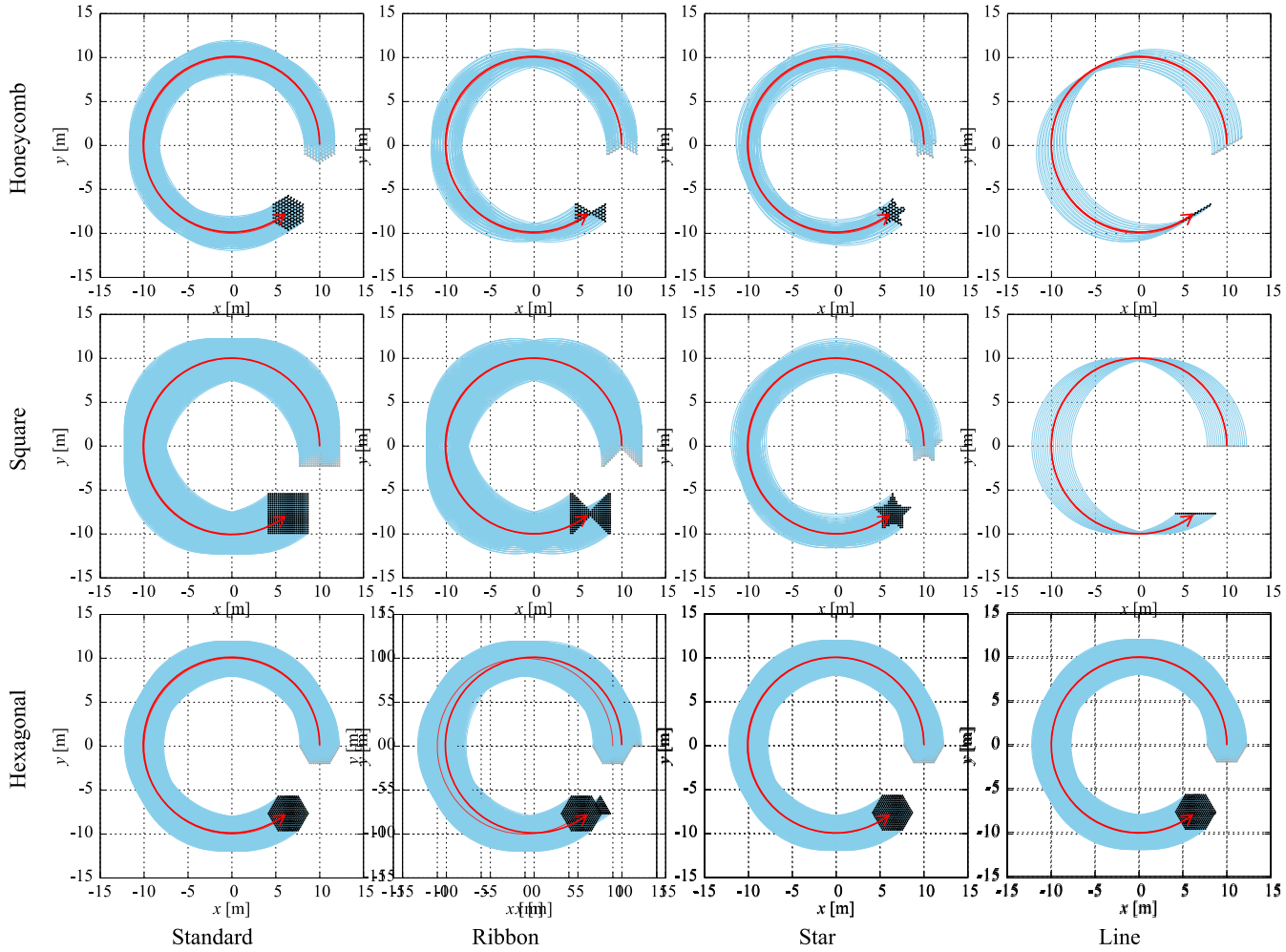


Figure 7: Movement trajectories of robots during collective movement in a circle using the formations shown in Figure 6. Each leader moves along the red arrow.

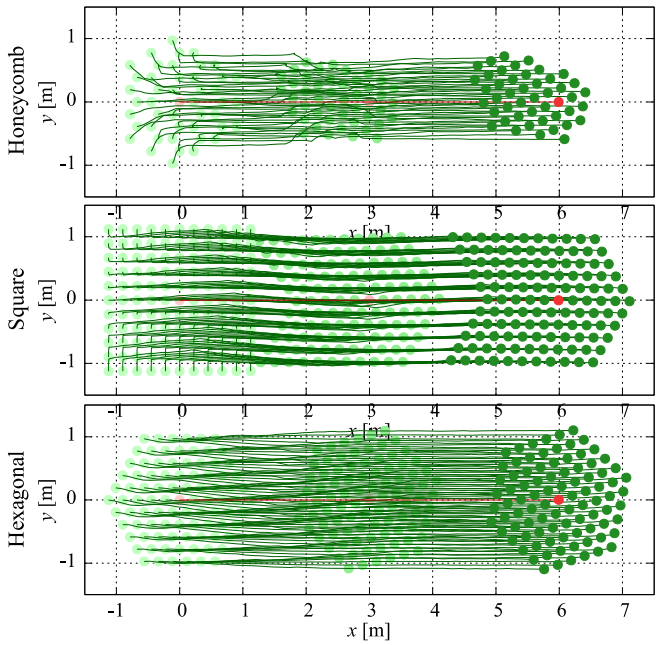


Figure 8: Movement trajectories of robots during collective movement in a straight line using standard formations for each lattice without the proposed angular potential

Subsequently, we confirm the structure-maintaining effect of the proposed method by comparing it with an existing collective movement method that can also maintain a high-density state. As already demonstrated, the proposed method organizes lattice structures using direction and distance potentials. In contrast, many existing collective movement methods control robots using interaction models without angular potentials. As a comparison, we employ a similar swarm behavior model for maintaining a high-density shape in an open space without obstacles that also applies the same Lennard-Jones potential and leader-follower control scheme as shown in our previous study [16]. We apply this existing method to a robot swarm initially arranged in a latticed structure and track their movement trajectories during collective movement. The robot swarm consists of a standard formation with five layers and moves in a straight 3 m line.

Figure 8 shows the movement trajectories of the robot swarm in standard formations with each lattice. The robot swarms arranged in honeycomb and square lattices devolved into a hexagonal lattice, while the robot swarm already arranged in a hexagonal lattice did not deform, as it was already in the highest-density state. These results mean that the angular potential applied in this paper maintains these regular latticed structures.

4.3. Evaluation of the scalability of the demonstrated formations

We demonstrated collective movement using robot swarms with a ten-layer swarm. We subsequently verify that the proposed method can provide scalable control by quantitatively evaluating the structural changes of the lattice structure during movement. To quantify structural change in the new potential, we combine the angle and distance errors between inter-neighbor positions during movement as follows:

$$\begin{cases} e_a = \dfrac{\sum_i^N \sum_j^{N_i} \| \sin^{-1} \sin(L\theta_{ij} - l_i\pi) \|}{2LNN_i} \\ e_d = \dfrac{\sum_i^N \sum_j^{N_i} \| r_{ij} - r_0 \|}{2NN_i} \end{cases} \quad (5)$$

where e_a is the angular error from the $l_i\pi$ direction, and e_d is the distance error from r_0, where $f^D(r_0) = 0$. These two metrics average values over all N robots in the swarm.

Figure 9 shows the average errors and their deviations during collective movement, with formation size varying in the demonstration from one to seventy layers. The overall result did not show any error divergence caused by lattice structure collapse, because these errors are both much smaller than the angles $\pi/12$, $\pi/16$, and $\pi/24$ [rad] that can be created by the angular potential for honeycomb, square, and hexagonal lattices, respectively, and the distance r_m [m], the radius with the greatest distance potential attraction where $f_i^{D\prime\prime}(r_m + r_0) = 0$. Where the number of robots is small, these values do not give accurate averages and are noisy. Thus, these series showed notably small values at small scales.

In line formations, robots have neighbors in chains for angular and distance comparisons regardless of the lattice type. That is, hexagonal, honeycomb, and hexagonal lattices all form a line. First, the distance errors changed with a similar magnitude regardless of the lattice structure used, because the proposed method applies the same distance potential to all latticed structures. In contrast, the angular errors were the largest for the honeycomb lattice and the smallest for the hexagonal lattice. This result means that the angular potential has a different performance depending on the lattice, as noted in the demonstration. The proposed angular potential has a convergence period based on the number of neighbors, L, to ensure that angular convergence occurs according to the angular properties of the lattice structures. Using this design, the convergence range of the angular potential is inversely proportional to L. Therefore, the angular potential of the honeycomb lattice has a wide convergence range while the hexagonal lattice has a narrow convergence range. To achieve the same shape-maintaining performance on all lattices, we should improve the angular potential to ensure that it works regardless of the lattice structure in question.

4.4. Limitations of the proposed method

In Figure 7 and 9, the entire robot swarm held formation. However, the swarm propagates local angular and distance errors to followers in subsequent layers. Therefore, the formation of the robot swarm should arrange the robots radially to ensure that the structural changes caused by the leader movement propagate quickly along the shortest possible route. The robot swarms used in this study, designed for practical applications, satisfy this requirement. To test the importance of radial error propagation, we consider collective movement under spiral formations that do not arrange robots radially. The followers in this formation were physically placed near the robots closer to the center. However, they only detect structural changes through neighbors in the spiral chain and ignore radial neighbors. Therefore, followers will have difficulty maintaining their shape owing to long delays when detecting structural changes in the formation. This study simulates the movement using spiral formations constructed with each lattice structure, as shown in Figure 10.

The robots arranged in a spiral pattern moved from left to right. Followers close to the leader maintained their structure, while followers on the outer periphery at higher layers experienced structure collapse. As expected, the proposed method could not maintain spiral formations in latticed structures. This limitation is unimportant for transportation and sweeping applications, because these applications rely on close packed lattice formations, rather than long-delay spiral formations. However, we will improve the proposed method to ensure robot swarms can be controlled with long delays in detecting structural changes, such as spiral formations for other applications.

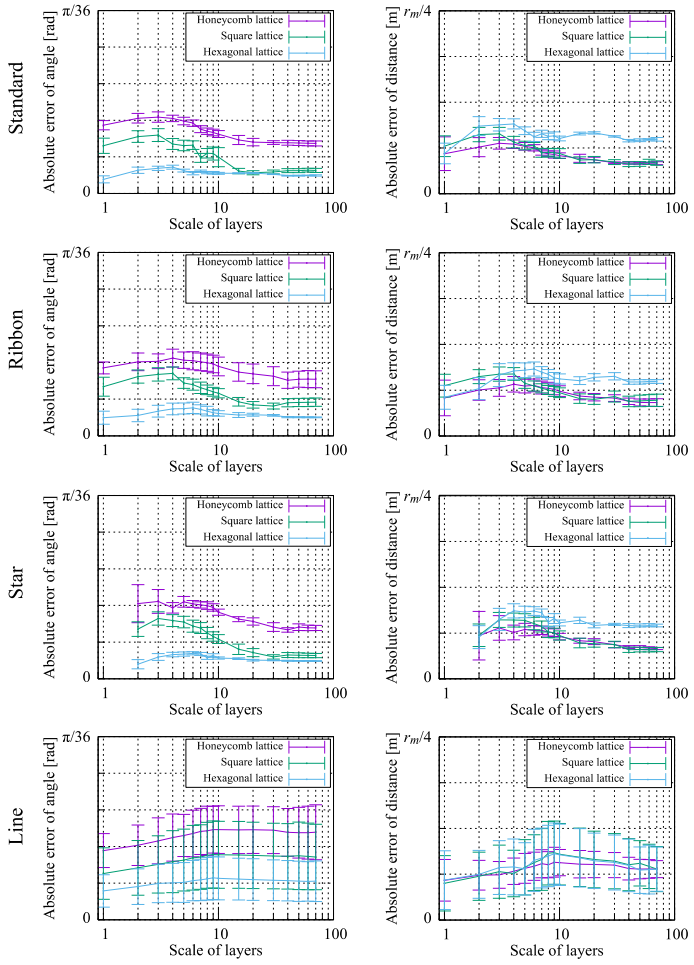


Figure 9: Angle and distance deflections during collective movements in various formations and lattices

5. Conclusion

Robotic cooperation and coordination is essential for a swarm robotic system when transporting long, large, heavy objects and observing over a wide area. Robots can efficiently complete these tasks by using a swarm with a suitable density for the task. Therefore, this paper proposed a structure-maintaining potential that can control swarm formations in regular tessellations and a collective movement method for a robot swarm using leader-follower control with a host system whereby an operator can handle the entire swarm as a single, large, and mobile robot. This study confirmed that the proposed method could maintain a structure with different densities during collective movement with four formations. The swarm control was evaluated and verified under various formation sizes regardless of formation lattice and scale, reacting to the movement of the swarm leader. The host system can lead a scalable robot swarm from a single robot using the proposed method of leader-follower control. Therefore, the system can distribute control to multiple robot swarm systems in various use cases, such as parallel transportation and environmental exploration. This approach allows robotic swarm systems to obtain higher resolution and larger scales. This approach can contribute to the flexible operation of robot swarms while reducing economic cost.

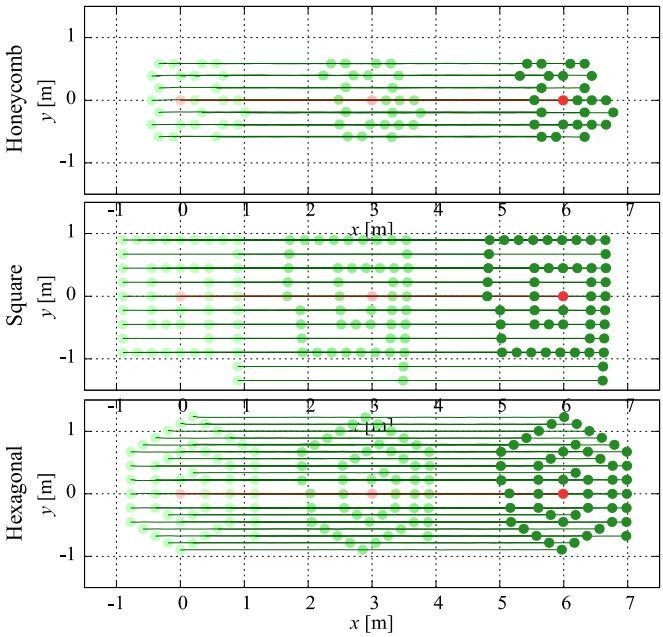


Figure 10: Movement trajectories of robots during collective movement in a straight line using spiral formations with long delays in the propagation of the movement information

In future work, we will develop autonomous collision avoidance and navigation for multiple robot swarms with different scales instead of multiple mobile robots in parallel distributed systems. We improve the proposed method to control even slow propagating formations of structural changes. In addition, the proposed method targets omnidirectional agents such as Omni-wheeled vehicles and multi-copters. Therefore, it does not consider nonlinear dynamics such as nonholonomic constraints. To implement this method on robots with various dynamics, we will develop a method that dynamically shares the direction of movement.

Acknowledgment

This work was supported by the Research Institute for Science and Technology of Tokyo Denki University (grant number Q21DS-03 / Japan). We would like to thank Editage (www.editage.com) for English language editing.

References

[1] M. Schranz, M. Umlauft, M. Sende, W. Elmenreich, "Swarm Robotic Behaviors and Current Applications," Frontiers in Robotics and AI, **7**, 2020, doi:10.3389/frobt.2020.00036.

[2] J. Wiech, V. A. Eremeyev, I. Giorgio, "Virtual spring damper method for nonholonomic robotic swarm self-organization and leader following,"

Continuum Mechanics and Thermodynamics, **30**(5), 1091–1102, 2018, doi: 10.1007/s00161-018-0664-4.
[3] Y. Kambayashi, H. Yajima, T. Shyoji, R. Oikawa, M. Takimoto, "Formation Control of Swarm Robots Using Mobile Agents," Vietnam Journal of Computer Science, **6**(2), 193–222, 2019, doi:10.1142/S2196888819500131.
[4] F. Dai, M. Chen, X. Wei, H. Wang, "Swarm Intelligence-Inspired Autonomous Flocking Control in UAV Networks," IEEE Access, **7**, 61786–61796, 2019, doi:10.1109/ACCESS.2019.2916004.
[5] J. Cheng, B. Wang, "Flocking Control of Mobile Robots with Obstacle Avoidance Based on Simulated Annealing Algorithm," Mathematical Problems in Engineering, **2020**, 2020, doi:10.1155/2020/7357464.
[6] A. Casteigts, J. Albert, S. Chaumette, A. Nayak, I. Stojmenovic, "Biconnecting a network of mobile robots using virtual angular forces," Computer Communications, **35**(9), 1038–1046, 2012, doi:10.1016/j.comcom.2011.09.008.
[7] T. Yan, X. Xu, Z. Li, E. Li, "Flocking of Multi-agent Systems with Un- known Nonlinear Dynamics and Heterogeneous Virtual Leader," International Journal of Control, Automation and Systems, **19**(9), 2931–2939, 2021, doi:10.1007/s12555-020-0578-3.
[8] A.D. Dang, H.M. La, T. Nguyen, J. Horn, "Formation control for autonomous robots with collision and obstacle avoidance using a rotational and repulsive force–based approach," International Journal of Advanced Robotic Systems, **16**(3), 2019, doi:10.1177/1729881419847897.
[9] A.M. Parrany, A. Alasty, "Introducing shell formation and a thermodynamics-inspired concept for swarm robotic systems," Robotics and Autonomous Systems, **148**, 2022, doi:10.1016/j.robot.2021.103939.
[10] Q. Wang, H. Zhang, "A Self-Organizing Area Coverage Method for Swarm Robots Based on Gradient and Grouping," Symmetry, **13**(4), 2021, doi:10.3390/sym13040680.
[11] A. Ekblaw, J. Paradiso, "Self-Assembling Space Habitats: TESSERAE design and mission architecture," in 2019 IEEE Aerospace Conference, 1–11, 2019, doi:10.1109/AERO.2019.8742122.
[12] P. Swissler, M. Rubenstein, "FireAnt3D: a 3D self-climbing robot towards non-latticed robotic self-assembly," in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 3340–3347, 2020, doi:10.1109/IROS45743.2020.9341116.
[13] Z. Yang, Z. Fu, G. Yu, J. Fei, H. Zheng, "A self-repairing approach for the M-Lattice modular robotic system using digital hormone model," Robotics and Autonomous Systems, **97**, 1–15, 2017, doi:10.1016/j.robot.2017. 08.001.
[14] J. W. Romanishin, K. Gilpin, D. Rus, "M-blocks: Momentum-driven, magnetic modular robots," in 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, 4288–4295, 2013, doi:10.1109/IROS.2013.6696971.
[15] A. V. Le, R. Parween, R. Elara Mohan, N. H. K. Nhan, R. Enjikalayil Abdulka- der, "Optimization Complete Area Coverage by Reconfigurable hTrihex Tiling Robot," Sensors, **20**(11), 2020, doi:10.3390/s20113170.
[16] K. Yamagishi, T. Suzuki, "Collective Movement Method for Swarm Robot based on a Thermodynamic Model," International Journal of Advanced Computer Science and Applications, **8**(11), 2017, doi:10.14569/IJACSA.2017. 081163.
[17] P. Bannur, P. Gujarathi, K. Jain, A. J. Kulkarni, "Application of Swarm Robotic System in a Dynamic Environment using Cohort Intelligence," Soft Computing Letters, **2**, 2020, doi:10.1016/j.socl.2020.100006.
[18] T. Kida, Y. Sueoka, H. Shigeyoshi, Y. Tsunoda, Y. Sugimoto, K. Osuka, "Verification of Acoustic-Wave-Oriented Simple State Estimation and Application to Swarm Navigation," Journal of Robotics and Mechatronics, **33**(1), 119–128, 2021, doi:10.20965/jrm.2021.p0119.
[19] A. Gutierrez, A. Campo, M. Dorigo, J. Donate, F. Monasterio-Huelin, L. Mag- dalena, "Open E-puck Range Bearing miniaturized board for local communication in swarm robotics," in 2009 IEEE International Conference on Robotics and Automation, 3111–3116, 2009, doi:10.1109/ROBOT.2009.5152456.
[20] N. Yu, A. A. Polycarpou, "Adhesive contact based on the Lennard–Jones po-tential: a correction to the value of the equilibrium distance as used in the potential," Journal of Colloid and Interface Science, **278**(2), 428–435, 2004, doi:10.1016/j.jcis.2004.06.029.
[21] M. K. Heinrich, M. D. Soorati, T. K. Kaiser, M. Wahby, H. Hamann, "Swarm robotics: Robustness, scalability, and self-X features in industrial applications," it - Information Technology, **61**(4), 159–167, 2019, doi:10.1515/itit-2019-0003.
[22] K. Huang, J. Chen, J. Oyekan, "Decentralised aerial swarm for adaptive and energy efficient transport of unknown loads," Swarm and Evolutionary Computation, **67**, 2021, doi:10.1016/j.swevo.2021.100957.
[23] A. S. Matveev, P. A. Konovalov, "Distributed reactive motion control for dense cooperative sweep coverage of corridor environments by swarms of non-holonomic robots," International Journal of Control, 1–14, 2021, doi:10.1080/00207179.2021.2005258.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 5, 35-40 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Assessing the Impact of Integrating Solar PV System using the Equal Area Criterion Method

Abdul Ahad Jhumka[1], Robert Tat Fung Ah King[*,1], Chandana Ramasawmy[2]

[1]*Department of Electrical and Electronic Engineering, University of Mauritius, Reduit 80837, Mauritius*

[2]*Advanced Mechanical and Electrical Services Ltd., Rose-Hill 71364, Mauritius*

ARTICLE INFO

Article history:
Received: 12 June, 2022
Accepted: 31 August, 2022
Online: 14 September, 2022

Keywords:
Equal area criterion
Transient stability
Critical clearing time
3-phase fault
Inertia

ABSTRACT

The impact of previous energy crisis gives an insight into what happens when the oil price crashes. Many countries were affected due the slow reaction in boosting the economic growth in key sectors such as oil-importing activities and the economic restructuring progress to face the challenges. But the present energy crisis resulting from the COVID-19 pandemic period and closure of many borders, has encouraged power distributors and generators to have recourse to renewable energy for grid integration. Countries such as USA, Germany, Italy, Spain and India are moving towards increasing the share of renewable energy on the grid. The increasing use of solar power systems over the past few years is being favoured for the decarbonization process. The percentage growth in integrating solar PV energy is forecasted to reach 23% in the future. This widespread application of renewable power energy sources (RES) such as wind and solar power comprise of many challenges namely power quality and stability. With this consequent increase in RES, synchronous generators are being displaced and replaced by power electronics grid interface, which reduces the overall rotating masses, hence the system inertia. Stability study in the presence of renewable energy is therefore an important aspect to be considered to meet the required power quality of the grid. This paper brings forward the use of equal area criterion (EAC) method for assessing the stability of the power system network with presence of renewable energy such as solar energy system in the generation portfolio. EAC provides an effective visual and analytical approach for transient stability analysis. The investigations have been performed within the steady and transient conditions in form of a 3-phase fault. The modelling and analysis were carried out using MATLAB/Simulink. Calculation of the critical clearing time (CCT) for stability assessment is the main contribution to knowledge. The increase in the CCT confirms that the Solar PV penetration to the grid will improve the transient characteristics of the national grid network.

1. Introduction

In 2020, the world encountered the outbreak of COVID-19 pandemic, which resulted in major energy upset across the world [1], as well the social life. A slight decrease in biofuels jobs from 2.5 million to 2.4 million was observed [2]. In reaction to this new challenge, the improvement of sustainable energy assets and renewable energy power system sources has shown its promising potential and effective strategy to meet the new energy demand. Globally, an increase of 30% in the connected capacity of solar energy was noted between 2019 and 2020. Japan, India and Republic of Korea were the largest contributing countries in favouring the growth of solar power energy [2]. Presently, in renewable generation portfolio, wind energy occupies a major share, while the connecting capacity of solar photovoltaic (PV) system is also increasing yearly throughout the world [3]. This growing exploitation of renewable energy sources (RES) such as wind and solar energy brings many challenges to power system operators. Besides the load variation and the stochastic nature of the RES, the integration of RES brings about a subsequent reduction in the system inertia by replacing the rotating masses of the generators with power electronics devices. Hence, stability of the power grid is being disturbed with penetration of RES. Power

[*]Corresponding Author: Robert Tat Fung Ah King, r.ahking@uom.ac.mu

electronics technologies play a predominant role in the connection of distributed generation and renewable energy sources into the power grid and are becoming more prevalent and growing rapidly as these applications are further integrated into grid-based systems [4]. The fast evolution of power electronics devices are mainly due to development of semiconductor switches such as insulated gate bipolar transistor (IGBT) [5]. Hence the importance of studying the impact of integrating renewable energy on the stability of the power system.

Transient stability analysis (TSA) can be studied in a time-domain simulation. The nonlinear behaviour of a power system dynamics can be solved mathematically using numerical integration of the nonlinear differential equations. Time-domain simulation provides flexibility, and it can consider all components of the power system. Equal Area Criterion (EAC) method provides a direct method for transient stability. This method was first introduced in the 1930s and 1940s to assess the transient stability of the classical model of a SMIB system [6]. It is used to determine the maximum load which can be acquired by the system without exceeding the threshold limit. The principle of this method consists of the basis that when the rotor angle (δ) oscillates around the equilibrium point with constant amplitude, thus maintaining transient stability. This present research work is an extended paper from the IEEE AFRICON 2021 conference proceedings [7]. A similar model is considered for further analysis on the transient stability study using the EAC method. The transient stability margin of the system and the critical clearing time (CCT) or the critical clearing angle were also considered in the analysis. Both parameters are essential elements in the stability analysis. The critical clearing angle was obtained through modelling of the swing curve on MATLAB/Simulink while the critical clearing time was obtained through mathematical derivation.

This research paper explains the concept of the EAC with connection of renewable energy using the positive phase sequencing and interpretation of results obtained. The main objective of this research paper is to investigate on the behaviours of a power grid in terms of stability in presence of RES. The method for developing the EAC consists of reducing the multi machine system into two sets, where one set is considered as the critical cluster while the other set is comprised of the remaining machines. CCT calculation for the stability studies in the presence of renewable energy contribute to knowledge of this research work.

This research paper is categorized into the background, theory, methodology, results and conclusions.

3. Background

This section describes the background works in the transient stability field. In [7], the authors studied the stability in terms of rotor angle analysis with connection of RES. The model consisted of synchronous generators and a solar PV system connected in an IEEE 9-bus arrangement. The investigation of the study was monitored under steady state condition and under transient condition such as a 3-phase fault. The result was compared to a conventional IEEE 9-bus arrangement. The author concluded that the integration of solar PV system brought an enhancement in the transient nature of the model under investigations due to a lower rotor angle deviation.

In [8], the authors discussed a unified approach on transient stability assessment of a multi machine power system with the initial objectives set as establishing the fundamentals of extended equal area criterion (EEAC), intrinsic limitations, basic formulations and applications to actual working conditions operating within a general threshold security limit. The proposed method consisted of using the equal area criterion as applied to a two-machine system. The EEAC thus established, gives the visual and analytical interpretation for transient stability analysis and sensitivity assessment. However, the limitation of the research is that it does not include any control, which could be applied to the system to enhance the transient stability. Other methods of applications of Artificial Intelligence (AI) to the system may enhance the transient stability particularly in presence of renewable energy.

The authors in [9] made use of a phase sequence exchange (PSE). The aim of the paper was to assess the efficiency of the PSE in a multi-machine system. The method consisted of installing two PSE devices positioned in a multi-machine system, with the first one installed at the outlet of each generator, and the second was the interconnected lines between the system. EEAC was used, and a control method was then proposed. It was observed that the use of the control methods maintained the synchronization, thus ensuring the system integrity and power quality. The efficiency of the PSE was checked against Kundur's two-area four-machine power system and the IEEE 39-bus system and compared with other emergency control methods.

In [10], the authors provided an overview of a revisited EAC with the aim of illustrating the different types of transient stability under occurrence conditions, namely type I, type II, type III and type IV. Four types of variables were discussed on the EAC, P_m, P_b, P_d, P_p where P_m is the mechanical input power while P_b, P_d, P_p are corresponding maximum power during the pre-fault, during fault and post fault. Conditions of always stable even during the fault and unstable states were investigated when $P_d<P_p$ and $P_d>P_p$. Type I was typical and widely accepted as common sense in the transient stability analysis. The study proved relevant to the development of relay protection for enhancing and maintaining the power system stability.

The authors in [11] investigated on the transient rotor angle stability during extreme conditions such as three-phase fault, the disconnection of generators, change of weather due to its stochastic nature and partial shutdown of PV generation [12]. Transient stability rating inclusive of the swing motion and damping are determined due to change of angular speed, frequency, and power output. Other factors such as critical clearing time (CCT) and loadability have been also studied in [13]. It was concluded that the transient stability rating of the grid was dependent on the extent of PV connection and type of transient events induced in the system.

In [14], the authors studied the impact of integrating photovoltaic generation penetration and together with the use of Flexible AC transmission (FACTS) devices such as STATCOM on the transient stability of a single machine infinite bus (SMIB) system based on the rotor angle stability. The system model was

created in MATLAB/Simulink software. It was concluded that with the penetration of solar PV along with the use of STATCOM was able to provide solutions to improve the behaviour of networks by damping the transient oscillations of the rotor angle for various fault conditions.

The literature survey described above shows some research gaps in the field of transient stability study in presence of renewable energy using the EAC method. Application of artificial intelligence (AI) in the transient stability has also received little attention in research. It is expected that the use of AI will enhance the use of RES by appropriate prediction of fault. However, the use of AI will not be considered within this research paper and same will be considered as further works. EAC will determine the critical clearing angle (CCA). The critical clearing time (CCT) will be determined using mathematical derivation and numerical integration. Both parameters are essential in the study of transient stability. The critical clearing angle is the maximum allowable change in the power angle δ before clearing the fault and without loss of synchronism. The critical clearing time is the maximum allowable time to remove a disturbance without interrupting the system integrity. To our knowledge, no studies of similar nature have been conducted for transient stability analysis.

4. Theory

This section explains the theory behind the single machine infinite bus (SMIB) and the equal area criterion (EAC).

4.1. Single Machine Infinite Bus (SMIB)

The diagram in Figure 1 is the equivalent representation of a SMIB in a power system through a line of reactance X_e.

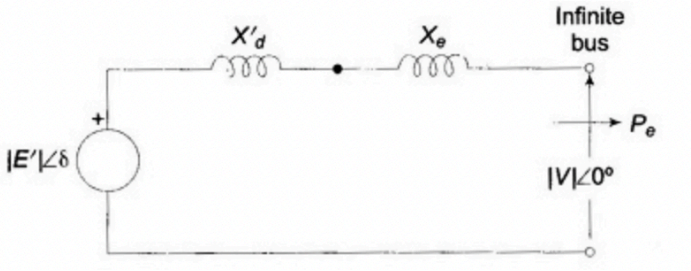


Figure 1: Arrangement of single machine

A generator operating at synchronous speed ω_m will develop an electromagnetic torque T_e and a corresponding electromagnetic power P_e. The input torque developed by the prime mover at the generator shaft is T_i, then under steady state conditions, $T_i = T_e$.

However, under disturbances, the input power, P_i is no longer equal to the output power, P_e. This gives rise to an accelerating power, ΔP which is equivalent to the accelerating torque. According to [15],

$$\Delta P = P_m - P_e = \frac{2H}{\omega_s}\frac{d^2\delta}{dt^2} \quad (1)$$

where

ΔP: Accelerating power (pu)

δ: Rotor angle (rad)

ω_S: Angular speed (rad/s).

H: Inertia (MJ/MVA)

P_m: Mechanical Power (pu)

P_e: Electrical Power (pu)

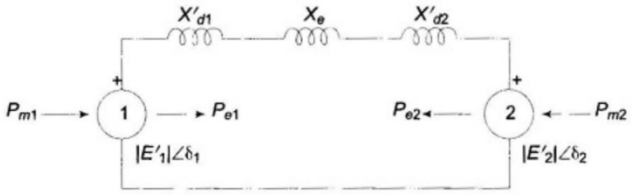


Figure 2: Two-Machine system

A two machines system operating at normal dynamic condition, i.e, steady state, the electrical power developed by the generator applications is equal to the power absorbed by the motor (network being lossless). Thus, at all time

$$P_{m1} = -P_{m2} = P_m \quad (2)$$

$$P_{e1} = -P_{e2} = P_e \quad (3)$$

The swing equations for the two machines can be written as per [13], which is detailed in (4) and (5),

$$d^2\delta_1/dt^2 = \pi f(P_m - P_e)/H_1 \quad (4)$$

$$d^2\delta_2/dt^2 = \pi f(P_e - P_m)/H_2 \quad (5)$$

Subtracting (5) from (4)

$$\frac{d^2(\delta_1-\delta_2)}{dt^2} = \pi f \frac{(H_1+H_2)}{H_1H_2}(P_m - P_e) \quad (6)$$

Therefore, the accelerating power can be written as

$$P_m - P_e = \frac{H_\eta}{\pi f}\frac{d^2\delta}{dt^2} \quad (7)$$

where H_η is the equivalent system inertia.

Electrical Power interchange can be expressed as follows

$$P_e = \frac{|E_1'||E_2'|}{X_d'+X_e+X_d'}\sin\delta \quad (8)$$

A two machines system has the same form of swing and power angle equations, with that of a single machine connected to an infinite bus [16, 17].

4.2. Equal Area Criterion (EAC)

Stability assessment for a SMIB may be determined based on the rotor angle deflection. It is also possible to define the stability graphically using the power-angle diagram, which is known as the equal area criterion (EAC).

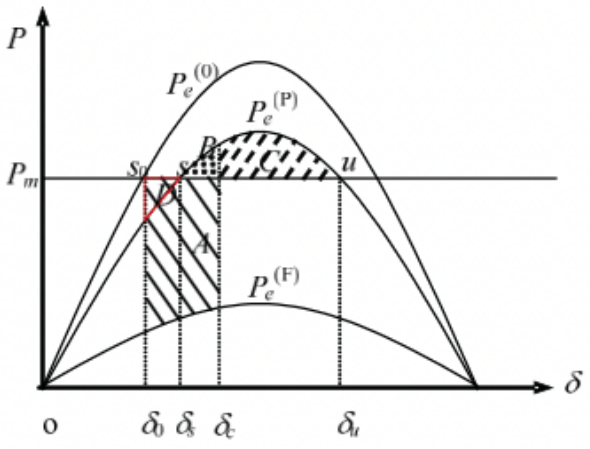


Figure 3: Power angle curve

According to [18], when

Area A < Area C: Stability is achieved within first swing operation.

Area A > Area C: System is not stable.

Area A = Area C: Machine is critically stable in the first swing.

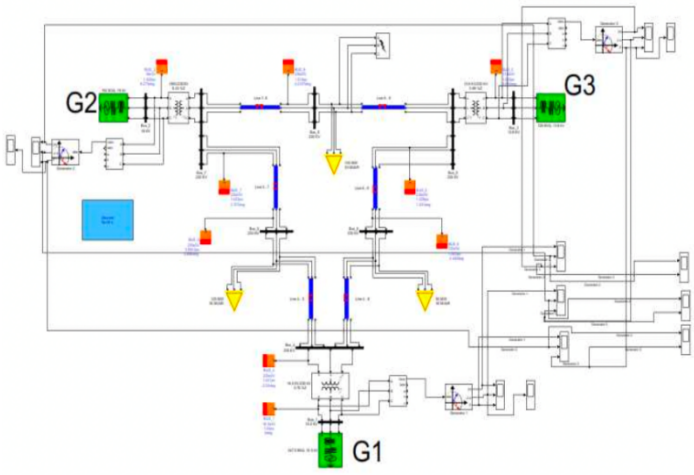


Figure 4: IEEE 9-bus arrangement

In case of a multi machine system, when

When Area A < Area C: the system is operating synchronously in the first swing.

When Area A > Area C: the system is operating asynchronously in the first swing.

When Area A = Area C: the system is operating in a critically synchronous condition within the first swing.

4.3. Methodology

The proposed system under study is comprised of an adapted IEEE 9-bus with the system frequency changed to 50 Hz. Figure 4 shows the arrangement of synchronous generators under investigation. The default values of the system's impedances were used during the study.

The power angle curve can be developed based on the impedance of the generator and the transmission line, which is shown in Figure 5.

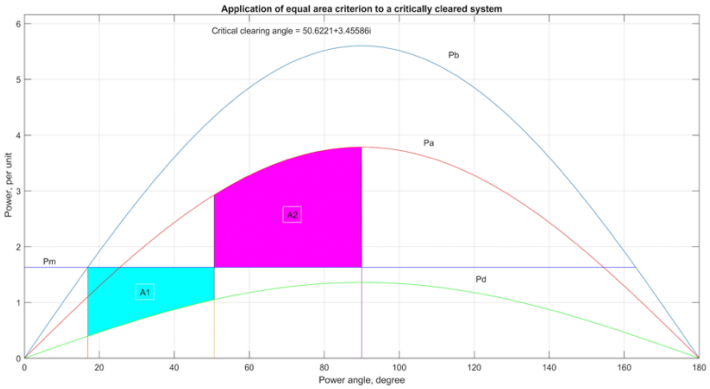


Figure 5: Power angle curve for an IEEE 9-bus arrangement

During the stability assessment of the system after the clearance of the fault, the accelerating energy (A1) and decelerating energy (A2) can be calculated using the principle laid down by [19],

$$A = \int_{\delta_2}^{\delta_1} \left(P_1 - P_2 \right) d\delta \tag{9}$$

In the present condition,

$A_2 > A_1$ which implies stable state of the system.

The critical clearing time of the system is calculated with integration of (7) above, which will result in (10) according to [20]

$$t_c = \sqrt{\frac{2H_\eta \left(\delta_c - \delta_0 \right)}{\pi f P}} \tag{10}$$

where t_c is the critical clearing time

H_η: Equivalent inertia (MJ/MVA)

δ_c: critical clearing angle (°)

δ_o: angle at steady state (°)

f: frequency (Hz)

P: Power (MVA)

The equivalent inertia H_η can be determined using the R method which is described in [21]. The resulting critical time for the generation mix of synchronous generator is 0.4 s for an estimated inertia of 33.125 MJ/MVA.

During the integration of the solar PV system, generator G2 is displaced from the grid as per the following arrangement in Figure 6.

The first step in developing the power angle curve for this new arrangement, a single axis machine was considered, together with system impedance. The network was reduced to an equivalent circuit diagram by applying Thevenin's theorem, which is given in Figure 7.

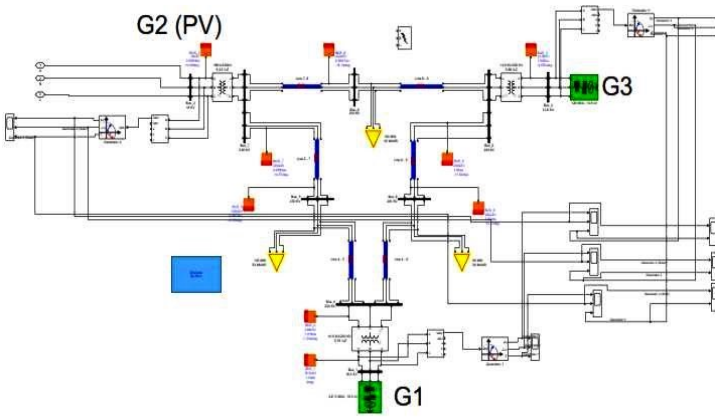


Figure 6: Solar PV system connected in an IEEE 9-bus arrangement

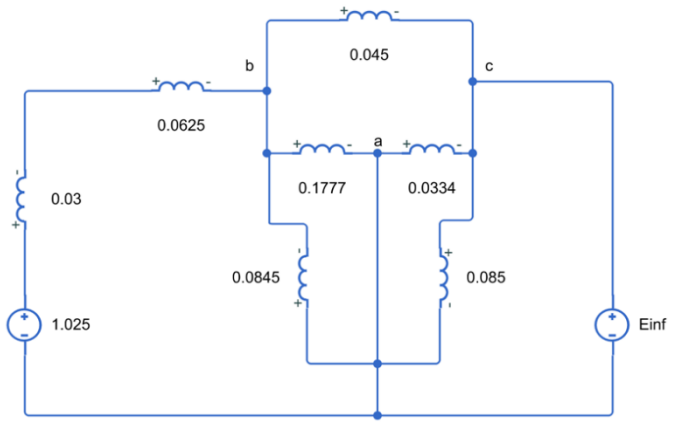


Figure 7: Equivalent circuit with solar PV integration

The result of the power angle curve is shown in Figure 8.

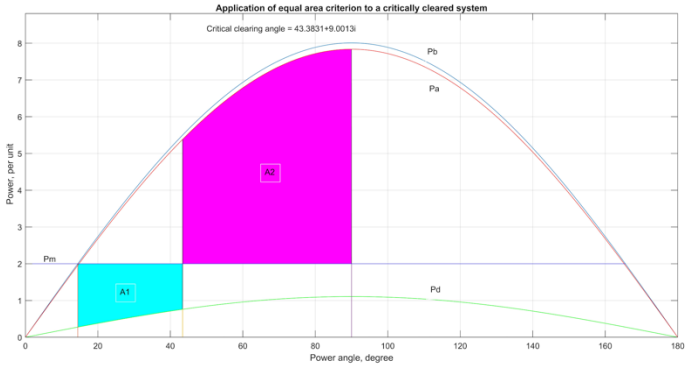


Figure 8: Equal Area Criterion with connection of renewable energy

The integration of solar PV system into the network has lowered the critical clearing angle from 50° to 43°. Using the principle as described above, it can be determined that A2 > A1, which confirmed the stable state of the system. Moreover, equation (9) in [22] was applied to determine the equivalent inertia. The estimated inertia was calculated using (11)

$$H_e = (f_{nom} C V_{dc} \Delta V_{dc}) / 2P_{res} \Delta f \qquad (11)$$

According to [22], $(\Delta V_{dc}) / \Delta f$ is a constant. Allowing a maximum change of frequency of $\pm$ 0.5Hz and a dc link voltage of 500 V, the constant can be taken at 1000. An estimated value of H_e = 6 MJ/MVA can be availed from the expression above. Upon substituting in (10) above, it is observed that the critical clearing time increased to 1.18 s.

5. Conclusion

As mentioned in the introduction, the exploitation of the solar energy is gradually increasing. It was demonstrated through this study that the system dynamics are impacted with the penetration of large-scale renewable generation. Thus, it is imperative to set the conditions for the steady state and dynamic performance of a power system for different installation scenarios of renewable power plants such as photovoltaic (PV) ones. This study can prove useful to power system operators and policy makers when it comes to the decarbonization of the power grid and its stability concept.

This research paper described the equal area criterion method as a means of transient stability analysis. Similar method can be applied in presence of renewable energy such as solar PV system. The study concluded that with solar power generating system forming part of the generation mix, an improvement of the transient stability of the grid were noted due to a decrease in the critical clearing angle from 50° to 43°.

The paper also brought about an innovative method of estimating the critical clearing time based on an estimated inertia with presence of RES. The critical clearing time was increased from 0.4s to 1.18s in the second test condition owing to the presence of solar PV system, thus confirming an enhancement in the transient stability. Further investigation needs to be worked on the application of AI to such model for fault prediction.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The assistance given by Advanced Mechanical and Electrical Services Ltd. in the development of this research paper was greatly appreciated.

References

[1] A.T. Hoang, S. Nižetić, A.I. Olcer, H.C. Ong, W.-H. Chen, C.T. Chong, S. Thomas, S.A. Bandh, X.P. Nguyen, "Impacts of COVID-19 pandemic on the global energy system and the shift progress to renewable energy: opportunities, challenges, and policy implications," Energy Policy, **154**, 112322, 2021, https://doi.org/10.1016/j.enpol.2021.112322.

[2] IRENA, ILO, Renewable Energy and Jobs – Annual Review 2021, International Renewable Energy Agency, International Labour Organization, Abu Dhabi, Geneva, 2021.

[3] I.S. Jha, S. Sen, K. Bhambhani, R. Kumar, "Grid integration of renewable energy sources," International Journal of Scientific and Technical Advancements (IJSTA), **1**, 1-5, 2015.

[4] M.E.T. Souza Junior, L.C.G. Freitas, "Power electronics for modern sustainable power systems: Distributed generation, microgrids and smart grids—a review," Sustainability, **14**, 3597, 2022, doi: 10.3390/su14063597.

[5] J.M. Carrasco, L.G. Franquelo, J.T. Bialasiewicz, E. Galvan, R.C. Portillo Guisado, M.A.M. Prats, J.I. Leon, N. Moreno-Alfonso, "Power-electronic systems for the grid integration of renewable energy sources: a survey," IEEE Transactions on Industrial Electronics, **53**(4), 1002-1016, 2006, doi: 10.1109/TIE.2006.878356.

[6] A. Bahmanyar, D. Ernst, Y. Vanaubel, Q. Gemine, C. Pache, P. Panciatici, "Extended equal area criterion revisited: a direct method for fast transient stability analysis," Energies, **14**, 7259, 2021, doi:10.3390/en14217259.

[7] A.A. Jhumka, R.T.F. Ah King, C. Ramasawmy, "Rotor Angle and Inertia Analysis in Presence of Renewable Energy Source," in 2021 IEEE AFRICON Conference, 1-6, 2021, doi: 10.1109/AFRICON51333.2021.9571011.

[8] Y. Xue, M. Pavella, "Extended equal-area criterion: an analytical ultra-fast method for transient stability assessment and preventive control of power systems," International Journal of Electrical Power & Energy Systems, **11**(2), 131–149, 1989, doi:10.1016/0142-0615(89)90021-5.

[9] Y. Li, S. Huang, H. Li, J. Zhang, "Application of phase sequence exchange in emergency control of a multi-machine system," International Journal of Electrical Power & Energy Systems, **121**, 106136, 2020, doi:10.1016/j.ijepes.2020.106136.

[10] Y. Sun, J. Ma, J. Kurths, M. Zhan, "Equal-area criterion in power systems revisited," Proc. R. Soc. A. **474**:20170733, 2018. doi:10.1098/rspa.2017.0733.

[11] S. You, G. Kou, Y. Liu, X. Zhang, Y. Cui, M.J. Till, W. Yao, Y. Liu, "Impact of high PV penetration on the inter-area oscillations in the U.S. eastern interconnection," IEEE Access, **5**, 4361–4369, 2017, doi: 10.1109/ACCESS.2017.2682260.

[12] S. Eftekharnejad, V. Vittal, G. T. Heydt, B. Keel, J. Loehr, "Impact of increased penetration of photovoltaic generation on power systems," IEEE Transactions on Power Systems, **28**(2), 893-901, 2013, doi: 10.1109/TPWRS.2012.2216294.

[13] B., Tamimi, C., Cañizares, K. Bhattacharya, "System stability impact of large-scale and distributed solar photovoltaic generation: the case of Ontario, Canada," IEEE Transactions on Sustainable Energy, **4**(3), 680–688, 2013.

[14] S. Fettissi, D. Labed, I. Labed, "Influence of high PV penetration and STATCOM on rotor angle stability of SMIB transmission system," Journal of Electrical Engineering & Technology, 849-857, 2018.

[15] M. Pavella, D. Ernst, D. Ruiz-Vega, Transient stability of power systems: a unified approach to assessment and control, Springer, 2000.

[16] P. Kundur, N.J. Balu, M.G. Lauby, Power system stability and control, McGraw-Hill. 1994.

[17] L. Fang and Y. Ji-lai, "Transient Stability Analysis with Equal Area Criterion Directly Used to a Non-Equivalent Generator Pair," in International Conference on Power Engineering, Energy and Electrical Drives, 386-389, 2009, doi: 10.1109/POWERENG.2009.4915162.

[18] S. Paudyal, R. Gokaraju and M. Sachdev, "Application of Equal Area Criterion Conditions in the Time Domain for Out-of-Step Protection," in IEEE PES General Meeting, 1-1, 2010, doi: 10.1109/PES.2010.5589303.

[19] H.-D. Chiang, F.F. Wu, P.P. Varaiya, "A BCU method for direct analysis of power system transient stability," IEEE Transactions on Power Systems, **9**(3), 1194–1208, 1994, doi:10.1109/59.336079.

[20] F.I. Khan, T. Aziz, "Transient Stability Evaluation of a Solar PV Integrated Industrial Microgrid," in IEEE Region 10 Symposium (TENSYMP), 535-540, 2019, doi: 10.1109/TENSYMP46218.2019.8971169.

[21] D. Zografos, M. Ghandhari, "Estimation of Power System Inertia," in IEEE Power and Energy Society General Meeting (PESGM), Boston, MA, USA, 1-5, 2016, doi: 10.1109/PESGM.2016.7741073.

[22] R, Ghosh, N.R, Tummuru, B.S. Rajpurohit, A. Monti, "Virtual Inertia from Renewable Energy Sources: Mathematical Representation and Control Strategy," in IEEE International Conference on Power Electronics, Smart Grid and Renewable Energy (PESGRE2020), 1-6, 2020, doi: 10.1109/PESGRE45664.2020.9070733.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 39-48 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Advantages of 3D Technology in Stereometry Training

Penio Lebamovski*

Bulgarian Academy of Sciences, Institute of Robotics, Sofia, 1000, Bulgaria

ARTICLE INFO

Article history:
Received: 30 August, 2022
Accepted: 25 October, 2022
Online: 13 November, 2022

Keywords:
Boundary method
Virtual reality
Stereometry
Stereo system
3D technology
3D printing

ABSTRACT

This paper presents a new software for stereometry training. Its name is StereoMV (Stereo Math Vision). It is results from a dissertation work on the topic "Stereoscopic Training System". It introduces virtual reality systems that are both immersive and non-immersive. The difference between them is in the equipment and the stereo effect they offer. A comparative analysis is performed between the virtual environments that are part of the software. 3D visualization is of four types: traditional, 3D active, 3D passive, and through immersive systems such as CAVE (Cave Automatic Virtual Environment) and HMD (Head Mounted Display). There are also four visualization modes in Java3D. Mixed Immediate Mode is used to implement stereoscopic projection using 3D active technology. Passive technology uses anaglyph glasses with a red and blue filter. Traditional visualization can be implemented from any computer. The system's most important function is exporting the objects in a file with the extension .obj. All this is thanks to a Java 3D library and a new boundary method involved in generating geometric objects, which sets the program apart from existing software solutions for training in the discipline of stereometry. The software can be used for stereo visualization of random 3D models from various scientific fields. Unlike planimetry in stereometry, a three-dimensional drawing would be much more difficult to understand, and that's why 3D technology comes to the rescue. This way, the student's interest will be strengthened, from where their spatial thinking will be further developed.

1. Introduction

"This paper is an extension of work originally presented in conference 2021 International Conference Automatics and informatics [1]".

One of the fastest-growing areas of the modern world is 3D technology, which in recent decades has found application in almost all spheres of society, including education. What until recently looked like science fiction is a reality today. The principle of 3D visual stereo is based on binocular parallax. One of the most important features of 3D technology is making the invisible/visible and the inaccessible/accessible. By means of it, new concepts are learned and higher levels of concentration are showed. The introduction of 3D systems in the education of the Republic of Bulgaria is currently hindered by many objective reasons, such as the lack of high-tech infrastructures and standards, as well as the insufficient awareness of potential consumers about the benefits of the new technology. These limitations can be overcome in the near future. Despite the constraints, experts predict the active introduction of 3D systems in secondary and higher education, and the first steps in this direction have been taken. The use of these systems in education will lead to a paradigm shift in how people view school and university education [2]. The application of this technological innovation in the education sector will improve how pupils and students will acquire new knowledge [3]. Three-dimensional modeling and animation of information will provide teachers with high-quality teaching materials and programs that will help students more easily absorb understand the material studied, increase their motivation and ability to learn large amounts of information.

Nowadays, solid knowledge is built faster through the so-called digital model, which can be realized through information technology and virtual reality devices [4] . The teacher's choice of appropriate software can become a potent tool in the learning process. In this way, students will be involved in the most innovative way possible to build knowledge. The proposed new software can participate in stereometry training. In contrast to planimetry, a problem arises related to understanding three-dimensional drawing. A large part of the students does not have well-developed spatial thinking. This problem can be solved by using appropriate 3D technology, such as (virtual, augmented, mixed reality, and 3D printing). It is assumed that the greater the immersion in virtual reality, the greater the interest in mathematics on the part of students.

The article aims are to present new software for teaching geometry in three-dimensional space. It is intended for middle and upper middle schools. Virtual environments that can visualize the software are both immersive and non-immersive. The system can export the shapes to a .obj file.

*Corresponding Author: Penio Dimitrov Lebamovski; p.lebamovski@abv.bg

https://dx.doi.org/10.25046/aj070605

2. 3D stereoscopic systems

Stereoscopic systems in a virtual environment are available with and without immersion. Each of them has specific characteristics that distinguish them from the others. Some systems require more money to invest in hardware but at the same time offer a very great experience. These are immersive systems such as CAVE and HMD. Non-immersive systems are desktop computers and can be 3D passive or active technology [5]. In one case, the active virtual environment is presented on a special stereo display, and 3D active (shooter) glasses are required. With passive technology, only a pair of anaglyph glasses is needed. Compared to other environments for VR, immersion is relatively less, but still, good experiences are offered. Their advantage is that they are relatively cheaper. And therefore, they are widely used. That is why they are suitable for application in education in the case of training in streometry.

3. Softwares for training in stereometry

After a study in European and Bulgarian education regarding the use of dynamic software in mathematics classes, it was found that the following programs have been used so far: Geogebra, Dalest, and Cabri3D. Thanks to these software solutions, students become engrossed in the presentation and study of mathematics material. The programs help the student to develop his spatial imagination and logical thinking and to accumulate new knowledge and skills. GeoGebra is an interactive program designed to learn and teach algebra, geometry, and statistics from elementary school to university. The program can work in many modes, such as a mode for 2D graphics i.e. planimetry, and a way for 3D graphics - stereometry, at the same time, there is also a mode for algebraic calculations related to the mentioned sections of geometry. The DALEST software project results from the joint work of five European universities: Cyprus, Great Britain, Portugal, Greece, and Bulgaria. The project aims to create dynamic 3D software for young people learning stereometry in middle school for young students. The CABRI 3D software system allows the construction of geometric figures in three-dimensional space and manipulations with them. The main functions of CABRI 3D and GEOGEBRA are constructing lines, surfaces, and 3D figures; translation, inversion, rotation, sections; polyhedra constructions, etc. Of the three considered software for training in the discipline of stereometry, only one of them has the possibility of implementing stereoscopy through anaglyph projection through passive 3D technology, and that is the GEOGEBRA program. The softwares uses trigonometry and the extrusion technique to construct geometric figures (polygons, prisms and pyramids) and rotational bodies. CABRI 3D and GEOGEBRA use an advanced traditional method, which is again based on trigonometry, to construct the geometric objects. An essential part of programming is regular polygons. On their basis, the complex polyhedra as prism and pyramid are constructed. As is well known, they are the basis of the geometry discipline in secondary school. The traditional way of creating them is based on trigonometry (sine, cosine, and radius). Several vertices and radius characterize them. This paper presents a new method for generating a regular polygon. It defines them by the number of vertices and the length of the side. The traditional way to create a polyhedra is by extruding a polygon defined by a radius. Here, extruding is the most used technique in 3D programming, and it is not appropriate to use when constructing geometric figures studied in mathematics. In this way, meaningful information about the vertices of the geometric body is lost. Therefore, thanks to the new method, the geometric objects are mathematically more accurate since they do not use the mentioned technique.

4. Extrude/sweep 2D graphics

One of the main techniques in 3D programming to create 3D models is extruding 2D and 3D graphics. Various arbitrary objects can be formed with it. It is usually applied to curves and polygons. The idea is that through a certain movement, such as (rotation, translation, and trajectory defined by an arbitrary curve) a complex geometric construction can be generated. Math software uses this technique in developing stereometric figures. A regular hexagonal prism generated by the traditional method is shown in Figure №1. The base is a polygon with 6 vertices and a radius equal to 2. The volume body is created by planar sectioning and setting a rule in the case of prism extrusion by translation. The diagram Figure №2 shows how the stereo system visualizes a regular polygon with new method without extrusion. First, it is necessary to enter input data, such as the number of vertices and side lengths. Here, the side length parameter is the same for each polygon. The values of parameter I vary between (3, 4, 5 and 6). Each shape is drawn using coordinate arrays. The values of the vertices are of type double. They were followed by visualization of the figure and calculating the perimeter and face.

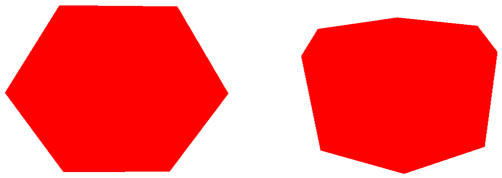


Figure 1: Traditional method and prism

5. Methods for stereoscopic visualization

The principle of 3D visual stereo is based on binocular parallax. The two images with parallax information are used, each of the images being seen by the corresponding eye, resulting in different incarnations. There are two methods to distinguish it [6]: First one with the help of display, and second one with stereo glasses.

6. 3D stereoscopic Display

3D stereoscopic displays are classified into two categories depending on the equipment:

1. Autostereoscopic, which does not require additional equipment to generate a 3D stereo image.
2. Stereoscopic displays: These are assisted by other devices.

In stereoscopic displays, the 3D image is obtained from two different images for the left and right eyes, respectively [7]. The two images are similar, with a slight offset from each other. The beginning was given in the 20s of the last century, mainly in films,

cinema halls, and television. Nowadays this technology already has a mass application, for example: in education and video games. Stereoscopic 3D technologies are divided into active (LC liquid crystal) and passive (anaglyph and polarizing). Anaglyph technology uses two color-filtered images and relatively inexpensive glasses. The colors that are commonly used are: red-cyan, red-green and blue-red. With this system, no additional equipment is needed. It is enough to have a pair of anaglyph glasses. The main disadvantage of anaglyph glasses is that a lot of colors can be lost. The active (shooter) system is a 3D stereoscopic technique that sends the left image to the left eye and the right image to the right eye, respectively [7]. One of the most used immersive 3D systems is the HMD (Head Mounted Display) virtual helmet [7]. This is a device that is mounted on head, with a small optical display in front of each eye. The virtual helmet is used in various virtual and augmented reality applications. It is easy to operate and can be installed on any platform. This system consists of two optical parts: an input and an output pair [7]. The first pair magnifies the image on the microdisplay, while the outer pair projects the color image to the viewer. II (integral imaging display) This type of display offers 3D imaging without needing additional glasses or tracking devices. In recent years II has attracted much attention from many researchers due to its full color moving 3D imaging full parallax, which sets II apart from other 3D display technologies [7]. Volumetric display is used and developed to present 3D volumetric or animated images in space. This technology is mostly used in military, medical, and academic applications.

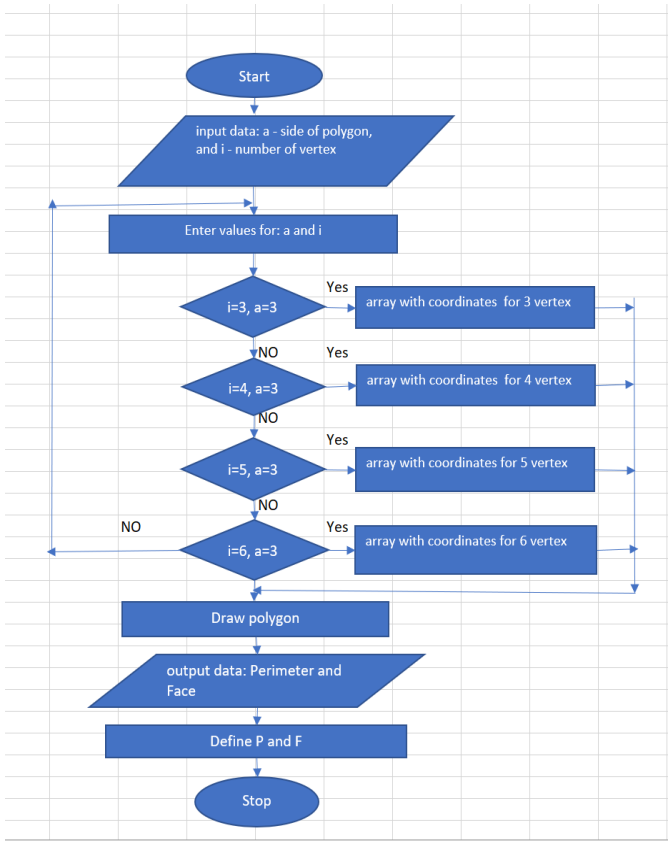


Figure 2. Diagram of polygons: new method

7. 3D Technology

The stereoscopic vision techniques that are part of the proposed new software are: HMD, active glasses, passive glasses (anaglyph) and CAVE (Cave Automatic Virtual Environment). There are several options for visual stereo. The first one uses the so-called active glasses, which alternate between the eyes in sync with the monitor. Stereoscopic perception requires each eye to receive a separate view of the scene. The view should be at a different angle and by according to the position of the head. The distance between the eyes is called interocular distance. The second way to generate stereoscopic objects on stage is by having two serial and different monitors [8]. This is precisely the way the HMD is ready to work, as each eye has its monitor. Here again, each eye has a different view, but there is no switching between the eyes and the two monitors are constantly on [8]. The display with which the three-dimensional perception is visualized is of two types: head-mounted or room mounted [8, 9]. Most displays are single screens without head tracking. Such systems are the simplest forms of room-mounted displays. In the case of the HMD, the screens are attached to the head (i.e. when the head moves, the screens follow it). There are several options for visual stereo. The first uses the so-called Head tracking that allows head movement instead of the standard mouse, joystick, keys, and other controllers.

7.1. HMD vs Screen

The full view specification can be divided into two parts, namely the view position and the view volume [8, 9]. Here it is necessary to calculate two matrices depending on the specification. When using the HMD, the camera needs to depend on the movement of the head, so that when the user's head moves and rotates, the 3D view moves together with it. It is important that the relationship between the eyes and the screen don't mustn't change. Therefore, the projection matrix remains constant [9]. When using room-mounted, the movement of the head changes the projection matrix but does not control the camera. This is a problem when watching stereo. With room mounted it is not necessary to use head tracking unless we are in stereo mode. Appearance settings depend on the location and orientation of the eyes. If the user's head changes position, the view also changes [9]. Dynamically recalculating the projection matrix and the view matrix is a very annoying task. Java allows tracking and supports features for automatic head tracking [8]. Where the view is considered a camera. A projection matrix represents the section definition. A visual matrix represents the position of the view. The first of the matrices are defined in the coordinate system, where the eye is located at the beginning of the view to the negative direction of the application z-axis (applicate) and up to the ordinate. On the other hand, in the view/camera, the eye is located at the beginning, looking down at z and up at y.

7.2. 3D Active, Passive and Immersion Systems

Anaglyph technology does not require expensive equipment, unlike 3D active technology. Here the user must be equipped with only one pair of glasses. The colors most often used are red for the left eye and blue for the right [10]. This projection is most often used when viewing magazine photos. The idea is to filter the

eye images based on color. The second technique requires two monitors for each eye to generate a stereo image. Here, as in the first technique, each eye must receive a certain image. The virtual reality helmet - HMD - works on such a principle. The difference between the two techniques is user interaction. How interactivity is carried out through the helmet is similar to non-immersive systems: through a mouse, joystick, keyboard, etc. The main feature that differs between the two systems when interacting is "Head Tracking" - a 6DoF device (six degrees of freedom) that offers 3D information in the form of location and orientation. In this way, a higher-level interaction is performed, offering a relatively more significant emotional effect on the part of the average user. The stereometry software proposed in this article can be implemented by immersive and non-immersive virtual reality devices, such as 3D active (shooter technology), 3D passive (anaglyph), and immersive systems (CAVE, HMD). With active technology, frames are alternated sequentially, first for one eye and then for the other [11]. The frequency of these frames is usually 120 Hz, 60 Hz per eye [11]. For the structure to be perceived by the viewer, special active glasses are used, which darken the glass of the eye which doesn't have to see the frame that was displayed. Here, darkening alternate first in one eye and then in the other. The use of a synchronization device is also required. One of the systems that offer the most excellent immersive effect is the CAVE system. In this 3D technology, the projectors are placed between the walls of a cube-shaped room. As with the active technique, the user must use glasses[5, 12].

8. Java stereoscopic mode

In Java3D, rendering modes are four in number **Table1**, and choosing a particular mode becomes very easy [9]. It is recommended that the programmer sticks to the first two modes.

8.1. Retained mode

This is the default rendering mode. The application remains in a continuous rendering state in this mode unless the method Canvas3D.stopRender of Java3D API (Application Programming Interface) is called. Also, in this mode, Java3D rendering will build a structure for handling geometry and operations related to 3D transformations[9].

8.2. Compiled Retained mode

Bitwise assignment in Java is one of the most confusing operations [9]. The following can be given as an example:

TransformGroup tg = new TransformGroup(); tg.setCapability(TransformGroup.ALLOW_TRANSFORM_WRITE); It should also be noted that there is a reference relationship between the classes: TransformGroup and Switch. These two classes can only be used in the first two modes and partially in the fourth mode. This is one of the main differences, which is not unimportant. The second mode gives Java the freedom to perform many optimisations. The most important of which are scene graphics compression and geometry grouping. This mode requires calling the compile() method of BranchGroup; this method optimizes the graph itself by grouping and sorting [9]. Another significant difference between the two modes is grouping, which is an advantage when there are many figures on the scene.

8.3. Immediate mode

Immediate mode - in this mode, applications intended for users have more freedom, which is at the expense of performance. This mode provides great flexibility, however rendering is at the lowest level when using it. A programmer using this mode only wants to take advantage of the geometry classes[9]. The first two modes (compression of geometry, grouping of geometry, and scene graph are more highly developed than the third.

8.4. Mixed mode

The fourth mode allows mixing the first two with the third mode. The mixed mode preserves the continuous rendering from the first two modes. For realization of the system is used this mode.

9. 3D printing and .OBJ file

3D printing, also known as additive manufacturing, creates three-dimensional solid objects corresponding to a given digital file. In this process, the object is created through successive layers of material until the 3D model itself is generated [13]. Each layer is a thinly sliced horizontal section of the object [13]. Before getting to the actual printing of the three-dimensional shape, it is necessary to go through the following steps.

1. Virtual design of the object, through a 3D modeling program (for a new object), through a 3D scanner (when an existing object needs to be copied) or, in this case, through the stereoscopic system.
2. The next step is to export to a .obj file
3. In the end, add to the printing device.

Table 1: Java 3D stereo Mode

Render stereo mode	Canvas3D	Stereo type
Retaind mode	*Canvas3D*	*Active3D*
Compiled – retained mode	*Canvas3D BranchGroup.compile()*	*Active3D*
Immediate mode	*Canvas3D.stopRenderer()*	*hybrid*
Mixed Immediate mode	*User Canvas3D: preRender(), postRender(), postSwap() and renderField()*	*Active3D*

Today, digitizing a real object into a 3D model becomes as easy as, for example, taking a photo. To prepare the .obj file for printing, the software slices the final model into hundreds of horizontal layers [13]. Next is uploading the 3D file to the printer; the object is created layer by layer [13]. The device makes the object by blending the layer. There are different types of technologies for working with 3D printing. One of the most used formats for storing polygon mesh information is the .obj format. One file can store data about indexed walls, normal vectors, and texture coordinates. The required data is stored in text format and

can be edited using a text editor, for example, Notepad. Each line contains the following characters: • "v" - vertices • "vn" – normal vector • "vt" – for texture • "f" – for the faces. Next comes the setting of specific values. Each of the sides is formed based on a triangle and a quadrilateral as shown on Figure 6 and Figure 7 . The row data for each wall contains a description of the vertices that make up the wall by specifying the vertices. Depending on the primitive being used.

10. Results

The software presented in this article, named StereoMV runs under Windows operating system. The object-oriented programming language Java, and more specifically, the Java3D library, was chosen as the means of implementation. The software interface uses JDK (Java Development Kit) version No. 8. This article presents 3D technologies that can be involved in teaching the discipline of stereometry, which is a branch of geometry in three-dimensional space. Besides the fact that technologies are divided into systems with and without immersion, they can be divided into two more groups: room-mounted and head-mounted. The main parameters of the stereoscopic system are:

- Software functionality

Two types of geometric figures are studied in the study of stereometry in secondary school:

1. Ribbed geometric figures – polygon, prism, cube, parallelepiped, and pyramid
2. Rotational bodies – cone, sphere, torus, and cylinder

The stereoscopic system can generate both types of geometric objects using the new boundary method for the first objects. For rotary bodies, the traditional extrusion method is used translational (cone and cylinder) and rotational (sphere and torus). The new program enables objects to be presented in three ways:

1. Transparent - makes it possible to understand the body's geometry, such as information about vertices, sides, angles, and polygons.
2. Solid – this presentation gives the finished look to figure
3. Mesh and solid body can be combined at the same time. It is possible to paint the walls of a polyhedra in different colors.

The program can perform algebraic calculations, such as finding faces, volumes, polygon side length, and more. This is made possible by a numerical calculation mode. Geometric objects can be viewed in different ways: 1. Traditional visualization 2. Full screen 3. Stereoscopic visualization (anaglyph, active, HMD, CAVE) 4. Mode for development of geometric objects

- Additional opportunity and contribution

Thanks to a new boundary method for generating geometric figures and a Java 3D library, it is possible to add additional features that are missing in stereometry software. For example, visualization can be given through virtual environment systems. StereoMV's most significant contribution is that it can convert geometric objects into a .obj file for printing or adding to an augmented or mixed reality device. The software enables arbitrary 3D models to be visualized through immersive and non-immersive virtual reality systems.

- Technical means for the realization of the stereo system

To operate with the system: VR systems are available with and without immersion. They differ in different ways from each other. For those whose stereo effect is relatively more minor, the following equipment is required: a special video card, such as NVIDIA QUADRO. Powered 3D monitor and glasses, as well as a projector. With passive technology, only a pair of anaglyph glasses are needed. Immersion systems require much more funds to invest in equipment, but simultaneously allow for great exciting experiences, unlike non-immersion systems.

- Objects for stereo visualization

The main Java3D API class that can be used to construct stereometric figures that are part of a stereo system is Shape3D, which defines the geometry and appearance of polyhedra and solids of rotation objects. In programming, these objects are called Java Nodes. A stereoscopic system can convert them to a .obj file. When developing the system in a program, it is necessary to have two methods: one creates the geometry, and the second the visualization.

- The stereoscopic system StereoMV

The core Java3D API class that enables stereo rendering via immersive and non-immersive systems is Canvas3D. It represents the canvas where objects from three-dimensional space are visualized. In Java3D, there are four modes for stereo visualization using active technology. The system uses Mixed Immediate Mode. The drawing of stereometry objects (edges and solids of rotation) are in Immediate mode, while static objects are in default mode, for example (app background, the three-dimensional coordinate system, the geometric network, axis and vertex labels. The createSceneGraph() method adds the objects that are part of the default mode. The geometric figures are in a class inheriting Canvas3D in which the mentioned four methods are added in Table №2.

- Appearance and mathematical description of geometric objects

In Java3D, several basic primitives can be used to construct complex geometric shapes, such as those from stereometry. When creating, for example, polyhedra such as a pyramid and a prism, it is necessary to traverse them in the same direction: clockwise or counter-clockwise.

LineArray – this array allows the creation of a mesh body by connecting a sequence of vertices. Here, the definition of a pair of vertices and topology is necessary.

TriangleArray – this array enables the construction of polygons that are formed by triangles. A sequence of triangles can generate any regular polygon. Here a sequence of three-vertex polygons is set, and the topology is defined.

QuadArray - objects are created using quadrilaterals. Arrays with four-vertex sequences and topology are defined.

In the results described in this article for constructing a quadrilateral prism - in this case is used all three types of primitives are used. The Appearance class from the Java3D API is used for object visualization. With its help, the following characteristics can be determined: color, type, and transparency of the considered primitives. The class using which the .obj file is generated is Geometryarray. It is possible to rebase the .obj file into a Shape3D object using the mentioned class. It is necessary to use an array of values of the vertices of the geometric figure. To generate a quadrilateral prism, the article uses only the information about the vertices and the primitive, which is used in the case of a triangle and a quadrilateral. The values are set using the COORDINATES parameter. If it is necessary to add more information to the creation of the .obj file, such as normal and texture ie "vn" and "vt" two more parameters -NORMALS and TEXTURE must be defined to the program code. Thanks to the normal, it is possible to color the objects with a specific color using a material. Here there are some crucial parameters, with the help of which the coloring is performed: ambient, emissive, specular, diffuse, and shininess. For the first four, the values range between 0 and 1. Based on these primitives, 3D geometric models (polygons, polygons, and surfaces based on curves) are constructed using a programming language or a 3D modeling program.

- Implementation of 3D active and passive technologies

For the realization of the 3D passive technology through the new stereoscopic system, it is necessary to do the following:

1. In the renderField() method, adding two identical geometric shapes with a slight deviation from each other. This is done as for the dynamically changing figures in Immediate Mode. Likewise, for static objects that are in retain() mode.

2. Next is setting the appearance for the two objects in blue and red, respectively

3. Finally, transparency is set, which varies between 0 and 1. The user can specify both the object's transparency and offset, representing the stereo effect. This will be done using a slider.

About 3D active Technology: In this technology, the two stereo images are created using the video card and the monitor. For this reason, a stereo effect cannot be defined here via an interface. It is set in advance by the stereo system developer, with values for the left and right eye, respectively: -0.01 and +0.01. It is possible to change the color of the system background through the interface. The background is white by default when the user is in traditional rendering mode. While in passive and active technology, the color is black. The goal is to make the objects stand out. Thanks to the Immediate Mode mode, it is possible to change the values of the parameters of the geometric figures using a text field, a slider, and keys. The proposed stereo system can generate, and at the same time visually present, both: java nodes in the case of polygons and polyhedra and files with the extension .obj. This way, arbitrary 3D models from various scientific fields can be visualized using immersive and non-immersive virtual reality devices. In this way, an accurate and complete picture of the studied object would be acquired. The system can import and export 3D files. Figure № 8 shows a regular hexagonal prism generated using geometry software. The object is represented in three ways: (3D active, 3D passive, and 3D traditional) perspectives. Also presented is a three-dimensional heart model for stereo visualization, downloaded from the following site: https://www.blendswap.com. It is preferable when observing a 3D file that the background offered by a virtual reality system is black. The aim is to make the model stand out in this way.

The visualization mode used through 3D active technology to implement the stereoscopic system is MixedImmediate. Geometric shapes created using the new boundary method are in Immediate mode. They are called once in the renderField() method. A Table №2 shows the methods for the 3D active technology mode. The most important of the four methods is renderField() - in this method, the 3D geometric objects that will be visualized by 3D active stereo visualization are called. It calls each of the objects that will be in Immediate mode.

A pair of glasses is required to realize 3D passive technology through anaglyph projection. With color, filters are blue and red for the right and left eyes, respectively. Each eye needs to have its own display for implementation through immersive HMD (virtual helmet) technology. For this purpose, two objects of type Canvas3D are created. Unlike planimetry, where the drawing of the geometric figure drawn on a piece of paper or board matches 100% with its original. Drawing the bodies from stereometry makes it very difficult for students, especially those who do not have well-developed spatial thinking. The problem is that the drawing transferred to a board, or a piece of paper can hardly be understood. This is where 3D technology, including (virtual reality, augmented reality, mixed reality, and 3D printing) comes to the rescue. In this way, students' interest in mathematics would be strengthened, and their spatial imagination would be further developed. The stereoscopic system realizes two types of geometric objects, which are studied by stereometry.

1. Edged geometric objects (polygon, prism, and pyramid) as shown on Figure 4 and Figure 5
2. Cylindrical bodies (cylinder, cone, sphere, and torus)

The system enables the geometric object to be presented solid and wireframe. The second presentation will enable the student to understand the object's geometry. At the same time, the dense presentation will allow its finished look. There is also a combination option between the two expressions.

Two methods are used to generate the geometric objects that are part of the stereoscopic system (traditional and new author's boundary method). Both cases aim to generate a regular polygon, which is the basis for drawing the pyramid and prism. The traditional method is based on trigonometry using (radius, cosine, and sine). Characterizes the polygon with number of vertices and radius. The new method defines the polygon by the number of vertices and the side length [14, 15]. It is based on Isaac Newton's method of limits and Cavalieri's indivisibles. It uses only the relation of parallel segments. At the same time, it gives a more accurate result since it is not necessary to use the technique of extrusion and trigonometry involved in generating objects in programming and, in this case, in stereometry. The objects created

in this way allow visualization through an immersive and non-immersive system and their representation as a .obj file. To illustrate the method, a regular hexagon will be considered as an example Table 3 and Figure 3. The same rules apply to polygons with the number of vertices increasing to infinity. The ratio of the parallel segments AB:FC =1:2 is determined, the length of the side of the polygon is represented by the parameter **a**. The regular polygon is placed in the center of the coordinate axis. The value of the height of the trapezoid FCBA is of value **h**. The Pythagorean theorem can determine it for triangle BCH; therefore, BH = a/2*sqrt(3). The height value is involved at four vertices of the polygon in the case of 6 vertices (A,B,D,E), defining values along the "**y**" ordinate (1,2,4,5).

- Prism construction

For the parametric representation of a regular hexagonal prism, it is necessary to do the following. The base is set, a polygon with 6 vertices with the length of the base through the parameter a. The parameter **h** sets the height of the prism. The bottom base is placed in the center of the coordinate axis, from which it follows that the values along the Z axis will be equal to zero. Vertices define the base numbered 1 to 6. The top base has vertices 7 to 12. The height rises along the z-axis. Therefore the values will be different from 0. Additionally, two more parameters h1 and h2 will be added, respectively. With the help of this, an inclined prism can be constructed; if they have a value equal to zero, the prism will be straight. An inclined prism follows at values where one parameter is zero and the other non-zero on Table 4.

Table 2: Java Mixed mode methods

Mixed – mode methods	***Pipeline***
preRender()	Allow you to prepare any data structures for rendering
postRender()	After 3D rendering operation
postSwap()	Called once the rendered frame has been made visible to the user
renderField()	Useful in stereoscopic imaging.

- Right straight/oblique prism

The algorithm for building a correct straight/slanted prism is as follows:

1. A regular prism is selected from the object panel. The type of the two bases is chosen, which are regular polygons with 3 to 6 vertices.
2. A length value is set for the main edge, then the multiwall is raised along the y-axis to set the height.
3. The abscissa (or Z coordinate) value is checked. If the value is a non-zero number, the final result will be a slanted prism; otherwise, the prism is straight.

The primitives used to generate a regular polygon are a triangle, a quadrilateral, and a line for mesh representation. The .obj file information in the article is based on vertices and walls as shown on figures 6 & 7. The traditional way to color is by setting a material. Thanks to the new boundary method, it is possible to color the walls using the mathematical description of the objects using adjacent vertices expressed by the side length of the polygon and a suitable primitive. A disadvantage of the system is that when generating the .obj file, it is necessary to use a material, i.e., its body is set in one color.

Table 3: Regular polygon

№ vertex	**X**	**Y**	**Z**
1	-a/2	a/2*sqrt(3)	0
2	a/2	a/2*sqrt(3)	0
3	a	0	0
4	a/2	-a/2*sqrt(3)	0
5	-a/2	-a/2*sqrt(3)	0
6	-a	0	0

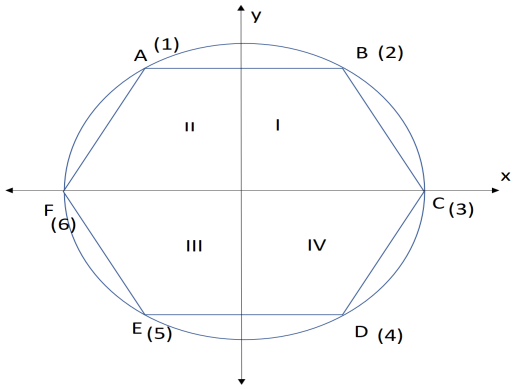


Figure 3: regular polygon with six vertex

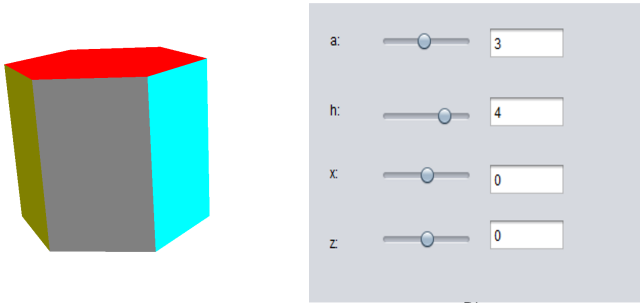


Figure 4: right regular prism

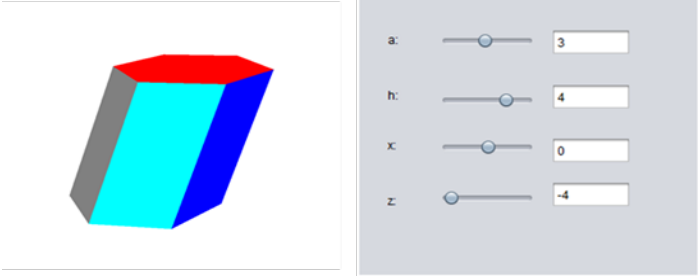


Figure 5: Tilted regular prism

```
# StereoMV
mtllib StereoMV.mtl

v 3 4 3
v 1 4 3
v 3 4 1
v 1 4 1
v 3 0 1
v 1 0 1
v 3 0 3
v 1 0 3
f 2 3 4
f 2 1 3
f 8 7 5
f 8 5 6
f 7 1 3
f 7 3 5
f 5 3 6
f 3 4 6
f 4 2 8
f 4 8 6
f 2 1 8
f 1 7 8
```



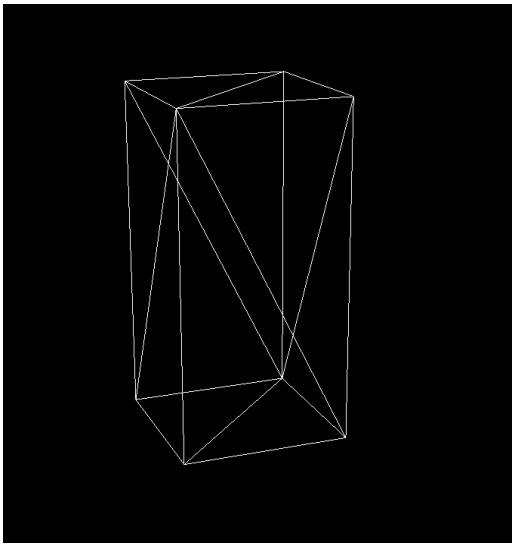


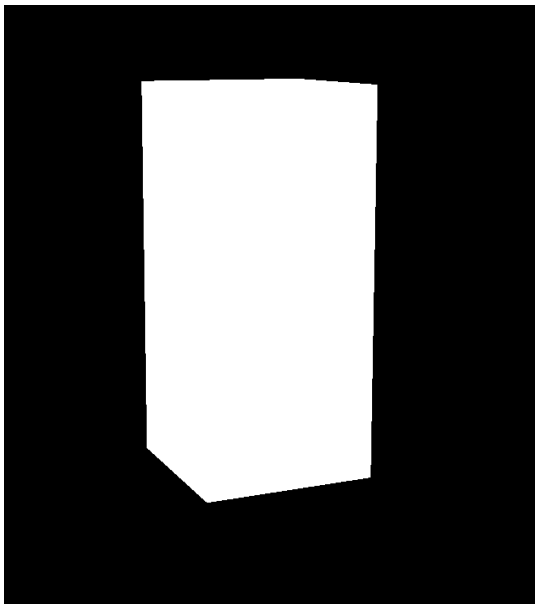
Figure 6: prism generated by primitive triangle

Table 4: Prism

Number of vertex	X	Y	Z
1	-a/2	a/2*sqrt(3)	0
2	a/2	a/2*sqrt(3)	0
3	a	0	0
4	a/2	-a/2*sqrt(3)	0
5	-a/2	- a/2*sqrt(3)	0
6	-a	0	0
7	-a/2-h1	a/2*sqrt(3)-h2	h
8	a/2-h1	a/2*sqrt(3)-h2	h
9	a-h1	0-h2	h
10	a/2-h1	- a/2*sqrt(3)-h2	h
11	-a/2-h1	- a/2*sqrt(3)-h2	h
12	-a-h1	0-h2	h

```
# StereoMV
mtllib StereoMV.mtl

v 3 4 3
v 1 4 3
v 3 4 1
v 1 4 1
v 3 0 1
v 1 0 1
v 3 0 3
v 1 0 3
f 1 3 4 2
f 7 8 6 5
f 1 7 5 3
f 2 4 6 8
f 2 8 7 1
f 3 5 6 4
```



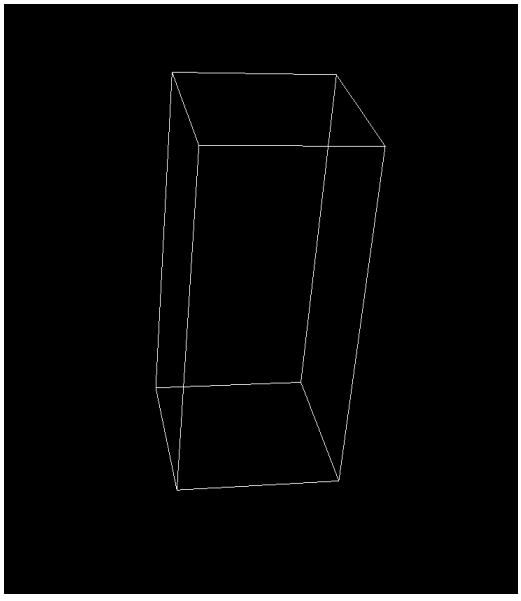
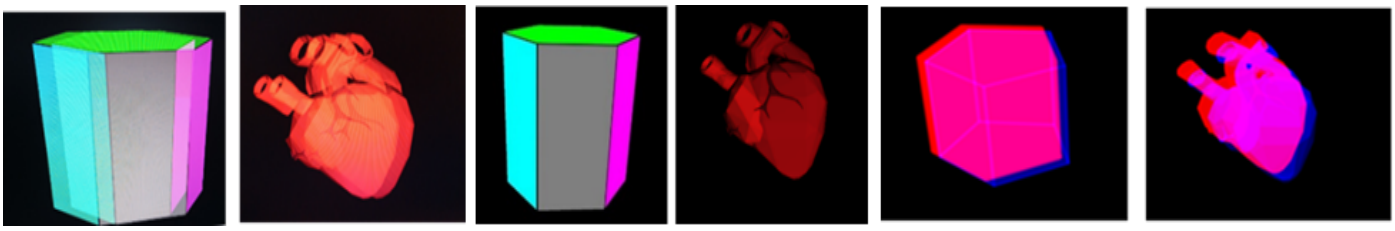


Figure 7: prism generated by primitive quad



(a)3D active b) 3D traditional c) 3D anaglyph

Figure 8: 3D Visualization

Table 5: Comparative analysis between 3D technologies realized by the new stereoscopic system

Type of Technology	Type of system	Advantage	Disadvantage
3D active/ 3D projector	Without immersion	- The colors are preserved - Great stereo effect - Short program code - Applies to .obj and nodes - Easy connection to the stereo system	- The stereo cannot be adjusted - Requires expensive equipment - There may be incompatibility with the video card
3D passive	Without immersion	- Applies to .obj and nodes - The stereo can be adjusted with a slider - Does not require expensive equipment	- Does not retain primary colors (blue / red only) - Small stereo effect - Long program code - Difficult to connect to the system
HMD/CAVE	With immersion	- The colors are preserved - Very large immersion effect - Applies to .obj and nodes - Easy connection to the stereo system	- They are very expensive - Long program code

11. StereoMV The stereoscopic system

StereoMV is the result of a dissertation work on the topic "Stereoscopic training system". It was originally developed for a non-immersive VR system using 3D active technology. Currently, the system can be visualized by an immersive system – an example being an HMD. A CAVE and semi-submersible FishTank system is also possible in the future. Table №5 provides a comparative analysis of the Virtual Reality systems discussed in this article. So far, the software is not yet finished. But at the same time, it continues to be further developed. The system can be used to visualize trimer models with VR from different domains. It also includes another direction - in medicine—the idea of creating stressful situations through a virtual environment. A mathematical analysis of recordings using a Holter device will be performed [16]. As well as a comparison of the data between recordings made by immersion and non-immersion systems. The goal is to be able to change heart rate and, from there, compare RR intervals to those at rest.

12. Conclusion

As is known, stereometry is a part of Euclidean geometry. Geometric figures are in three-dimensional space. Many students do not have a well-developed imagination to understand three-dimensional geometric objects. For the solution to this serious problem, a new mathematical software called StereoMV is presented in this paper. It allows students to view things in virtual reality. The main contribution of the system is that it can export geometric figures in files with the extension .obj. Then the three-dimensional model can be added to an augmented reality device or printed on a printer. The other innovation is that virtual systems with and without immersion can visualize the system. This will pique the interest of students. Studying the discipline of stereometry will become interesting, valuable, and enjoyable. This article performs a comparative analysis of virtual environment systems. Each of them differs in a specific way from one another. Some require more equipment, and others are relatively cheaper and offer an excellent stereo effect. Prolonged observation on the computer monitor through 3D active glasses can cause discomfort. Autostereoscopic monitors that do not require additional hardware come to the rescue. The Java programming language is used to implement the StereoMV stereoscopic system. Mixed immediate mode for stereo visualization through 3D active technology. Anaglyph projection is used for 3D passive technology with right and left eye color filters (blue and red), respectively. At this stage of system development, the immersive virtual reality that can visualize the software is the HMD (virtual helmet) device via *head* tracking. The system can export the objects in the .obj file extension. To add to a virtual or augmented reality device. As well as for 3D printing.

References

[1] P. Lebamovski, “Analysis of 3D technologies for stereo visualization,” in International Conference Automatics and Informatics, ICAI 2021 Proceedings, Varna: 206–209, 2021, doi:10.1109/ICAI52893.2021.9639534.

[2] Y. Lan, “Immersion, interaction and experience – oriented learning: Bringing Virtual reality into FL learning,” Language Learning & Technology, **24**(1), 1–15, 2020.

[3] A. Yeh, R. Nason, “Toward a Semiotic Framework for Using Technology in Mathematics Education : The Case of Learning 3D Geometry,” International Conference on Computers in Education Melbourne, Australia, 2004.

[4] M. Nikolova, P. Byalmarkova, R. Radoeva, “3D technologies as electronic educational resource 3D technologies as electronic educational resource,” 2015(November).

[5] D. Minkovska, “Virtualna realnost – edin nov pogled kam savremennoto obuchenie,” **56**(1), 64–72, 2013.

[6] J. Kang, B. Liu, “Application of 3D stereoscopic visualization technology in casting aspect,” China Foundary, (August), 308–313, 2014.

[7] N. Kim, M. Erdenebat, “3D Display Technology,” ©2013 Old City Publishing, Inc., June, 73–95, 2013.

[8] Y.Liang, H. Zhang, "Computer Graphics using Java 2D and 3D", Pearson Education, 2003.

[9] A.Terrazas, J. Ostuni, M. Barlow, "Java Media APIS: Cross-platform Imaging, Media, and Visualization", Sams, Indiana, 2002.

[10] K. Kerestes, “Making the anaglyph map,” 2000.

[11] "3D Технологии - Pro Screens, Oct. 2022", [Online]. Available:http://www.proscreens.eu/bg/3d-technologies

[12] M.G.O. Brien, R.M. Levy, A. Orich, “Virtual immersion : The role of CAVE and PC technology,” CALICO Journal, **26(2)**(January 2009).

[13] J.Q. Al-maliki, A.J.Q. Al-maliki, “The Processes and Technologies of 3D Printing,” International Journal of Advances in Computer Science and Technology, **4**(10), 161-165, October 2015.

[14] P. Lebamovski, “THE EFFECT OF 3D TECHNOLOGIES IN STEREOMETRY TRAINING,” CBU International Conference Proceedings, **202**, 68–74, 2021, doi:10.12955/pns.v2.155.

[15] P. Lebamovski, E. Petkov, “USAGE OF 3D TECHNOLOGIES IN STEREOMETRY TRAINING,” CBU International Conference Proceedings, **202**, 139–146, 2020, doi:10.12955/pss.v1.61.

[16] E. Gospodinova, M. Gospodinov I. Domuschiev, N. Dey, A. Ashour, “Nonlinear analysis of heart rate variability in Type 2 diabetic patients,” Fractal Geometry and Nonlinear Analysis in Medicine and Biology, **1**(4), 2016, doi:10.15761/fgnamb.1000123.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 4, 42-48 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Computer Vision Radar for Autonomous Driving Using Histogram Method

Hassan Facoiti[*,1], Ahmed Boumezzough[1], Said Safi[2]

[1]*Applied Physics and New Technologies Team, Sultan Moulay Slimane University, Beni Mellal, Morocco*

[2]*Laboratory of Innovation in Mathematics, Application and Information Technologies, Sultan Moulay Slimane University, Beni Mellal, Morocco*

ARTICLE INFO

Article history:
Received: 24 May, 2022
Accepted: 12 July, 2022
Online: 27 July, 2022

Keywords:
Radar
Autonomous driving
ADAS
Computer vision
Line lane detection
Histogram

ABSTRACT

Mobility is a fundamental human desire. All societies aspire to safe and efficient mobility at low ecological and economic costs. ADAS systems (Advanced Driver Assistance Systems) are safety systems designed to eliminate human error in driving vehicles of all types. ADAS systems such as Radars use advanced technologies to assist the driver while driving and thus improve their performance. Radar uses a combination of sensor technologies to perceive the world around the vehicle and then provide information to the driver or take safety action when necessary. Conventional radars based on the emission of electromagnetic and ultrasonic waves have been consumed in the face of the challenges of the constraints of modern autonomous driving, and have not been generalized on all roads. For this reason, we studied the design and construction of a computer vision radar to reproduce human behavior, with a road line lane detection approach based on the histogram of the grayscale image that gives good estimates in real-time, and make a comparison of this method with other computer vision methods performed in the literature: Hough, RANSAC, and Radon.

1 Introduction

Advanced driver assistance systems (ADAS) are multiplying with the emergence of new technologies and the fall in the price of sensors and computers. The ADAS systems not only act on the vehicle in the event of an emergency but also make it easier to drive the vehicle by delegating certain tasks to the vehicle. ADAS are electronic systems that have access to the restitution, traction, braking, and steering components of the vehicle, thus allowing drivers to benefit from assistance and/or to temporarily delegate driving to an automatic co-pilot under certain conditions of traffic [1].

In the literature on driver assistance systems, there are two types: informative systems and active systems. Informative ADAS: Anti-collision system, LDW (Lane Danger Warning), BSD (Blind Spot Detection system), Park assist, and DMS (Driver Monitoring System). Active ADAS: EBA (Emergency Brake Assist), AEB (Automatic Emergency Braking), VSL (Variable Speed Limiter), VCR (Variable Cruise Regulator), ACC (Adaptive Cruise Control), LKA (Lane Keeping Assist), and LPA (Lane Positioning Assist) [2].

The Lane Danger Warning (LDW) warns the driver in the event of involuntary lane crossing. The device uses for this an infrared sensor or a camera pointed toward the ground in front of the car. The sensor locates the white lines to the left and right of the vehicle and calculates the relative position of the vehicle in relation to these lines. As soon as the car (bites) the line, an alert is triggered. For this, we were interested in developing this device to create a computer vision radar to take the raw images and then processed, them so that an automatic driving system could be built [3].

Images are important information vectors that represent a considerable amount of data. Their analysis makes it possible to make the decision in terms of space management and locate the areas of interest and carry out processing in real-time [4].

Computer vision tools enable rapid and automatic extraction of qualitative and semantic information, performing relevant information extraction and intelligent interpretation operations on images [5]. For this purpose, the new radars for industrial use or onboard modern cars migrate to the use of computer vision technology thanks to its reliability and cheaper than other types of radars, since they are based on cameras.

In this paper, we present a histogram method processing image data provided by the camera to detect line lanes and compared it with different methods (HOUGH, RANSAC, and RADON) [6], [7]. This paper is organized into 4 sections: In section 1, we present the general processes to detect line lanes. In section 2 we illustrate how

[*]Corresponding Author: Hassan FACOITI, Hassan.facoiti@gmail.com

to find lanes from the track with the histogram method. In section 3, we present the detection of line lanes by the three methods indicated and their obtained results. In section 4, we compare data from experimental results to the presented methods on videos obtained under real conditions.

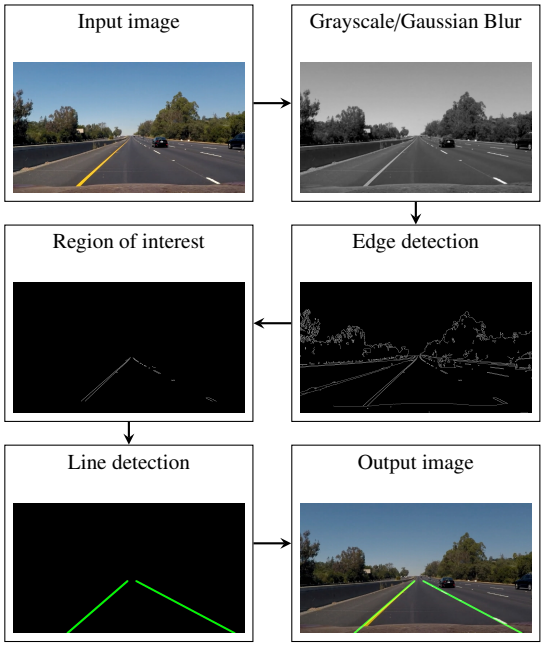
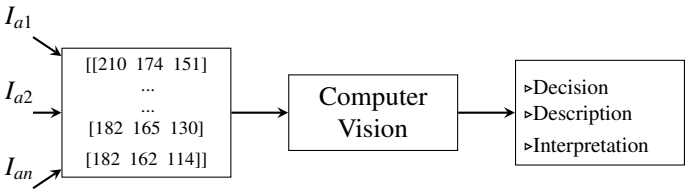


Figure 1: Diagram of the roadway detection & estimation method

2 General approach to the extraction of road marking

In the Figure 1 we illustrate the step-by-step the image processing methodology of the roadway detection and estimation method that achieves the objectives proposed in this paper. This processing is programmed by hybrid languages C++, python-OpenCV, and Matlab.



I_{an} : *analog images of the environment*

Figure 2: Block diagram of a computer vision system

2.1 Computer vision

The development of computing machines makes it possible to construct a vision system for visual perception capable of naming the objects that surround it, in order to provide the necessary knowledge without ambiguity.

The computer vision system Figure 2 is based on the processing of a sequential sequence of digital images (in the form of pixels) taken by a camera (raw information). Each pixel of the digital image represent an information giving an indication of the amount of light and color coming from the surrounding space [8].

2.2 Frame per second

The general challenge of vision applications is costly in computational time due to the complexity of the algorithms developed. For this reason, the first step to reduce the calculation time is to reduce the processed data depending on time, for this, we have adapted our image capture program to take one image (frame) per second without affecting the data.

2.3 Perspective transformation (Bird-Eye-View)

The Bird-Eye-View transformation technique is to generate a top view perspective of an image, as illustrated in Figure 3. This technique can be classified in digital image processing as a geometric image modification. The Bird-Eye-View transformation can be divided into three stages. First, represented the image in a system of offset coordinates, second rotate the image, third project the image on a two-dimensional plane. The basic diagram of the transformation is given in Figure 4.

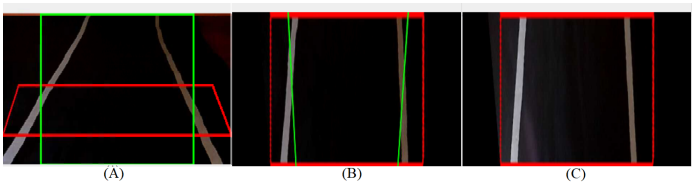


Figure 3: (A) Image originale, (B) Bird-Eye-View image, (C) Output image

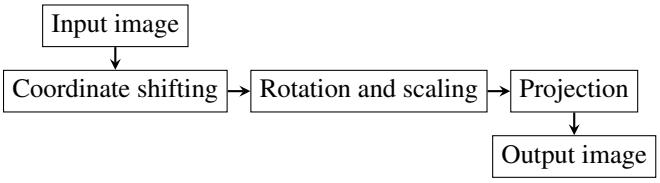


Figure 4: Bird-Eye-View block diagram

2.4 Edge detection

Edge detection is based on Canny Filter, it is a computational approach to detection of marking contours, which represent a fast approach but sensitive to noise because it is accentuated by derivation [9], [10].

After performing a Gaussian smoothing on the image, in order to remove the impurities, then we apply the Canny filter, which calculates the norm of the intensity gradient and the angle of the normal to the gradient (direction of the contours) for each pixel of the smoothed image.

The formulas to apply for each pixel are described as follows:

$$G_r(x,y) = \sqrt{G_x^2(x,y) + G_y^2(x,y)} \quad (1)$$

$$\theta = \pm \arctan(\frac{G_x}{G_y}) \quad (2)$$

$G_r(x,y)$ represent the function of the intensities of pixels, in all directions, x and y.
G_x : represents the gradients in x.
G_y : represents the gradients in y.
G_x et G_y are two convolution masks defined previously, one of dimension 3 x 1 and the other 1 x 3:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \end{bmatrix} \quad G_y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \quad (3)$$

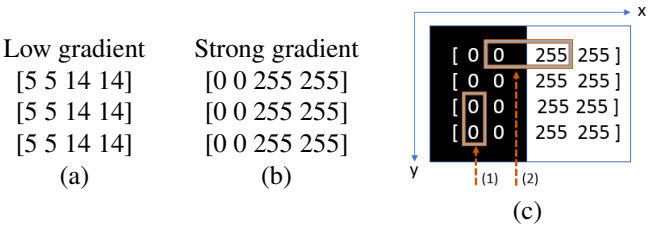


Figure 5: (a) Low gradient. – (b) Strong gradient. – (c) Corresponding image of strong gradient (b). – (1) No change, derivative = 0. – (2) large change, large derivative (gradient).

Canny's method is based on the derivation variation of $G_r(x,y)$ to determine the image contours, it draws the edge with a big change of intensity (large gradient) and is displayed in the form of white pixel outlines Figures 5 and 6.

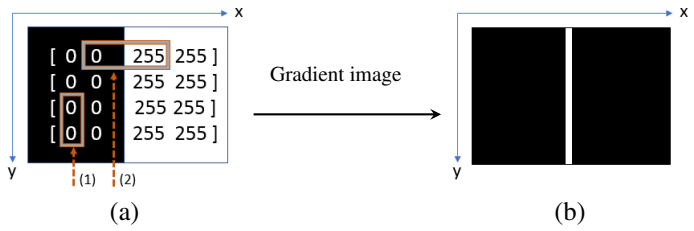


Figure 6: (a) Image. – (1) No change, derivative = 0. – (2) large change, large derivative (gradient). – (b) Image contours in white pixels.

The completely black areas Figure 6(b) correspond to a small variation of intensity between the adjacent pixels. While the white line represents a region of the image where the intensity varies considerably and exceeds the threshold. The threshold is to refine the filtering of the weak contours and to keep only the significant contours, by using two thresholds: High (S_H) and Low (S_L). If the value of a contour is higher than the highest threshold, it is preserved. And if the threshold is lower than the low threshold, the corresponding pixel becomes black.

At this location, two transformations are made of the image obtained after perspective transformation Figure 7(A) to detect the edges, the first is a threshold operation Figure 7(B) and the second is the Canny method Figure7(C). At the end, the two resulting images are confused to guarantee the detection of the edges of the image Figure 7.

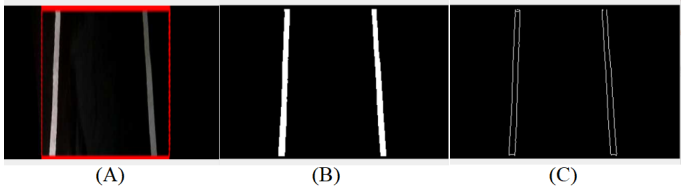


Figure 7: (A) perspective Image, (B) Grayscale image, (C) Canny image

2.5 *Region of interest*

Automatic detection of the region of interest on an image often has multiple areas is a difficult problem to solve. The proposed method Figure 8 is to determine the region of interest is based on the creation of a digital mask of the image, this mask consists of two types of pixels: 0 represent the black and 255 represent the white, a triangle (or rectangle) pixels of values 255 identical to the region of interest that we are trying to extract and the rest of the pixels have the value 0 Figure 8(b). The goal is to do the and-logic operation between the bit of each pixel homologous to the image and this mask, thus masking the complex image to show only the region of interest corresponding to the mask Figure 8(c).

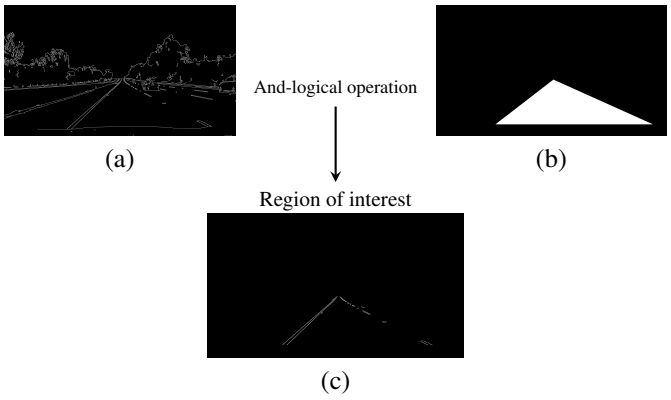


Figure 8: (a) Canny image. – (b) Image mask. – (c) Region of interest.

As the pretreatment steps mentioned and detailed in [8], a region of interest (the red rectangle) is created as illustrated in Figure 9.

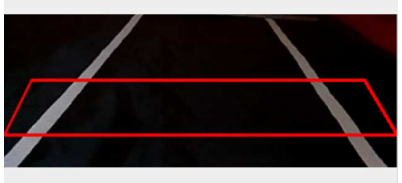


Figure 9: Region of interest

3 Find Lanes from Track using the histogram method

3.1 Histogram method

The histogram method makes it possible to find the exact position of the road line lane from the grayscale image compared to the position of the radar system. This method is based to create a region of interest forms a rectangle in the final image Figures 10(1) and 11(C), divide into different strips.

Algorithm 1: Histogram algorithm

Input : FrameFinal;
initialization : DynamicAreas.size();
for *int i=0; i < FrameFinal.size().widht ;i++* **do**
  RegionOfInterestLane=FrameFinal(rectengle.size());
  divide(255, RegionOfInterestLane,
   RegionOfInterestLane);
  push all the intensity valueus in DynamicAreas;
  histogrameLane.push_back(int)(
   sum(RegionOfInterestLane)[0]);
end
LeftLaneParametre= max_element(histogrameLane.begin(),
 histogrameLane.lenght()/2);
LeftLanePosition=
 distance(histogrameLane.begin(),LeftLaneParametre);
RightLaneParametre=
 max_element(histogrameLane.lenght()/2,
 histogrameLane.end);
RightLanePosition=
 distance(histogrameLane.begin(),RightLaneParametre);
Output : Final_image, LineLanePosition;

In one step, there will be total pixels and each pixel has an intensity of 0 if is black or 255 if is white, we can calculate multiply each step of pixels by its intensity and get the resultant intensity and store this intensity in dynamic areas Figure 10(2), then replace each intensity equal to 255 by X Figure 10(3)(4) and save the coordinates value to find the line lane position. Finally, draw these lines in green and their average on the final image Figure 11(C).

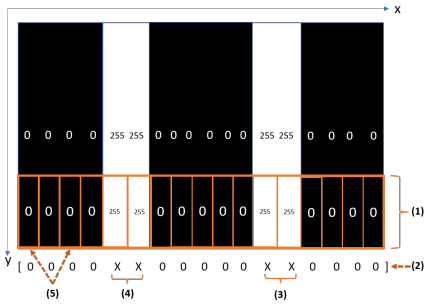


Figure 10: (1) Region of interest, (2) Dynamic areas, (3) Right lane position, (4) Left lane position , (5) Pixels strips

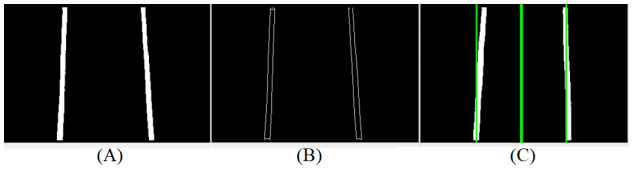


Figure 11: (A) Grayscale image, (B) Canny image, (C) Final image

3.2 Calibrage

We take the average of the lines detected as a reference, it is the instruction of our automatic closed-loop system which must be followed by the car, it is the principle of autonomous driving. Each time the car deviates from this average (the deviation is indicated by a blue line), the system asks to perform the calibration Figures 12, 13 and 14.

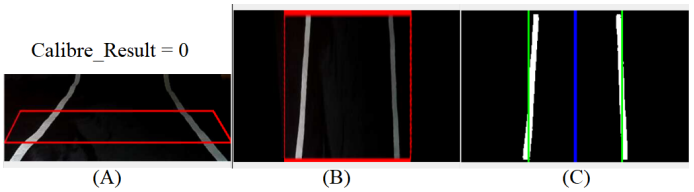


Figure 12: The right position: (A) Original image, (B) Bird eye viw image, (C) Result Image

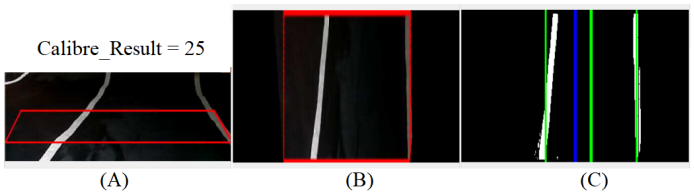


Figure 13: Deviation to the left: (A) Original image, (B) Bird eye viw image, (C) Result Image

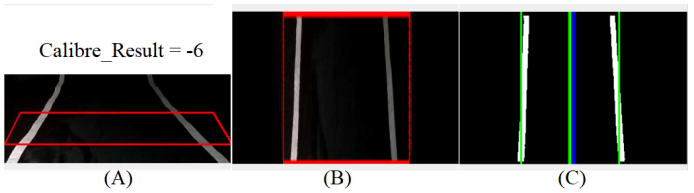


Figure 14: deviation to the right: (A) Original image, (B) Bird eye viw image, (C) Result Image

4 Computer vision methods find lanes from the track

4.1 HOUGH Transform

The Hough Transform is an efficient form in recognition tool, the practical application to detect in a camera image the presence of parametric curves from a set of characteristic points, essentially uses the spatial information of the characteristic points, that is, their

position in the image. The generalized Hough transform can detect other forms [7].

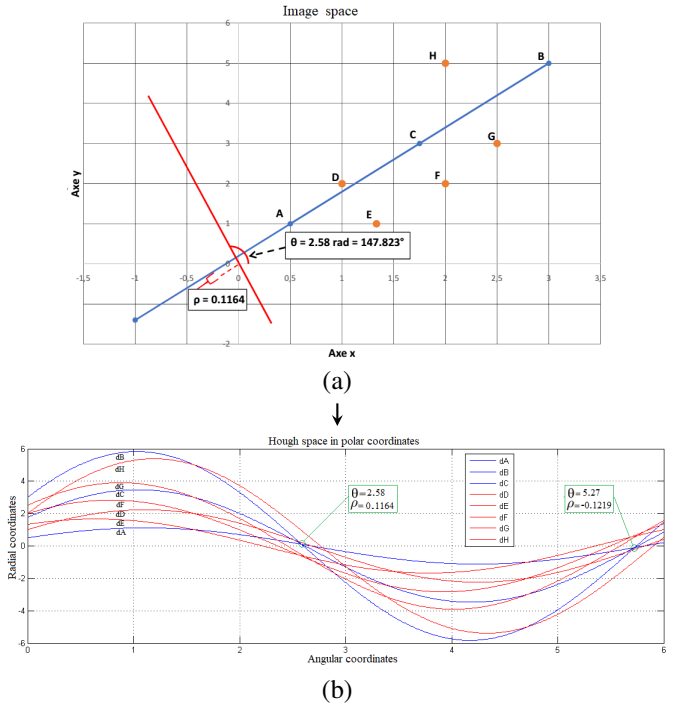


Figure 15: Principle of Polar System. – (a) Distribution of noisy points. – (b) Hough Transform

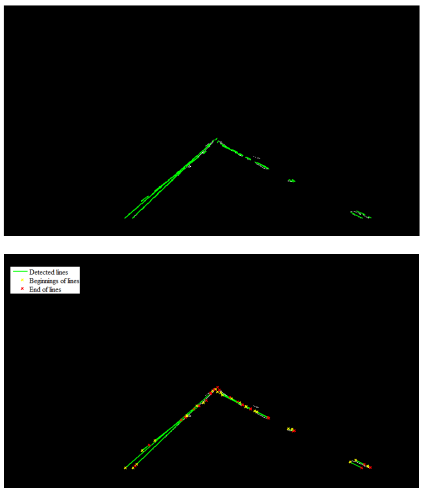


Figure 16: HT Detected lines lanes

The Hough transform algorithm based on the polar system uses an accumulator matrix that represents the space (ρ,θ) [11]–[14], of dimensions (L,C) where L is the number of possible values of ρ and C the number of values of θ. In a distribution of noisy points (x_i, y_j) of the Figure 15(a), each point becomes a sinusoid of the equation "(4)" in the parameter space (ρ,θ). The Figure 15(b) shows, at the end of the accumulation, the sinusoids corresponding to the points of the same line intersect at the point (ρ,θ) setting this line.

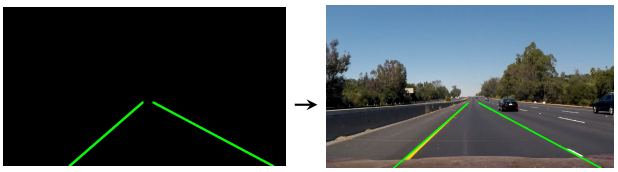
$$x\cos(\theta) + y\sin(\theta) = \rho \tag{4}$$

The Figure 16 shows the execution of this algorithm on the image of the regions of interest Figure8(c), so that the accumulator correspond to the votes, which obtained the highest values of the parameters correspond to the lines of the contours of the road way in the treated image.

4.2 *Optimization*

The detection of roadside contours by the Hough transform is robust as well as the information of the positions of the lines of these contours, this information must be optimized, so that to give a single line in each limit of the road lane Figure 17(a), the objective is to invest in the engineering of Advanced Driver Assistance Systems (ADAS) to build an embedded control system of the vehicle position. The optimization algorithm is summarized in "Algorithm. 2".

Finally, the experimental results as they appear in the output image Figure 17(b) show that the detection algorithm followed in this paper improves visibility of the roadway and reduces noise.



(a) (b)

Figure 17: (a) Optimization of line detection. – (b) Output image

Algorithm 2: Optimization algorithm

```
Slope_interception_left = [ ];
Slope_interception_Right = [ ];
for line in linesDET do
    % linesDET: the lines detected by the transform;
    % each linesDET detected is a 2D array containing the
      coordinates in [[x1, y1, x2, y2]];
    Convert 2D coordinates of linesDET to 1D;
    % 2D [[x1, y1, x2, y2]] ⇒ 1D [x1, y1, x2, y2];
    Calculate the slope a_i and the interception b_i of each
      linesDET;
    if slope a_i < 0 then
        % note that the y axis is reversed. Values increase in
          descending;
        Slope_interception_left = [a_i,b_i];
    else
        Slope_interception_Right = [a_i,b_i];
    end
end
Left_line_parameter = average.(Slope_interception_left);
Right_line_parameter = average.(Slope_interception_Right);
```

4.3 Probabilistic voting method RANSAC

The RANSAC (Random Sample Consensus) method is a probabilistic voting method based on the use of minimal data to accurately estimate the parameters of a model, even if the data is noisy. This method has been proposed to reduce the calculation time of conventional voting methods such as the Hough Transform. The algorithm of this method is based on a number of iterations, at each turn, it randomly selects a subset of data (two random points), then finds the model for the selected data, then tests all the data by model and determine the relevant points according to the threshold. In the end, if the new model is better than the best model (based on the number of relevant points), then the new model becomes a best model. From these processes, the estimate preserves the parameters of the searched lines. Let "p" be the probability of obtaining a good sample, "s" the minimum number of points in a sample to estimate the parameters of the model and "r" the probability to have a valid point in all selected contour points. The minimum number "m" of random draws necessary to have a probability of the correct parameters is as follows:

$$m = \frac{log(1-p)}{log(1-(1-r)^s)} \tag{5}$$

Finally, in Figure 18 the vehicle circulation lane is determined by exploiting the RANSAC algorithm after 12 iterations, the execution of this algorithm on the same image returns different results because it is based on the random selection of the data, and all the results obtained are valid. The choice of 12 iterations amounts to minimize the calculation time in parallel to obtain a good estimate. After optimizing the results with the optimization algorithm "Algorithm. 2", we obtain a single line in each path limit which is clear in Figure 18(c).



(a) (b) (c)

Figure 18: (a) and (b):Track markings detected by our RANSAC algorithm after 12 iterations. (c):Grouping of line segments.

4.4 Radon Transform

The Radon transform is a mathematical technique developed by Johann Radon, it is defined on a space of lines L in R^2 according to two arguments (ρ, θ). The application of this transform on a two-dimensional function I(x,y) (image) is based on several projections of the image under parallel beams under different given angles, in order to calculate the line integrals of this beam in directions specified. The resulting image R(ρ, θ) of the projection is sum of intensities of the pixels in each direction, can be translated by [15]–[17]:

$$R(\rho,\theta) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I(x,y)\delta(x\cos(\theta)+y\sin(\theta)-\rho)\,dxdy \tag{6}$$

The robustness of the Radon transform lies in its ability to detect lines in noisy images. Like the Hough transform, it can also transform lines into easy-to-solve pics that correspond the parameters of the lines in the outlines of the image. This property of the Radon transform allows us to identify and detect the contour lines of the roadway in the image of the region of interest. After the optimization, we find very important and valid results Figures 19 and 20.

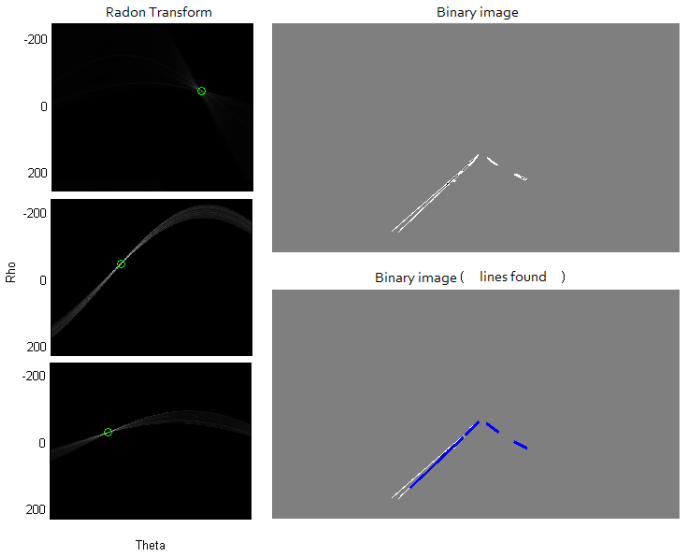
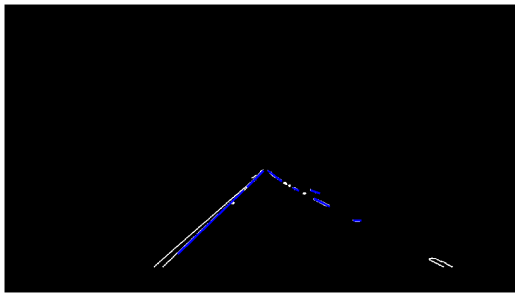
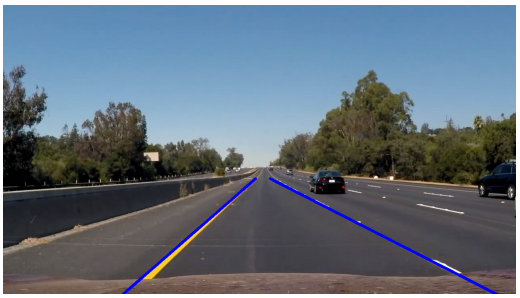


Figure 19: Radon algorithm lane detection



(a) . Detected lines



(b) . Grouping of lines

Figure 20: Radon algorithm results

5 Experimental data

We have based on the error (%) (false alarms) of the estimate of the slope and the interception of the detected lines lanes to prove the robustness of the algorithms presented and compare them, by using several tests on sequences of images (videos) of resolution (1280 x 720) in different conditions. Each image sequence has an almost constant width of the line lanes and also has real parameters (slope, interception). The false alarm rate determined after a series of experiments varies acceptably among the four methods compared to the real values Figure 21.

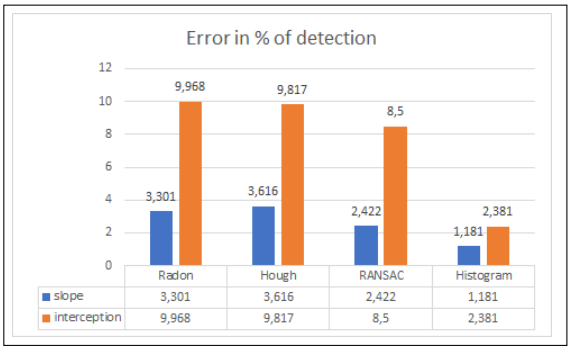


Figure 21: Detection error percentages for each of the four methods

Note, however, that the method based on the histogram presents a minimum error and minimum computing time compared Figure 22 to RANSAC, HOUGH, and RADON methods, due to the data processing being done at the rectangle of the region of interest after generating a perspective view from the above an image.

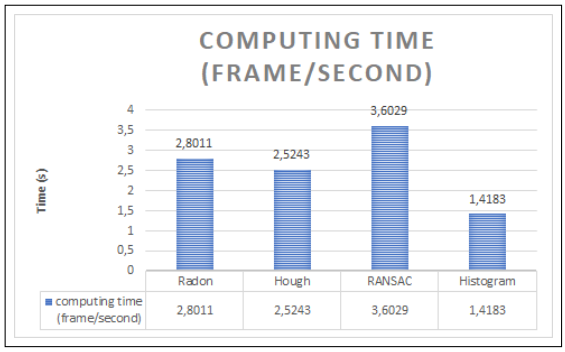


Figure 22: Computing time (frame/second)

6 Conclusion

The comparative study has shown that the proposed vision algorithms (Histogram, Hough, RANSAC, and RADON) are robust to detected contour points with a predefined model describing the curvature of the road, and are quite efficient and have a very efficient accuracy rate despite the fact that noise is present in the images of the road. But the Histogram method presents more efficiency compared to other methods.

The use of these methods is necessary to treat the real environment, but in terms of execution time is very important, which poses a problem of synchronization with the real time of a sequence of images.

In order to complete our study, we want to continue to improve vision algorithms able to be generalized to all roads and detect obstacles and remain robust to different conditions and compatible with real-time processing.

Conflict of Interest The authors declare no conflict of interest.

References

[1] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller and H. Winner, "Three Decades of Driver Assistance Systems: Review and Future Perspectives," in IEEE Intelligent Transportation Systems Magazine,**6**(4), 6-22, winter 2014, doi: 10.1109/MITS.2014.2336271.

[2] M. Nolwenn. "From driving assistance systems to connected autonomous vehicles". 2019. Doctoral thesis in Automation, Production, Signal and Image, Cognitive Engineering. Bordeaux.

[3] R.N. Mahajan and Rashmi A. M. Patil, "Lane departure warning system". International Journal of Engineering and Technical Research, **3**(1), 120-123, 2015.

[4] C.J. Jacobus, D. Haanpaa.:All weather autonomously driven vehicles. U.S. Patent **9**, 989-967. Jun 2018.

[5] M. Olivier, Y. Poncet. "Computer vision techniques applied to aerial images of floodplains".2001.

[6] M.A. Rahman, M. F. I. Amin and M. Hamada, "Edge Detection Technique by Histogram Processing with Canny Edge Detector," 2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII), 128-131, 2020, doi: 10.1109/ICKII50300.2020.9318922.

[7] M. Henri.:A panorama of the Hough transformation. Signal processing. **2**(4), pp 305-317, 1985.

[8] H. Facoiti, A. Boumezzough and S. Safi. "Comparative study between computer vision methods for the estimation and detection of the roadway", 2021 7th International Conference on Optimization and Applications (ICOA), 1-6, 2021. doi: 10.1109/ICOA51614.2021.9442633

[9] B. BESBES, Christele LECOMTE, Peggy SUBIRATS.:New lane line detection algorithm. XXIIe colloque GRETSI, Signal and image processing, September 2009.

[10] C. John, A computational approach to edge detection. Readings in computer vision. Morgan Kaufmann,184-203, 1987.

[11] R.O. Duda, E. Hart, Use of the Hough transformation to detect lines and curves in pictures. No. SRI-TN-36, Communication of the ACM, 15, 11-15, janvier 1972.

[12] H. Facoiti, S.Safi, A.Boumezzough.:Analyse, caractérisation et conception d'un système radar anticollision embarqué dans les véhicules. sciencesconf.org:icsat-2020:326086, 13-18, International Colloquium on Signal, Automatic control and Telecommunications, Caen, France, juin 2020.

[13] S. Haykin.: Neural Networks and Learning Machines. 3rd Edition, New York 2008.

[14] C. John.:A computational approach to edge detection. Readings in computer vision. 184-203, 1987.

[15] J.X. Wu, S.V. Kucheryavskiy, Linda G.Jensen, Thomas Rades, Anette Müllertz, Jukka Rantanen.:Image Analytical Approach for Needle-Shaped Crystal Counting and Length Estimation. Crystal Growth and Design, Bind 15, s.4876-4885 2015.

[16] H. Wang, Q. Chen.:Real-time lane detection in various conditions and night cases. Intelligent Transportation Systems Conference, 1226-1231, Canada, September 2006.

[17] K. Osman, J. Ghommam, M. Saad.:Vision Based Lane Reference Detection and Tracking Control of an Automated Guided Vehicle. 25th Mediterranean Conference on Control and Automation (MED), Valletta, Malta, July 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 4, 49-58 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

ARAIG and *Minecraft*: A Modified Simulation Tool

Cassandra Frances Laffan[*,1], Robert Viktor Kozin[1], James Elliott Coleshill[1], Alexander Ferworn[1], Michael Stanfield[2], Brodie Stanfield[2]

[1]*Computational Public Safety Lab, Department of Computer Science, Toronto Metropolitan University, Toronto, M5B 1Z4, Canada*

[2]*Inventing Future Technologies Inc. (IFTech), Whitby, L1N 4W2, Canada*

ARTICLE INFO

Article history:
Received: 10 May, 2022
Accepted: 19 July, 2022
Online: 27 July, 2022

Keywords:
Haptics
Octree
Pathfinding
*A**
Search and Rescue
Digital Games

ABSTRACT

Various interruptions to the daily lives of researchers have necessitated the usage of simulations in projects which may not have initially relied on anything other than physical inquiry and experiments. The programs and algorithms introduced in this paper, which is an extended version of research initially published in ARAIG And Minecraft: A COVID-19 Workaround, create an optimized search space and egress path to the initial starting point of a user's route using a modification ("mod") of the digital game Minecraft. We initially utilize two approaches for creating a search space with which to find edges in the resulting graph of the user's movement: a naive approach with the time complexity of $O(n^2)$ and an octree approach, with the time complexity of $O(nlogn)$. We introduce a basic A algorithm to search through the resulting graph for the most efficient egress path. We then integrate our mod with the visualization tool for the "As Real As It Gets" (ARAIG) haptic suit, which provides a visual representation of the physical feedback the user would receive if he were to wear it. We finish this paper by asking a group of four users to test this program and their feedback is collected.*

1 Introduction

As most people, readers and the general populace alike, are aware, the COVID-19 pandemic has hampered the plans of many researchers [1]; moreover, ongoing supply chain interruptions have compounded this issue. From research using specialized equipment to field testing, the past two and a half years have been full of ingenious "workarounds" to both workplace restrictions and material shortages. We feel our current project is no exception: in a time where our research, as introduced in our previous work [2], should be applied to tangible hardware and tested in physical environments, access to our lab and materials is limited. The need for corporeal results is clear and thus, we propose and implement a workaround in this paper. This work is an extended version of previously published conference proceedings: *ARAIG and Minecraft: A COVID-19 Workaround* which may be found here [3].

As touched upon in [2], the field of search and rescue is going to experience various changes over the coming decades. Canada, home to both this project and its researchers, is one of many countries currently feeling the metaphorical and literal heat of climate change's effects [4]. Natural disasters, such as the forest fires plaguing the Canadian prairie provinces [5, 6], will become more prevalent across the country and globally before climate action takes effect [4]. Thus, now is the time to act in terms of curbing future disasters while empowering emergency workers to safely and efficiently respond to high impact, low (though increasing) frequency events. While we cannot influence government actions, domestically or globally, we can certainly do our best to assist in preparing our first responders for future disasters.

One of the most dangerous situations a firefighter may face in disaster environments, particularly in enclosed spaces such as homes and buildings, is potentially growing disoriented and thus, lost [7]. In fact, there is ongoing research on this exact issue [7], as this is still an unsolved problem. The motivation for this research, in conjunction with the increasing incidence of natural disasters, is this issue: how can we utilize modern technology in a lightweight fashion to assist firefighters in navigating out of these low visibility environments? Due to the restrictions the pandemic has brought about, as well as ongoing supply chain issues, the initial answer to this question comes in the form of simulation.

Minecraft is a digital game which focuses on the exploration, and building, of randomly generated environments. The graphics, game mechanics and game world are simple: the player is placed, without warning or preamble, into the *Minecraft* environ-

[*]Corresponding Author: Cassandra Frances Laffan, George Vari Engineering and Computing Centre, 245 Church Street, +1 (647) 983-4070 & Cassandra.Laffan@ryerson.ca

ment, which is composed of "blocks", not dissimilar to voxels. The game is centred around navigating, manipulating and accumulating blocks while exploring a procedurally generated world. Since the world of *Minecraft* is so simple, and there are numerous frameworks which support modding the game (these programs will be referred to as "mods"), we have found it suits the above requirements for a simulation medium given our problem domain. Moreover, the game has a well established and active "modding" community [8, 9].

2 Related Work

There has been considerable research into simulations over the past few years, particularly for reasons outlined in this paper's introduction. It is of note, however, that the usage of digital games as an avenue for simulating experiments predates the pandemic and supply shortages. "Project Malmo", a project published in 2016 by Microsoft, is an earlier example of this. The authors write an Application Programming Interface (API) and abstraction layer for *Minecraft*. With it, they train an artificial general intelligence (AGI), to complete various tasks [10, 11]. The inclusion of [10] in our *Related Work* is due to their motivations for using *Minecraft* as their training medium:

- *The environment is rich and complex, with diverse, interacting and richly structured objects [10, 11].* The platform must offer a worldspace with which a user, or otherwise autonomous agent, can interact. This worldspace must be varied and robust.
- *The environment is dynamic and open [10, 11].* The platform needs to offer unique settings which allow us to mimic real-world environments.
- *Other agents impact performance [10, 11].* Other agents, such as AI or other humans, should be able to impact the simulation.
- *Openness [10, 11].* The platform should be cross-platform and portable.

Our ongoing research does not necessitate training AI in *Minecraft*. However, the above guidelines summarize why we believe the game is an optimal platform on which we can test our algorithms and simulate our experiments. The third point in the above list, touching upon how other agents impact performance, is discussed again in various sections of this paper.

Research into simulation, space division and construction of point clouds, graphs and pathways is limited, especially in the realm of search and rescue. Thus, we explore a more generalized approach to 3D pathfinding which has a lower time complexity than more conventional approaches to navigating Euclidean space. In that regard, the authors in [12] explore pathfinding using 3D voxel space in the digital game *Warframe*.

In [12], they propose utilizing an octree, comprised of voxels or "octants", to split the 3D world of *Warframe* into a low time complexity, searchable space which allows for more efficient pathfinding. Every octant is the centre point of a corresponding voxel. The program explores all 26 nearby voxels to find the next best space to move to. The authors determine the next available voxels via the following constraints: are there obstacles in the closest voxel? If so, the voxel is left out of the potential path as the agent cannot occupy the same space as another object. Is the voxel out in the "open", away from cover? It is undesirable for an agent to be out in the open, as it leaves it vulnerable to enemy attacks.

In [13], the authors implement an approach to processing point clouds in Euclidean space, much like what we are attempting to accomplish in our own research; in this case, they wish to navigate through buildings and other structures. The researchers observe that sorting and naively navigating through unordered point clouds can have needlessly high time and space complexities. They propose using an octree to circumvent these issues, allowing new buildings to be mapped internally with a lower demand for computation time. It should be noted that the point clouds these researchers are using are already constructed before data processing; our research differs in that, not only are we constructing the point clouds as a user navigates through the worldspace, but points are not pre-processed.

The authors in [13] use these point clouds to determine the location of obstacles and throughways, such as furniture or doors, respectively. They concern themselves more with identifying and labelling specific objects and terrains, such as stairs, whereas we are more concerned with efficient navigation. More precisely, our focus is on whether or not the firefighter's elevation has changed, as well as the most efficient egress path through a given point cloud and resulting graph.

Further on in this paper, we implement an A* algorithm for searching the resulting graph with appropriate edges drawn between nodes. Two papers which are referenced in this publication for the implementation of A* are: *Algorithms and Theory of Computation Handbook* [14] and *Artificial Intelligence: A Modern Approach* [15]. Both publications act as guidelines for the implementation of not only A*, but other algorithms which are discussed in the *Future Work* section. The latter also provides guidance for both the time and space complexities of A*, which are necessary to explore given the problem domain.

3 Methodologies

The next sections are divided and ordered in a way mimicking the timeline of this project. First, methods for modifying *Minecraft* are outlined, as this is a necessary step for producing meaningful datasets and egress settings for our experiments. Next, we explore dividing the worldspace of *Minecraft* and ensuing datasets into a searchable graph with edges (at times referred to in this work as "adjacencies"). Splitting the worldspace and creating a graph in a timely manner emulates the urgency necessary in the real world environment for firefighters. Pathfinding through the graph follows the graph creation, as is the logical progression of events; again, efficiency in time complexity is discussed since urgency is one of the most important factors in creating the egress path for a first responder. Navigating through the *Minecraft* worldspace follows, alongside using the ARAIG visualization tool. Creating directions and output for the suit, while important, needs to be done with the intention of making them intuitive for first time users. Thus, a short survey is conducted with a small group of users to get initial

feedback on the simulated system.

4 Modding *Minecraft*

Minecraft follows a server-client network model. This means there are two avenues for modding the game: server-side or client-side [16]. The server-side handles logic, game state and updates from clients. The client-side handles rendering the game state and sending updates. This separation of concerns is important. If we wish to illustrate this concept in *Minecraft*, a creature, its location and where it moves is handled by the server. The information about said creature is sent to the client where it is rendered. The client can also send updates, for example, when the player wishes to perform an action such as jumping or hitting a block. In this respect, as it can only modify what it has access to, the client cannot control where the aforementioned creature is and the server cannot control how the aforementioned creature appears.

Server-side modding tends to be relatively straightforward, since the network interface is well understood and documented. There are numerous "community sourced" implementations and application programming interfaces (APIs) which extend the *Mincecraft* server. The API we use for this mod is Spigot [17], which provides a way to run code on events, such as player movement. It also enables programs to react to said events, namely cancelling the movement, recording coordinates or even running actions, such as smiting the player with a bolt of lightning.

Client-side modding is often more complex than its server-side counterpart. To extend the game, client-side mods have to "hook" directly into the "vanilla" client using a variety of complex methods. An example is runtime "bytecode" manipulation, where the Java runtime environment is used to modify compiled code while it is running [9]. *Minecraft* client code is often complex in nature as it deals largely with in-game rendering. In addition, the code is obfuscated, making it difficult to read and understand.

There are two big projects that make modding *Minecraft* easier: Forge [18] and Fabric [8]. Forge is more established and has a larger scope of supported modding functionalities. However, the consequence of this is that the framework takes longer to update and is less light-weight. Fabric, in contrast, is newer and lighter, using modern techniques and providing more low level control. In this project, we opt to use Fabric, as we prefer to keep the mod lightweight, allowing for a more agile approach to mod development.

Originally, we attempted to create a server-side mod. Our goal was determining if location data and its visualization is viable in *Minecraft*. Despite our success in proving its viability, the limitations of server-side modding in regards to visualization quickly made themselves known. As previously mentioned, the server has limited control as to what the client "sees". While the server can spawn creatures or create particles for visualization, we require more customization. Given these circumstances and our requirements, the client-side model is best suited for our mod. This allows us to use the same code to render our custom visualizations that the client uses to render the game.

Our *Minecraft* mod visualizes a graph data structure by drawing the edges as lines and the nodes as numbers in the *Minecraft* worldspace. This gives the user the ability to visualize the recorded coordinate points and the connections created between them. Coordinate points closer to the real world are now easier to collect; there is no need for random numbers or predetermined coordinates.

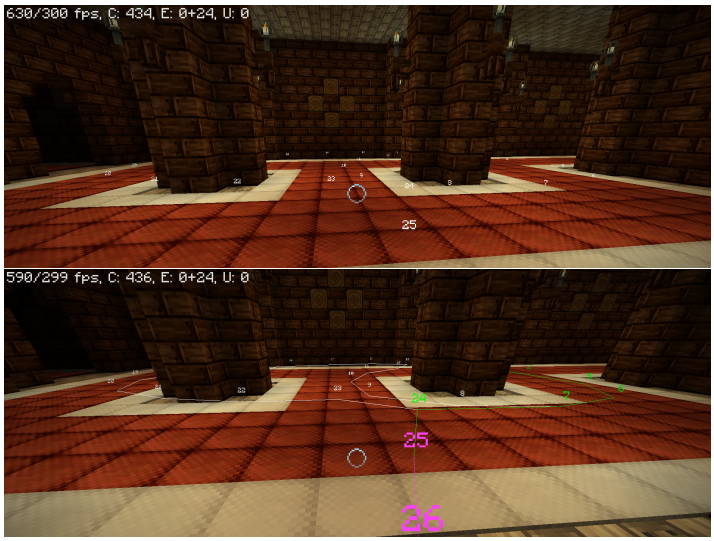


Figure 1: An example of how the *Minecraft* mod visualization tool renders the player's pathway before and after the pathfinding algorithm is run.

As stated above, *Minecraft* is a Java based game. However, in order to seamlessly interface with the ARAIG simulation software, the pathfinding program is in C. Thus, the *Minecraft* mod and the pathfinding algorithms must be split into separate programs, with communication being done over a network socket using a simple protocol. This protocol can be described as a request-response byte encoded message protocol. The first byte is the message type and the rest is the message body in the request. In contrast, the reply is comprised of only a message body. For example, to get the current location from the mod, the pathfinding program sends a *GET_LOCATION* byte identifier. The server then responds with 4 big-endian byte encoded floating point numbers containing to the x, y, z, and yaw components respectfully.

The control flow of the mod is as follows:

1. **Capture location data:** The mod has two commands, */start* and */stop*, the user can enter in the "chat bar" to control the recording of coordinate points. This emulates a GPS device by providing the coordinates of a wearer to an external machine. The location data is stored as a list of coordinate points. While the user is recording his path, a numbered node is placed in the worldspace and recorded as a set of coordinates.

2. **Transfer captured location data:** The pathfinding program then requests all of the recorded coordinate points from the mod, allowing processing to begin.

3. **Transfer live location data:** Additionally, the pathfinding program queries the user's live location data from the client. This allows for live tracking of the user along the path so that egress instructions are updated accordingly in real time.

4. **Send draw updates:** Finally, in response to the captured and live location data, the pathfinding program sends draw

commands to the client. The commands include: drawing or deleting a line between points, changing the colour of a given line, and changing the colour of the points.

While our initial belief of having the data collection and visualization as separate from the pathfinding program may result in unnecessary complexity, the resulting mod is, in fact, the opposite. Using two separate programs for data processing and data visualization allows the code to be more organized and decoupled. One of the benefits for this is the pathfinding program could be restarted and debugged without needing to restart the *Minecraft* client. Since the two are decoupled, both can be developed at different paces.

5 Creating the Graph

Two approaches are outlined here for graph creation, which is necessary for finding an egress path given points in 3D space: a naive approach which has a time complexity of $O(n^2)$ and an octree approach which runs in $O(nlogn)$ time.

5.1 Naive Implementation

Algorithm 1 is the initial approach for creating the shortest egress pathway possible given the user's nodes.

Algorithm 1: *add_node*(*graph*, *node*) Adds a node to the graph. Then, adds it to the adjacency arrays of any existing nodes within a radius *R*.

```
Input: A graph with all previous nodes already inserted
       and connected: graph, a newly created node to
       insert into the graph: node
foreach existing_node ∈ graph.nodes do
    if distance(existing_node, node) ≤ R or
     existing_node == graph.nodes[−1] then
        /* Appends the node to the adjacency
           array of a given node in the graph
           */
        existing_node.adjacencies ← node
end
graph.nodes ← node
```

When a new node is created, the *add_node* function is called to add it to the graph. The node is compared to every existing node already in the graph; if the node is within *R* radius of the node it is being checked against, it is added to the second node's adjacency list. This approach is simple and intuitive, yet inefficient. As mentioned previously, its time complexity is $O(n^2)$. The algorithm works as a proof of concept but it quickly becomes evident that with sufficiently large n, a more efficient approach is necessary. See Figure 2 for a visualization of this function.

5.2 Octree Implementation

An octree object has three member variables associated with it. First, the *bounds* variable, which contains the boundaries for the 3-Dimensional (3D) box that the octree resides in; these are stored as a set of coordinates. Next, it contains a *children* array. This array contains either zero or eight octants. These octants are the result of splitting the *bounds* object into eight equally sized boxes.

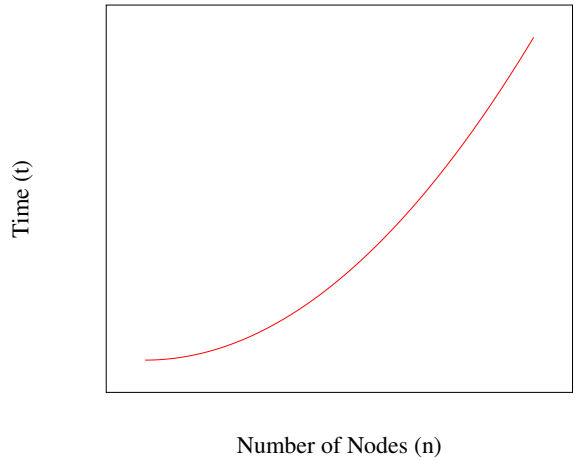


Figure 2: Plotting number of nodes n against time t where $t = n^2$.

Algorithm 2: *make_octree*(*nodes*, *bounds*) Creates an octree given a set of points in 3D space.

```
Input: An array of nodes: nodes, the current boundaries for
       the octree: bounds
Output: An octree containing either eight octree "children"
        (octants) or N ≥ leaf nodes
initialize octree
initialize boundaries
if !nodes then
    octree.children ← NULL
else
    if length(nodes) ≤ N then
        octree.nodes ← nodes
    else
        /* Splitting the space into octants  */
        boundaries ← split(bounds)
        initialize octree.children
        for i = 0; i ≤ 8; i + + do
            initialize next_nodes
            foreach node ∈ nodes do
                if node ∈ boundaries[i] then
                    next_nodes ← node
            end
            octree.children[i] ←
              make_octree(next_nodes, boundaries[i])
        end
    end
end
return octree
```

Finally, an octree object has a *nodes* array; it remains empty unless the octree object is a leaf in the greater data structure. It should be noted that, unlike the naive approach above, this algorithm is run after the subject has ended the path recording. The pseudo-code for the octree creation is illustrated in Algorithm 2.

The initial octree object is created with its bounding box characterized by the minimum and maximum (x, y, z) coordinates of the whole graph. The function *make_octree* is then called on an array

of every node and the aforementioned boundaries. Then, one of three things must happen: first, if the array of nodes is empty, the octree's *children* array is set to $NULL$ to indicate that it is empty. Second, if the array of nodes exists, and contains less than or equal to the amount of allowed nodes in a given octant space, the octree's *nodes* array is populated and the octree object becomes a "leaf".

Finally, if neither of the previous two conditions are true, the function recursively calls itself. A new set of octants is created by splitting the boundaries into eight identically sized cubes. Next, for each new octant, each node is compared to its bounds and, if it is contained within the octant's boundaries, it is placed in a new array. Once the array is populated with the appropriate nodes, the *make_octree* function is recursively called on it and its respective boundaries. The resulting octree is added to the *children* array for the current octree.

The construction of this octree creates an efficient, though spatially complex, search space which allows for the appropriate edges in the graph to be created. The pseudo-code for edge creation, as a series of octree searches, is in Algorithm 3.

Algorithm 3: $find_adjacencies(octree, node)$ Populates a node's adjacency array with nodes within a radius R.

```
Input: An octree object: octree, the current node that we
       wish to find the adjacencies for: node
if octree.nodes then
    foreach oct_node ∈ octree.nodes do
        if node ∉ oct_node.adjacencies and
          distance_between(oct_node, node) < R then
            node.adjacencies ← oct_node
    end
else
    for i = 0; i < 8; i++ do
        if node ∈ octree.children[i].bounds then
            find_adjacencies(octree.children[i], node)
    end
end
```

As Algorithm 3 demonstrates, $find_adjacencies$ takes a fully formed octree and a node as arguments. It is called on every node in the graph once. The function checks if the octree object is a leaf, and, if it is, the selected node is compared to each node in the octree's *nodes* array. If any nodes are within the given distance R, they are added to the adjacency array for the node. Conversely, if the octree object is not a leaf in the octree, the bounds for each member of its *children* array are checked against the coordinates of the node. If the node is within the given child's bounding box, the $find_adjacencies$ function is called on the child and the node. This is done recursively until all of the appropriate octree branches for a given node are explored.

It should be noted that the *average* case runtime for creating the octree and recreating its path is $O(nlogn)$, where n is the total number of nodes in the system. The worst-case runtime for this octree approach is the same as the above *naive implementation*, which is $O(n^2)$. The worst case for this algorithm occurs if and when a node is compared to every leaf in the octree and thus, every other node in the system.

The bottleneck for efficiency in this case is not the creation of the octree, which is in fact linear time, nor is it any given single search of the octree, which is also $O(logn)$. Rather, the bottleneck is that the search must be performed for each node in the graph. Thus, the time complexity for this algorithm is $T = n + nlogn$, which is the sum of both the runtime of the octree construction and the creation of each node's adjacency list. Idiomatically, the runtime is $O(nlogn)$. See Figure 3 for a visualization of this function.

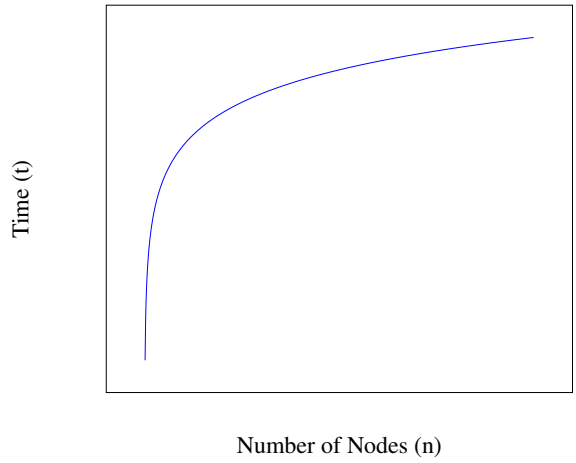


Figure 3: Plotting number of nodes n against time t where $t = nlog(n)$.

6 Egress Path Creation

This section assumes one of the previous two graph creation algorithms was run and there now exists a graph with appropriate nodes and their respective adjacencies. Algorithm 4 creates a path back to the first node of the graph for the user. See Figure 4 for an example overview of a reconstructed path.

Algorithm 4: $create_pathway(graph)$ Creates a stack containing the nodes comprising the most efficient path back to the beginning of the graph.

```
Input: A graph with appropriate adjacencies already
       created: graph
initialize stack
initialize node ← graph.nodes[0]
while node ≠ graph.nodes[−1] do
    stack.push(node)
    node ← node.adjacencies[−1]
end
/* Gives the directions to the user in order
   */
initialize next_node
while stack do
    next_node ← stack.pop
end
```

6.1 Naive Search

Algorithm 4 navigates through the graph by "jumping" to the last neighbour in the current node's edges. Both of the previous algorithms place the latest recorded neighbour at the end of the adjacency array. The naive algorithm accomplishes this by appending new

nodes to the graph as they are created and updates the adjacency arrays accordingly by comparing each existing node to the new node. The octree algorithm emulates this behaviour by comparing a node's ordering ID to the neighbour's ID. If the node's ID is less than the incoming neighbour's ID, the node is appended to the adjacency array. Thus, a stack is the only thing necessary when recreating the shortest path with which the user can navigate back to the beginning of the graph. The time complexity for this approach is linear, $O(m)$, where m is the number of nodes in the egress path; the worst case runtime is $O(n)$, where n is the number of nodes in the graph.

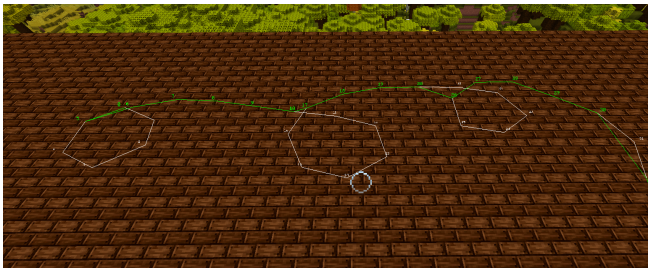


Figure 4: An aerial view of a reconstructed path.

Algorithm 5: A* Search

Input: A graph with all adjacencies drawn: *graph*
Output: A path containing the nodes from the user to the closest goal node: *path*

```
initialize priority_queue
priority_queue.enqueue(start_node, 0)
while !priority_queue.is_empty do
    u ← priority_queue.deqeueue
    if u == graph.end then
        final_node ← u
        break
    else
        foreach adjacency ∈ u.adjacencies do
            if adjacency.g + distance(adjacency, u) <
              adjacency.g then
                adjacency.previous ← u
                adjacency.g ←
                  u.g + distance(adjacency, u))
                f ← adjacency.g + adjacency.h
                priority_queue.enqueue(adjacency, f)
            if adjacency.visited then
                adjacency.reExpansions + +
            else
                adjacency.visited ← true
            end
        end
    end
end
return path(start_node, final_node)
```

6.2 A* Search

The issue with the previous implementation is that it does not take into account a future feature for this system: multiple users. Even worse, there are some instances of path creation in which jumping to the highest neighbouring node will not actually create the most efficient pathway back for a single user. An edge case which would cause this undesirable behaviour, for example, is if the user continuously navigates in "loops" or interconnected circles. Algorithm 5 is a basic implementation of A*, which replaces the naive approach above. While the naive approach is linear in its time complexity, the average time complexity for A* is $O(b^d)$, where b is the branching factor and and d is the depth of the solution [15].

The heuristic function utilized by Algorithm 5 is the Euclidean distance value from any given node to the final node. This allows for prioritization of nodes, leading the user in a straighter line towards the goal. This heuristic, given only one goal node, is admissible [14], meaning it does not overestimate the cost of a given node to the goal node [14].

7 Giving Directions in *Minecraft*

This section assumes the egress path has already been constructed. As the *Modding Minecraft* section states, the game was created and released in the early 2010 [19]. The programmers behind the original Java-based game made some unusual implementation decisions, particularly in regards to its coordinate system. In essence, it is a right-handed system with unconventional axes. The three issues which we circumvent later in this paper, in order to provide accurate navigation, are:

1. The y-axis is the measure of how high or low the player is relative to the ground, as opposed to the conventional z-axis for this task;
2. The angles between points in the XZ-plane are given clockwise, instead of the conventional counter-clockwise that is generally utilized for trigonometry;
3. The positive Z axis, which is South, is 0 radians in *Minecraft*. Consequently, North, which is negative Z, is π radians.

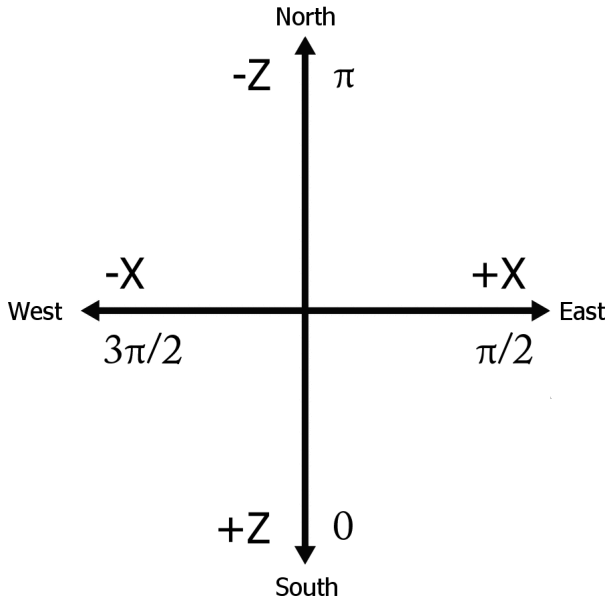


Figure 5: A diagram of the *Minecraft* coordinate system.

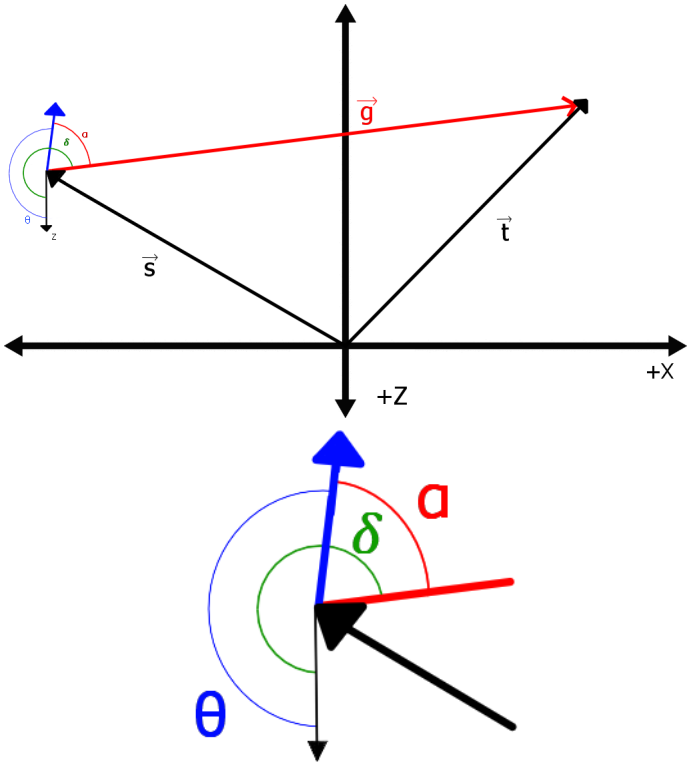


Figure 6: An example of the resulting system of vectors when a player's yaw is facing a different direction than the next desired node. θ and the blue arrow are the player's yaw. $\vec{t}$ is the node vector, $\vec{s}$ is the player's position vector.

Once the shortest path has been created for the player, we need to give him directions to the next immediate node in that path. To do this, we calculate the angle of the next node relative to the player's yaw. We shall call this angle α. First, we take the vector $\vec{t}$, which is the vector pointing to the desired node, and the player's position vector, $\vec{s}$, and calculate $\vec{g}$, which we will refer to as our *direction vector*:

$$\vec{g} = \vec{t} - \vec{s}$$

Once we have our direction vector, we calculate β, which is the angle of $\vec{g}$ relative to the native coordinates to *Minecraft*. For this, we use 2-argument arctangent:

$$\beta = atan2(-g_z, g_x) + \pi$$

Once β is calculated, we have to adjust the angle so that it is relative to the flipped coordinate system that we are now working in. We will call this angle δ:

$$\delta = 3\pi/2 - \beta$$

The final step to find the next desired direction is to calculate the difference between the resulting δ and θ, where θ is the player's yaw. This is α:

$$\alpha = \delta - \theta$$

For the sake of convenience, in order to ensure that the resulting angle is easy to use when giving directions from the suit in the next node, we may add 2π to normalize the value of α. Please refer to Figure 6 for a visualization of an example in this system. Thus, the algorithm used to calculate the next desired directional instruction, relative to the player, is outlined in Algorithm 6.

Algorithm 6: Returns the angle of the node relative to the player's position and yaw.

Input: A player's coordinates and yaw: *player*, the desired node: *node*

$g_x \leftarrow node.x - player.x$
$g_z \leftarrow node.z - player.z$
$\beta \leftarrow atan(-g_z, g_x) + \pi$
$\delta \leftarrow (3\pi)/2 - \beta$
$\alpha \leftarrow \delta - player.yaw$
if $\alpha < 0$ **then**
  $\alpha \leftarrow \alpha + 2\pi$
return α

A public repository including these algorithms may be found in [20]. The functionalities which are specific to the ARAIG suit, as discussed below, have been removed for the sake of licensing.

8 Integration with ARAIG

The "As Real As It Gets" (ARAIG) suit, as outlined in IFTech's specifications in [21], has numerous vibratory and stimulus sensors. In order to provide a distinct set of visual cues on the simulated suit, we utilize the vibratory sensors. This way, the player can quickly translate the visual instructions from the simulated suit to following directions inside the *Minecraft* environment. To ensure the system was quick to learn for new users, the program only outputs four directions (which are given in relation to α as calculated in the previous section):

1. **Forward:** $\pi - 1/2 < \alpha < \pi + 1/2$. The user's abdomen and pectorals are stimulated, indicating to them that they should move forward.

2. **Left:** $\pi/2 - 1/2 < \alpha < \pi - 1/2$. The user's left shoulder is stimulated, indicating to them that they should turn left.

3. **Right:** $\pi + 1/2 < \alpha < 3\pi/2 + 1/2$. The user's right shoulder is stimulated, indicating to them that they need to turn right.

4. **Turn around:** If α is not within the previous three ranges, the user is not facing the correct direction. Thus, the user's back is stimulated, prompting them to turn around.

Once the program is running, the user is given a set of initial instructions. The directions relayed to the user are updated every 500 milliseconds given the user's yaw and location. Taking the conditions outlined in Figure 6 as an example, the user's suit output would appear on the screen as shown in Figure 1.



Figure 7: An example of how the ARAIG suit simulation software appears given the conditions in Figure 6.

The ARAIG simulation software integration has been omitted in the supplementary GitHub repository. If readers wish to use this software, they are encouraged to reach out to IFTech at [22]. Once readers receive the appropriate permissions to use the ARAIG visualization software and SDK, they are free to contact Cassandra Laffan or Robert Kozin for access to the full version of this mod and its functionalities.

9 User Testing

A small series of tests were designed to examine whether or not this system is intuitive and quick to learn for new users. The sample size for this study is limited by the non-disclosure agreement (NDA) which protects the ARAIG visualization tool. Consequently, users tested in this study are only those with access to the researchers' machines. As a result, we survey four users of varying backgrounds, all of whom have access to one of the computers with the visualization software available.



Figure 8: Screenshots of the burn house as constructed in *Minecraft*.

The user tests take place in a "burn house", which is a structure built to the specifications as outlined in [23]. Burn houses are standardized buildings in which firefighters may practice navigating structures and fighting fires in a physically simulated environment here in North America. This is a logical testing environment as the system is being designed with first responders in mind. The main goal of this "pilot study" is not necessarily to evaluate how quickly a user may exit a building given different circumstances, but to observe how the average user interacts with the system. We also gather feedback on possible system improvements in hopes of making it more intuitive to new users.

There are four categories of navigation tests, all of which have the same set up and goal: the user is tasked with navigating to the top of the tower, then retracing their pathway down. The navigation upward is not timed as it generally took 60 seconds±1 second; what was timed was the user navigating back to where they started. The users are told it is acceptable to both stray from the path if they were lost or quit for the same reason. The categories for testing are as follows:

1. **Control Run:** The control run times the users navigating to their starting points at the bottom of the tower with high visibility and no navigational assistance.
2. **Low Visibility:** This run is much like the control run, with no navigational assistance. However, distance of visibility for the users is greatly decreased, as per Figure 8.
3. **High Visibility with Path Recreation:** This category of testing allows the users their full field of vision. Their goal of retracing their path is assisted with output on the simulated ARAIG suit on a neighbouring screen. The users are informed that they could opt not to acknowledge the suit's output if they find it to be confusing or a hindrance to their task.
4. **Low Visibility with Path Recreation:** This set of tests has the users navigate up to the top of the tower and back down with low visibility. They are given the output of the simulated ARAIG suit on a neighbouring monitor to assist them in this task. Much like in the previous category, they are informed that if they felt the suit is acting as a hindrance to their task, they can opt to ignore its output.



Figure 9: An example of a low-visibility environment created by our mod.

Users are also asked to give any and all feedback they believe is pertinent to the experiment.

10 Results

Users were given time to practice controlling the player character in the game before running the tests. The results for each run are

shown in Figure 10. Generally, the users were more efficient when not following the directions given by the suit in high visibility situations. However, whether or not the use of the suit made it easier for the user to find their way back to the start in low vision environments seemed entirely dependent on the user's experience with *Minecraft* in the past and the path they took. This will be further discussed below.

	User A	User B	User C	User D
Control Run with High Visibility	0:51:12	0:23:03	0:47:29	0:19:03
	0:53:2	0:55:07	0:22:16	0:49:11
	0:52:5	0:45:13	59:33:00	0:39:33
Control Run With Low Visibility	1:35:14	1:02:29	0:17:35	0:26:27
	1:30:24	0:39:1	1:12:07	1:10:23
	1:25:29	59:18:00	1:36:09	1:35:16
High Visibility with Path Recreation	0:61:3	54:26:00	1:17:42	0:22:16
	0:56:4	1:46:48	2:57:00	1:05:12
	0:55:6	0:21:17	2:17:59	0:58:45
Low Visibility with Path Recreation	1:15:03	1:42:14	1:30:04	0:46:06
	1:20:32	1:26:45	1:51:01	1:23:36
	1:12:12	1:40:28	36:09:00	1:28:29

Figure 10: The results of our short pilot study. Measurements are given in Minutes:Seconds:Milliseconds.

User feedback was as follows:

- Users A, B and D all remarked that having the suit in the low vision environment made navigating somewhat easier if they did not initially take erratic paths.
- Users B, C and D all said that dividing their attention between *Minecraft* and the ARAIG simulation software made navigating efficiently very difficult.
- Users B and C insisted that completing the task would be much easier while wearing the suit.
- User B suggested that instead of stimulating the user on the back to prompt him to turn around, to instead have it indicate that the user should go forward. This would emulate a "pushing" motion.

When we asked users if they would find the system useful when wearing the physical ARAIG suit, all of them responded that yes, it would be more helpful.

11 Discussion

On average, the time taken to navigate the high visibility portion is shorter unassisted versus navigating with the assistance of the simulated ARAIG suit. This is supported by the feedback from most users: splitting their attention between two monitors may be distracting and much more difficult than simply guessing their return path. The shortest pathways, as reflected in Table 1, are either the user taking advantage of the fact there were a few optimal routes to the top of the building from the ground, or them falling down multiple flights of stairs.

Users A and D both have previous experience playing *Minecraft*. They found the ARAIG suit's contributions to their navigation back to their starting point to be beneficial. Users B and C, on the other hand, expressed that the simulated suit detracted from their ability to navigate, as they were already focusing heavily on how to navigate in the low light environment. Shorter pathways up, which were generally just a simple race up the stairs, are reflected in the low visibility runs.

12 Conclusion

In this project, we continue the work we first presented in [2], where we propose a system for assisting first responders in navigating out of low visibility environments utilizing the ARAIG haptic suit. In order to circumvent various obstacles due to the pandemic and supply chain interruptions, we opt to simulate the ARAIG functionality and path recreation in the digital game *Minecraft*. To do so, a mod for the game integrating the ARAIG visualization software is written. This mod tracks a user's movement through the *Minecraft* worldspace. Once the player's points in space are recorded, two different implementations of creating a graph from these points are proposed: the naive approach, which directly compares every node to every other node, and an octree approach, which divides the worldspace into octants, allowing for efficient edge creation. Then, two search algorithms are implemented and compared for finding the most efficient egress path in the resulting graph. The first implementation is another naive approach, which simply jumps from a node to the last recorded node in its edge array. The second implementation is a basic A* algorithm, which, while having a higher time complexity in a worst-case scenario, does not fail in special circumstances.

Finally, we have four users test our software and give us constructive feedback based upon their experiences playing the game in conjunction with the visualization tool for the ARAIG suit. Users generally agree dividing their attention between two programs is difficult and the task would be much easier to complete if wearing the physical ARAIG suit. This feedback is useful for when our research evolves to include firefighters, as we do not want firefighters to feel as if they are disconnected from, or do not understand, the physical input from the suit.

12.1 Future Work

Succeeding this leg of our research are various steps we plan to implement. First and foremost, as we note above, we are looking to implement multi-user functionalities. In doing so, further usage of A* is necessary and the naive implementation for finding the egress path simply will not work. There are numerous variants and alternatives to A* [14], including iterative deepening A* (IDA*)

and recursive best first search (RBFS). The next immediate step is thus finishing the *Minecraft* mod, implementing various search algorithms and comparing them on large datasets.

Before we begin testing in real-world environments, we intend on bridging the gap between these two concepts. As our *Results* section mentions, users lamented having to split their attention between two screens while navigating with the simulated ARAIG suit. Further testing in the *Minecraft* environment includes having a user wear the physical suit while navigating the game world. This would allow the user to focus his full visual attention on *Minecraft* while receiving instructions from the ARAIG garment.

Other steps in this research include integrating the algorithms with software which interfaces with the physical world, such as Google's AR Core API [24]. Using our algorithms in more controlled, less noisy real-world environments will allow us to refine these algorithms before we begin integrating them with noisier data such as LiDAR, sonar or a 3D camera data. Finally, we can eventually begin work on integrating these algorithms fully with the physical ARAIG suit and supplementary sensors, as first introduced in [2]. We will then test this technology in a physical burn house as presented in [23].

Conflict of Interest The authors declare no conflicts of interest.

Acknowledgements The researchers acknowledge the funding provided by the Toronto Metropolitan University in conjunction with the Natural Sciences and Engineering Research Council (NSERC).

A special thanks to Jinnah S. Ali-Clarke and Aaron Gill-Braun for providing invaluable support and insight for our research. Another special thanks to Erin MacLellan for her invaluable editing contributions to this paper.

References

[1] S. R., "Covid-19's impact felt by researchers," 2021.

[2] C. F. Laffan, J. E. Coleshill, B. Stanfield, M. Stanfield, A. Ferworn, "Using the ARAIG haptic suit to assist in navigating firefighters out of hazardous environments," 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2020, doi:10.1109/iemcon51383.2020.9284922.

[3] C. F. Laffan, R. V. Kozin, J. E. Coleshill, A. Ferworn, B. Stanfield, M. Stanfield, "ARAIG And Minecraft: A COVID-19 Workaround," in 2021 IEEE Symposium on Computers and Communications (ISCC), 1–7, 2021, doi:10.1109/ISCC53001.2021.9631428.

[4] W. J. Ripple, C. Wolf, T. M. Newsome, P. Barnard, W. R. Moomaw, "World Scientists' Warning of a Climate Emergency," BioScience, **70**(1), 8–12, 2019, doi:10.1093/biosci/biz088.

[5] J. P. Tasker, "Elizabeth May says climate change, extreme events like Fort McMurray fire linked — CBC News," 2016.

[6] S. Larson, "Massive fire north of Prince Albert, Sask., is threatening farms and acreages — CBC News," 2021.

[7] W. Mora, Preventing firefighter disorientation: Enclosed structure tactics for the Fire Service, PennWell Corporation, Fire engineering Books & Videos, 2016.

[8] "Fabric," 2015.

[9] "Fabric Mixin Framework," 2015.

[10] M. Johnson, K. Hofmann, T. Hutton, D. Bignell, "The Malmo Platform for Artificial Intelligence Experimentation," in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, 4246–4247, AAAI Press, 2016, doi:10.5555/3061053.3061259.

[11] S. Adams, I. Arel, J. Bach, R. Coop, R. Furlan, B. Goertzel, J. S. Hall, A. Samsonovich, M. Scheutz, M. Schlesinger, S. C. Shapiro, J. Sowa, "Mapping the Landscape of Human-Level Artificial General Intelligence," AI Magazine, **33**(1), 25–42, 2012, doi:10.1609/aimag.v33i1.2322.

[12] D. Brewer, N. R. Sturtevant, "Benchmarks for Pathfinding in 3D Voxel Space," in SOCS, 2018.

[13] F. W. Fichtner, A. A. Diakité, S. Zlatanova, R. Voûte, "Semantic enrichment of octree structured point clouds for multi-story 3d pathfinding," Transactions in GIS, **22**(1), 233–248, 2018, doi:10.1111/tgis.12308.

[14] M. J. Atallah, M. Blanton, 22.4, CRC Press, 2010.

[15] S. J. Russell, P. Norvig, Artificial Intelligence: A modern approach, Pearson, 4th edition, 2022.

[16] "Protocol FAQ, URL: https://wiki.vg/Protocol FAQ." .

[17] S. Team, "Spigot, URL: https://www.spigotmc.org/." .

[18] "MinecraftForge Documentation, URL: https://mcforge.readthedocs.io/en/latest/." .

[19] Notch, "Minecraft 0.0.11A for public consumption," 2009 URL: https://web.archive.org/web/20150716115516/http://notch.tumblr.com/post/109000107/minecraft-0-0-11a-for-public-consumption.

[20] C. F. Laffan, R. Kozin, "Octree Path Finding Algorithm," 2021. URL: https://github.com/cassLaffan/Minecraft Pathfinding.

[21] "ARAIG As Real As It Gets, URL: https://iftech-technologies.com/downloadthe- sdk/." .

[22] B. Stanfield, M. Stanfield, "Download the ARAIG SDK, URL: https : / / www . firefacilities.com/fire-training-towers/tower-models/thecaptain/." .

[23] "Fire Department Training Building - Multiple Fire Fighter Trainees," 2020. URL: https : / / www . firefacilities.com/fire-training-towers/tower-models/thecaptain/.

[24] "Build new augmented reality experiences that seamlessly blend the digital and physical worlds, URL: https://developers.google.com/ar." .

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 49-57 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Profiling Attack on WiFi-based IoT Devices using an Eavesdropping of an Encrypted Data Frames

Ibrahim Alwhbi Alharbi[*,1], Ali Jaber Almalki[2], Mnassar Alyami[1], Cliff Zou[1], Yan Solihin[1]

[1]*Department of Computer Science, University of Central Florida, Orlando, 32816, USA*

[2]*Department of Computer Science and Information Technology, University of Bisha, Bisha, 67714, Saudi Arabia*

ARTICLE INFO

Article history:
Received: 19 August, 2022
Accepted: 15 October, 2022
Online: 13 November, 2022

Keywords:
Internet of Things
Privacy Attack
Eavesdropping

ABSTRACT

The rapid advancement of the Internet of Things (IoT) is distinguished by heterogeneous technologies that provide cutting-edge services across a range of application domains. However, by eavesdropping on encrypted WiFi network traflc, attackers can infer private information such as the types and working status of IoT devices in a business or residential home. Moreover, since attackers do not need to join a WiFi network, such a privacy attack is very easy for attackers to conduct while at the same time invisible and leaving no trace to the network owner. In this paper, we extend our preliminary work originally presented at the CCNC'22 conference by using a new set of time series monitored WiFi data frames with extended machine learning algorithms. We instrument a testbed of 10 IoT devices and conduct a detailed evaluation using multiple machine learning techniques for fingerprinting, achieving high accuracy up to 95% in identifying what IoT devices exist and their working status. Compared with our previous work in , the new approach could achieve IoT device profiling much quicker while maintaining the same level of classification accuracy. Moreover, the experimental results show that outside intruders can significantly harm the IoT devices without joining a WiFi network and can launch the attack within a minimum time without leaving any detectable footprints.

1. Introduction

This paper is an extended version of the paper published in Alyami, Mnassar, Ibrahim Alharbi, Cliff Zou, Yan Solihin, and Karl Ackerman. "WiFi-based IoT Devices Profiling Attack based on Eavesdropping of Encrypted WiFi Traffic." In 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC), pp. 385-392. IEEE, 2022 [1].

The emerging smart infrastructures are integrated with the Internet of Things (IoT) devices and their applications to make daily life easier for individuals and improve the public environment [2]. To this end, IEEE 802.11 Wireless network (WiFi) is a significant development that helps connect a wide variety of IoT devices such as smartphones, smart TV, home automation, intelligent vehicles, surveillance cameras, health monitoring, and many more [3]. The increase in applications increases the attention of attackers that find the loopholes and gain maximum knowledge of users' private in- formation. The connected devices, digital systems, and sensors that play a vital role in people's daily life cause a significant threat of privacy leakage of private information [4]. For example, the Mirai malware attack, which triggered distributed denial-of-service (DDoS), generates attacks on WiFi-connected IoT devices and ap- plications [5]. Additionally, worms in smart bulbs gave attackers access to all adjacent IoT lights that were compatible [6]. To this end, the infrastructure and IoT applications must incorporate pri- vacy protection during intelligent network design and development phases. Figure 1 shows an overview of IoT ecosystems and relevant scenarios.

Considering the aforementioned privacy threat, this paper focuses on the attacker's ability to fingerprint the IoT devices in the WiFi network. The existing work mostly assumes the attacker is inside the network where the attacker has to either join the network prior to fingerprinting or wiretap the network link of the WiFi network [7]-[9]. This assumption cannot be satisfied in the real world by most attackers as most WiFi networks have secured password protection, and very hard to impossible for attackers to have physical access to their WiFi routers/access points.

To this end, in this paper, we conduct a brief investigation of whether an outside attacker can identify a network's IoT devices without having a joined in a WiFi network. Compared to the entities in the network, the outsider attacker may face several

*Corresponding Author: Ibrahim Alwhbi Alharbi, ia@knights.ucf.edu

difficulties, i.e., the attack may be prevented from analyzing the plaintext of packet payload due to the WiFi data-link layer encryption. On the other hand, the outside attacker captures the fingerprint of the data packet, which would be noisy (especially when more nodes are present in the network). Moreover, the captured data cannot obtain the IP address or the port information; thus, it is difficult to assume whether the hypothesis is correct or not.

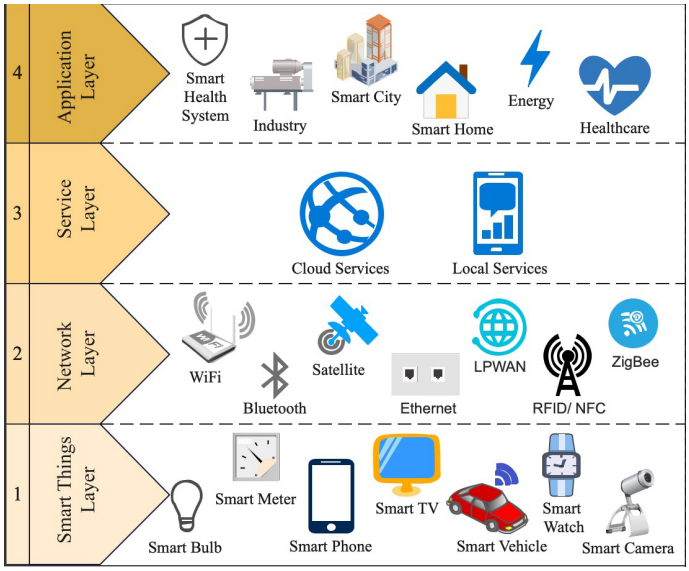


Figure 1: The overview of IoT ecosystem and applications scenarios.

However, if the hypothesis is correct, there could be major threats as follows:

- The attack is possible on all WiFi networks which are in close proximity to the attacker. So it is easy to attack those with weak passwords.
- The attacker only needs to drive or walk a short distance to begin analyzing traffic, where no specific preparations are required beforehand.
- The attack is usually untraceable because it leaves no traces. Thus it is undetectable neither for users nor for forensic investigators.

The hypothesis needs to be investigated in detail based on the above points.

To this end, we demonstrate that the outside attacker cannot only fingerprint the IoT devices, but it is a straightforward process that one can achieve depth information about the network devices. The significant contributions of the proposed research are summarized as follows:

- We conduct a detailed investigation and prove that fingerprinting the IoT devices from the outside of the network eavesdropping is not only feasible, but it's a straightforward process.
- We consider nine real-world IoT devices and capture the out-of-network WiFi traffic in two modes, idle and active, using the sniffing tool capable of single-channel and multi-channel monitoring.
- We explore the time-series data and train a machine-learning algorithm to profile nine real-world IoT devices and present their prediction accuracy. The experimental results prove that the fingerprint can be possible by achieving high accuracy of up to 95%.

The rest of the paper is organized as follows: In section 2, we present a brief literature review that motivates us to conduct this research work. In section 3, we first define the problem statement and then present the threat model and assumptions of the proposed research. Section 4 briefly explains a detailed procedure for capturing WiFi traffic from outside the network. In section 5, we conduct data processing and analysis for profiling attacks based on machine learning. Section 6 shows the experimentation and proves that the hypothesis above is true. Finally, in section 7, we conclude the proposed research and define the future work.

2. Related Work

Thanks to the IEEE 802.11 protocol and the development of Wi- Fi-enabled devices, instant access to the internet is now possible everywhere near a public AP through a Wi-Fi connection [10]. Bil- lions of people's lives have been significantly impacted by this development throughout the world. However, when a massive num- ber of people are involved, there are higher chances of misuse or exploitation [11]. The general public can now be followed and profiled due to security vulnerabilities brought on by the ease of use of the 802.11 protocol suite and the number of open public Wi-Fi hotspots [12].

Since the early days of the internet, device identification has been one of the primary targets for classifying network traffic [13]. Several studies have been conducted on WiFi and ethernet, proving that various information can be extracted from IoT devices using traffic classification, such as the device type and the device's activi- ties [14]. The existing literature predicts that a collection of TCP/IP level packets would allow observation of network activity. Those techniques are used to extract the device's potential information. In [15], the authors claim the possibility of a single attribute signature for a variety of IoT devices by employing a port number. Moreover, the deep learning technique is used to perform device fingerprinting for flow volume features. However, this technique is not suitable for the network which is accessed from the outside adversary because the IP traffic is encapsulated in the upper layer that encrypts all significant network features such as cipher suites, protocol, and port number. To this end, we employ a unique collection of attributes that are straightforward to extract the features even from outside the net- work. In [16], the authors used hardware fingerprinting and extracts clock skew measurement. This major focus in this study is more on the hardware rather than the device-specific classification which is considered in the proposed research. In [7], the authors used traffic analysis of WiFi, Bluetooth, and ZigBee to identify the status of the devices and proposed a defensive strategy based on traffic spoofing. In order to gain the traces of WiFi, the authors use a rogue access point with tcpdump, which means the authors assume that all the adversaries are either a part of the network or the adversary already contains significant knowledge about the network. However, the proposed technique considered that the adversary doesn't have any prior knowledge, and the attack is conducted from the outside of the network. There are many other

studies that have been conducted with a focus on WiFi traffic analytics from the outside of the network where the traffic is analyzed using off-the-shelf monitoring devices [17]. However, to the best of our knowledge, none of the existing studies reports the missing rate using those off-the-shelf sniffers or the lost frames as a result of channel hopping eavesdropping [18].

On the other hand, several studies have been conducted on the defense mechanism for those traffic analysis attacks [19, 20]. How- ever, the primary focus of the existing literature is either on location anonymity or website fingerprinting. Moreover, the techniques such as padding or traffic morphing result in less accuracy of the classifier [21]. Besides, those methods cannot rely on time-based classifi- cation because of their limited capabilities for obfuscating traffic patterns.

In contrast to comparable work mentioned above, the proposed research presents an alleged privacy attack against WiFi-based devices that rely on WiFi traffic monitoring from outside the network. We proceed with a precise and practical proof-of-concept attack on the assumption of a realistic threat scenario. We also discuss a possible defense against those attacks.

3. Problem Statement, Threat Model, and Assumptions

In this section, we first define the problem statement that serves as a strong motivation for the proposed research. Afterward, we present the threat model that needs to be covered by the proposed study. Finally, we discuss the considered assumptions while conducting the proposed research.

3.1. Problem Statement

The TCP/IP paradigm is used for communicating devices on net- works. The IP address at the Network Layer and the MAC address at the Data Link Layer can be used to identify the devices on the network. Spoofing identities have been used to get around these identifying mechanisms and access restricted resources. Using WiFi traffic analysis, an attacker can ”fingerprint” devices to determine private user behavior [22]. For instance, by continuously watching the camera’s bitrate, the attacker could ascertain the movements of objects inside a building [23]. Moreover, the attacker can predict which vulnerabilities are available to exploit depending on the type of IoT devices in the network. However, extensive research has been conducted on fingerprinting the IoT devices using eavesdropping from the inside network [24]. This research performs a detailed investigation and proves that fingerprinting the IoT devices eaves- dropping from outside the network is not only possible but also a straightforward process. The developers of the IoT devices must need to consider some extra security constraints to overcome those privacy threats.

3.2. Threat Model

In the proposed research, we consider that the attacker aims to target the information of IoT devices using a targeted WiFi network. More- over, the attacker is also interested in the number and type of unique IoT devices, e.g., Laptop, Smart TV, Light Bulb, etc. Moreover, the attacker is also interested in getting the mode of those devices, such as idle or active. The attacker intends to gain maximum sensitive information by gathering the devices’ data. For example, the device type may reveal potential vulnerabilities to software/hardware status. The number of devices may reveal the customers in business, the number of employees, or the family size. The type and number both can reveal the status of socioeconomic.

Considering this as a potential threat model, we aim to perform a detailed investigation that fingerprinting the IoT devices eavesdrop- ping from outside the network is a straightforward process. This threat should be considered in the first place.

3.3. Assumptions

In the proposed research investigation, we assume that the attacker continuously observes the network traffic outside the network using the targeted WiFi or access point. We also believe that the attacker is physically in the signal range of the access point, so he can per- form eavesdropping using a sniffing tool and gather the nearby WiFi network traffic. We assume that the attacker can’t join or break the network. To this end, in the proposed research investigation, we prove that fingerprinting the IoT devices from the outside of the network eavesdropping is possible. Moreover, the existing research focuses on the IoT devices operated at 2.4GHz; we consider the same. However, the proposed study can be applied to 5GHz as well.

4. Verification of Collective Movement

In this section, we first present the system architecture, and then we discuss how the attacker captures the network traffic from outside of the network. After that, we present the pre-processing of the captured data.

4.1. System Architecture

In the proposed investigation, we consider the system architecture illustrated in Figure 2 where the attacker uses to access the WiFi network. The system architecture consists of two stages, offline and online. The first stage (offline) is the attacker’s profiling model training and building stage, where an attacker uses his computer and many IoT devices to conduct experiments in order to build the profiling model of each IoT device. On the other hand, the second stage (online) is the attacking stage, where the attacker monitors a WiFi network, trying to identify all IoT devices in the WiFi network based on monitored data and profiling models built in the offline stage. In the first stage (offline), the attacker configures maximum IoT devices which are connected to the nearby WiFi gateway. The attacker accesses the network traffic using a sniffing tool, where the traffic data would be labeled as the device name using the MAC address. The collected data is then pre-processed, where we re- moved the noise (e.g., network traffic gathered from nearby WiFi networks, data link layer broadcast frames, WiFi protocol beacon frames), and dumped the valuable features into a CSV file for apply- ing the machine learning techniques. In particular, we apply several machine learning algorithms and achieve accuracy up to 95% for device identification.

In the second stage (online), the attacker applies a sniffing tool and targets the victim’s access point for a short period of, for example, 30 seconds, and stores the traces for pre-processing. To this point, we never require prior knowledge of the IoT devices for pre-processing, which we will explain in detail later in the follow- ing subsection. Precisely, we use standard and statistical filtering techniques to eliminate the noise from the data frames which do

not represent the patterns of data. After that, we use Python scripting to extract the features from pre-processed data. Finally, we were able to predict the type of devices and their activity.

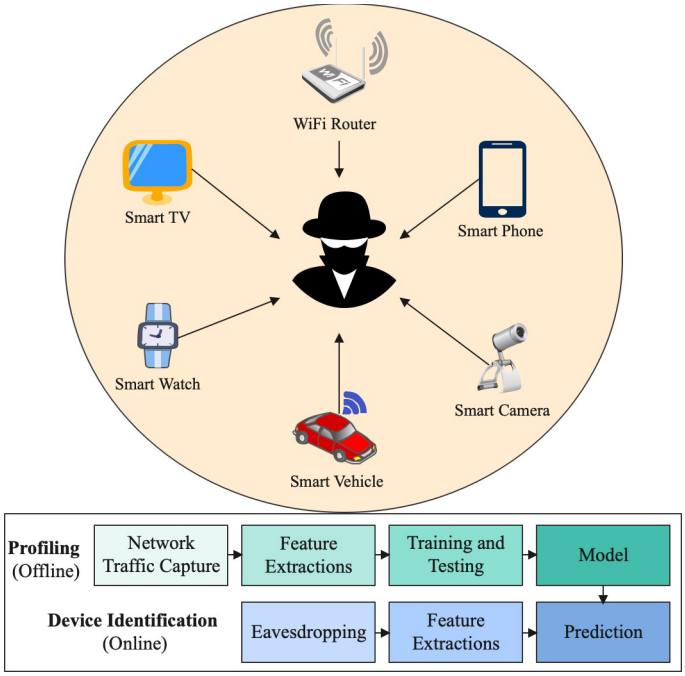


Figure 2: The attacker's perspective of profiling attack on the IoT devices

Table 1: Comparison of captures by out-of-network Airtool with the in-network Wireshark

Size Range of Packets/Frames	Wireshark	Airtool	
	#Packets	#Frames	#Data Frames
0-19	0	2428	0
20-39	0	5593	199
40-79	2441	0	0
80-159	260	2890	2883
160-319	108	239	239
320-639	173	194	190
640-1279	241	255	255
1280-2559	13574	13846	13846
Total	16797	25445	17612

4.2. Traflc Capturing From Outside of the WiFi Net- work

In order to capture the data frames from outside of the WiFi net- work, we first test to use the two most popular sniffing tools, Kismet and Airodump-ng. Kismet stores the network traces as SQLite3 database, whereas Ariodump-ng dump the traces into a capture file format such as pcap. The output of pcap is used to perform packet inspection as the output is in a compatible format, where the packet inspection can be done via a network analyzer such as Wireshark. We use those sniffing tools because of their capabil- ity to sniff raw 802.11 frames. Besides, both of them are able to monitor single-channel and multi-channel using frequency hop- ping. For the hardware, we use an external wireless adapter (Alfa AWUS036ACM) as the built-in WiFi cards don't serve our purpose because they are programmed to accept the data packets which are particularly addressed to the machine's interface card.

Once the captured traffic is analyzed, we observe that a signifi- cant proportion of captured packets contradicts the elephant-mouse internet traffic phenomenon [25]. The elephant flows of 1500 bytes were unable to see for all the available devices, including the video packets captured from a Camera or a smart TV. The proposed in- vestigation shows that the aforementioned tools can only capture a limited range of packets in terms of their sizes. For example, they are able to capture the packet up to the frame size of 472 bytes; how- ever, this size is enough for particular applications such as signal intelligence. Due to such limitations of Kismet and Airodump-ng, we consider another sniffing tool called Airtool5. The Airtool snif- fer is a MAC's built-in sniffing tool that can passively sniff WiFi traffic and store the traces in a pcap format which can be further analyzed using Wireshark. We simultaneously run Wireshark on a different laptop connected to the network to record its own in- coming/outgoing network traffic to the AP in order to confirm the accuracy of the traffic caught by Airtool which is running on an out-of-network MacBook. The traffic between the second laptop connected to the network and the AP was considered for comparison of those two traces.

Because Airtool also records additional control and management frames at the data-link layer (most frames have sizes between 0 and 39), which are absent from Wireshark's in-network traffic captur- ing, we discovered that Airtool collects more frames than packets recorded by Wireshark, as shown in Table 1. Every frame that is collected by Airtool is translated into a WiFi data-link layer frame,

Whereas every frame that is captured by Wireshark is translated into an Ethernet II frame. As a result, for the same WiFi packet, the Wireshark capture is smaller than the Airtool interpretation of the data-link layer frame. This explains why the 2441 packets in the Wireshark capture that are between 40 and 79 bytes all show up in the Airtool capture's higher packet size range (80 to 159 bytes). Additionally, the comparison holds even after excluding all control and management frames from the Airtool capture (as seen in the last column), so there aren't any apparent missing packets according to Airtool. As a result, we employ Airtool as our testbed for evaluation.

4.3. Data Pre-Processing on Captured Data

Once the encrypted WiFi traffic data is captured using the Airtool software, the output in pcap format is analyzed using the Wireshark tool. In particular, the following steps are taken to analyze the captured data:

- We start with the traffic broadcasting in both directions to the MAC address of the WiFi network under investigation. This is required since Airtool could potentially monitor WiFi traffic from many neighboring APs. Only data frame types are kept since all other control, and management MAC-layer frames do not adequately depict the profiling data pattern.
- We export the pcap files into the csv files for the following steps number 3 and 4.
- We eliminate noisy frames that some MACs produced. Since they often only appeared as a single frame, these noise frames are simple to filter out. By just keeping traffic frames with bi-directional communication traffic, they are filtered away.
- To make dataset labeling easier, we swap out the MAC ad- dresses for the relevant device names and their operational status. This step is included only in the offline training stage.

- The dataset required for both offline training and performance testing is eventually obtained using a Python script that ex- tracts and calculates statistical characteristics.

5. Data Processing and Analysis

In the section, we first discuss the observable Data Fields, and then we discuss data analysis. Finally, we discuss device profiling on time-series data.

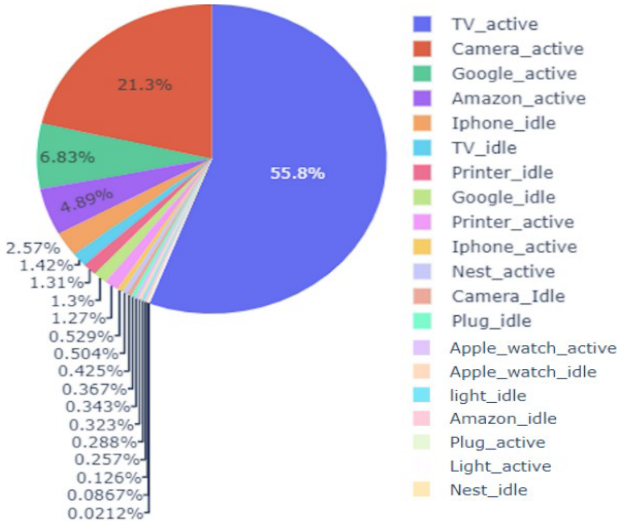


Figure 3: The total number of data packets captured with respect to the IoT device. Many devices' names show whether they are in active or in an idle state.

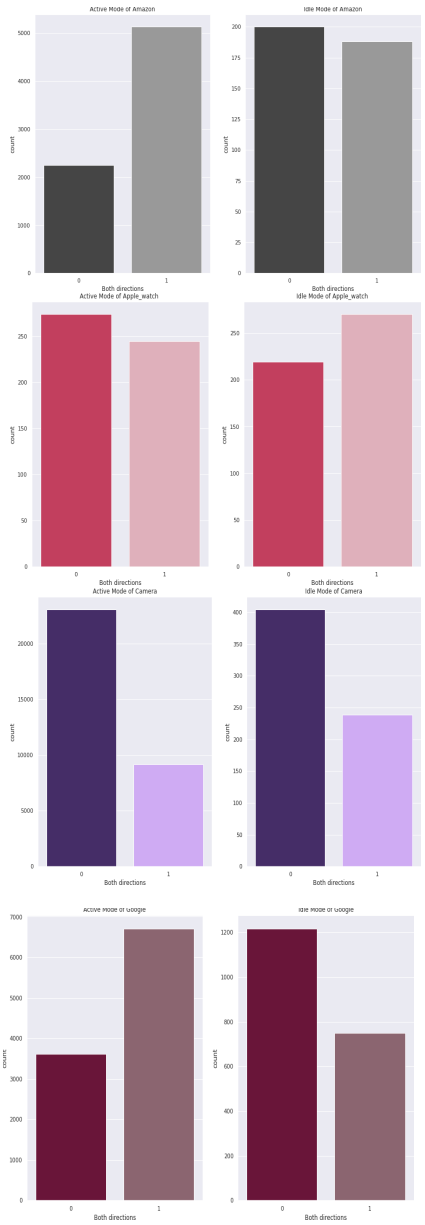
5.1. Useful Data from Network Monitoring

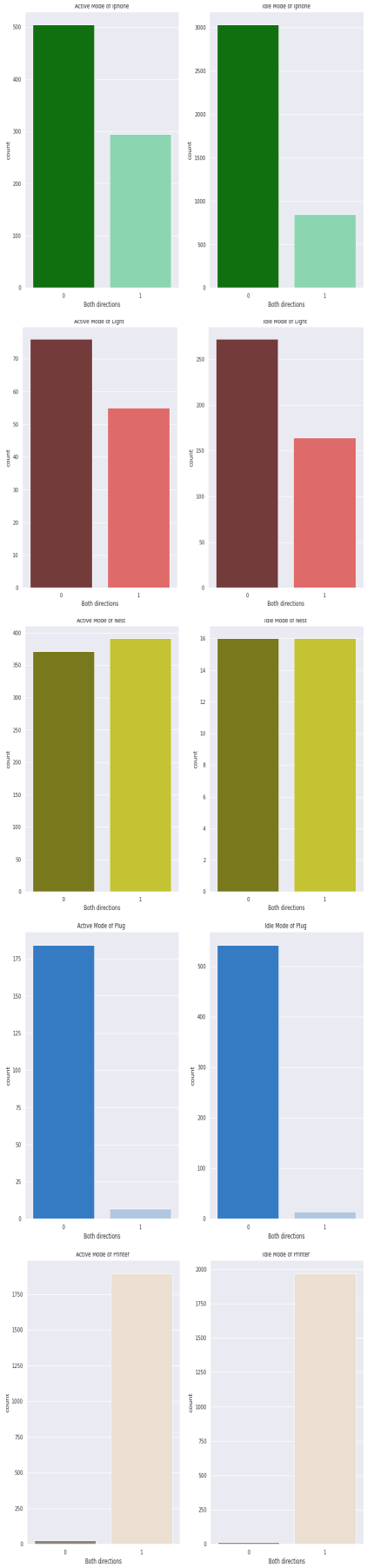
As we have discussed earlier, the process of out-of-network monitor- ing. Here, we present the observable data fields on a secured WiFi network where everything above the data-link layer is encrypted because of WiFi protocol WPA-PSK. Because of this encryption, the only observable data is the MAC-layer frame header, signal strength, and the observation timestamp. The frame header pro- vides the source and destination's MAC addresses, frame size, and frame type. However, the signal strength cannot be used as it has been affected by several factors such as neighboring WiFi net- works, deflection, absorption, and reflections of the surrounding objects. Therefore, in the proposed investigation, we neglect the signal strength and only use the MAC layer frame header.

5.2. Data Analysis

As discussed earlier, we consider 10 different IoT devices and col- lected their data through out-of-network monitoring. In Figure 3, we show the total percentage of captured data from each device. To provide insights into the monitored traffic, we measure the working and idle status of all the IoT devices. From Figure 3, we can observe that the number of received packets from TV is more than the other IoT devices such as the camera and Google. All of the other devices' captured data are comparatively way less than those three devices. Consequently, the camera and google both display different behavior with regard to packet sizes, as shown in Figure 4. Specifically, the camera and google both appear to send the majority of their packets at a fixed size of 170 bytes and 140 bytes, respectively, as shown in Figure 5. We will discuss this part in the following section.

As shown in Figure 3, we believe that the attacker can easily build the signature. Moreover, the attacker can also change the status of those IoT devices. As we can see in Figure 3, there is a huge difference between the active and idle states of the devices which means the AP initiated communication to send an off signal to those devices. On the other hand, due to the notable decline in flow when switching to the idle state, the working condition of other devices with better network capabilities and memory storage, such as iPhones, printers, and Amazon, is impressively noticeable. For instance, Amazon will only get a small number of packets when it is idle because the user is not searching the Internet. Similarly, we show a detailed transmission of data packets from each IoT device in Figure 4. In particular, we show the total number of packets sent from the access point to the IoT device, which is represented by 1, and the total number of packets sent from IoT devices to the access point, which is represented by 0. Moreover, the figure depicts both the total data captured in the active and idle states of all the devices.





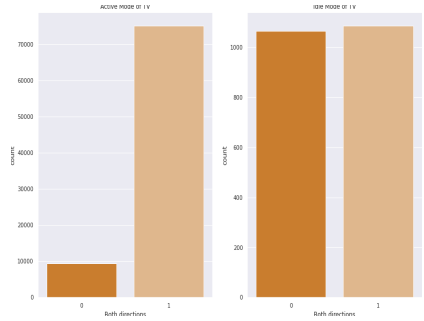


Figure 4: Total number of packet transmission from access point and IoT devices, respectively. In particular, the figure shows the total number of packets sent from the access point to the IoT device, which is represented by **1**, and the total number of packets sent from IoT devices to the access point, which is represented by **0**.

5.3. Machine Learning Techniques

In order to execute our investigation, we choose several classifi- cation methods, and a popular machine learning model XGBoost [26]. We consider XGBoost because of its superior performance, especially for the problems of network classification among other popular machine learning models. Because our data is captured in two different sequence sizes, so we consider precision, recall f-1 score, and support vector machine (SVM) as performance metrics to tackle the time series data.

5.4. Profiling on IoT Devices using Time-series Data

Processing time-series data is simple. The captured traces from the Data-link layer is converted into a string of three-feature items. The monitored frame is then converted into three nu- meric values: the size of packet P, the direction of packet X, and the arrival time Y, where 0 represents the transmitted pack- ets and 1 represents the received packets by an IoT device. Fol- lowing those numeric values, we can obtain the series of data such as {P0, X0, Y0}, {P1, X1, Y1}, ..., {Pn, Xn, Yn}. Figure 5 shows the heatmap of the correlations for each feature in the dataset, where 1 shows the maximum value, and 0 shows the minimum.

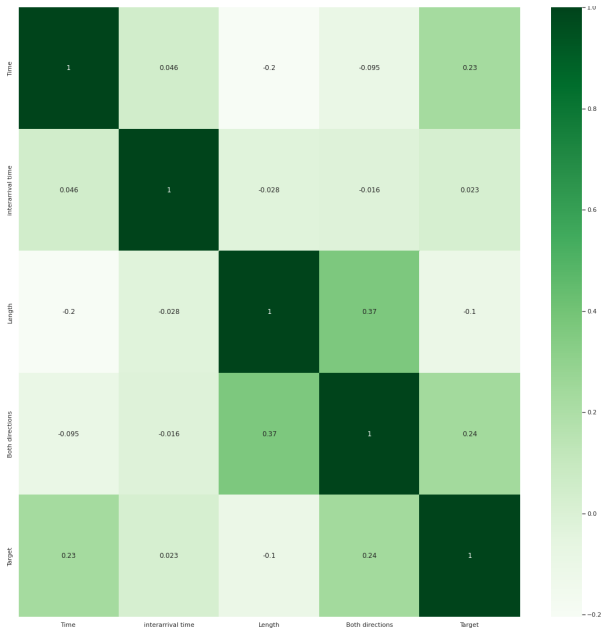
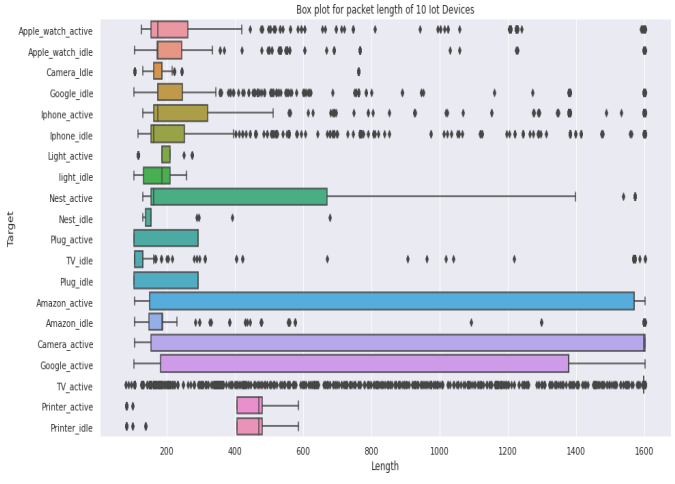


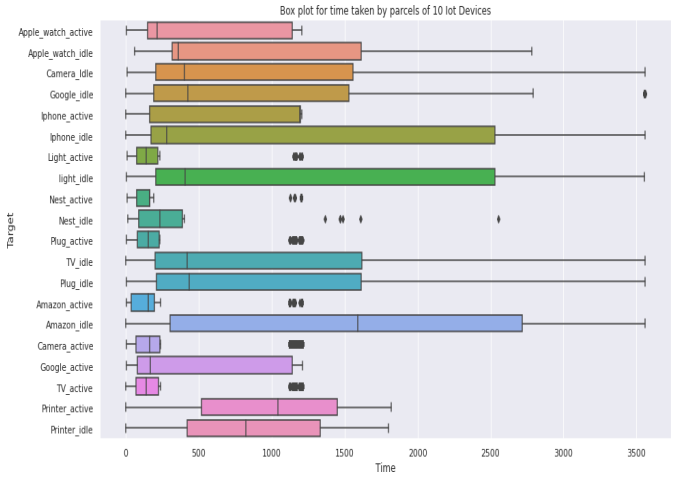
Figure 5: Correlation of features in a dataset

This method's main disadvantage is that it needs a lot of packets to supply all the data points needed for categorization or machine learning training. In our case, we are dealing with a heterogeneous system monitored inside a specific time window; certain devices (such as Smart TV) generate a large volume of data packets while others only produce very sparse packets (such as smart light bulb). For instance, if we compare how long it takes the light and TV to collect a 100-packet series, the TV just needs one second of visible data while the light needs approximately 30 minutes.

To overcome this challenge, we use a two-level categorization technique, starting with a traffic intensity threshold. Devices are divided into two groups in the first level according to whether there is a high or low volume of traffic. Then, in accordance with the volume of device traffic, we use an appropriate sequence size. Using ML algorithms, the second level determines the prediction proba- bility. A prediction is made if the probability rises above a certain threshold; else, the data is labeled as an "unknown" device. Once the dataset is created, we extracted the following features from the monitored traffic in each time window:



(a) Box plot for packet length frames of 10 IoT devices.



(b) Box plot for the time taken by packets of 10 IoT devices1.

Figure 6: Packet length and time-taken in transmission or reception of 10 IoT devices

- Packets transmitted and received by the access point to and from the IoT devices, respectively.
- The difference in inter-arrival time.
- Total number of bytes in the transmitted and received packets.
- Variance of sizes in transmitted and received packets.
- The average number of consecutively transmitted or received packets before seeing a received or transmitted packet, respectively.

6. Results and Discussion

In this section, we first present the testbed settings and the evaluation metrics. Afterward, we show the IoT devices' packets transmission and their reception in terms of packet length, and time. Finally, we show the evaluation results and prove that the outsider intruder can significantly harm the IoT devices without joining the WiFi network.

6.1. Testbed and Evaluation Metrics

With the help of a WiFi router, we built up a testbed with 10 distinct IoT devices. We use AirTool to capture the WiFi data frames be- tween all IoT devices and the WiFi router for an appropriate amount of time in order to collect enough data. Once the data is captured, we use a time-series format and randomly split the dataset into two groups, 20% for testing and 80% for training.

The following metrics are used to assess our classification mod- els: Precision, Recall, F1 Score, and Accuracy. Let's use the abbre- viation T to stand for true prediction, further subdivided into true positives and true negatives. The letters F stand for false prediction, which is further divided into false positives and false negatives. The following equations are used to measure the Precision, Recall, F1 Score, and Accuracy, respectively.

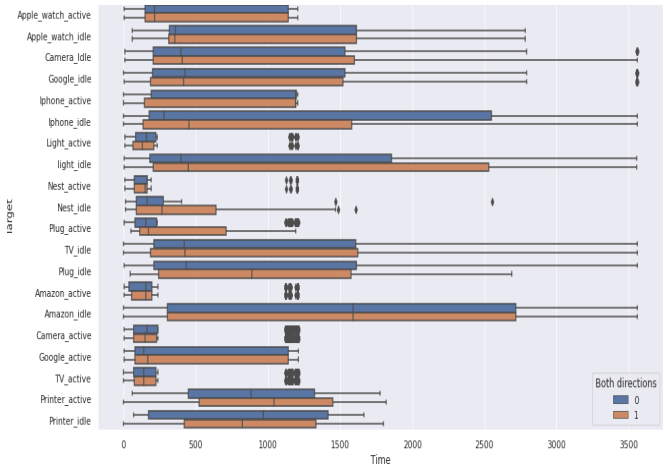


Figure 7: Packets transmission in both directions of each IoT device with respect to time.

- Variance of size distribution in transmitted and received pack- ets.
- Mode of transmitted and received packets.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{} \quad (2)$$

$$TP + FN$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

To access point and access point to IoT device) with respect to time. The line in the middle of the box shows the average length of the packet transmitted or received. In particular, 0 represents the packets transmitted from the IoT device to the access point, and 1 represents

$$Accuracy = T + N \quad (4)$$

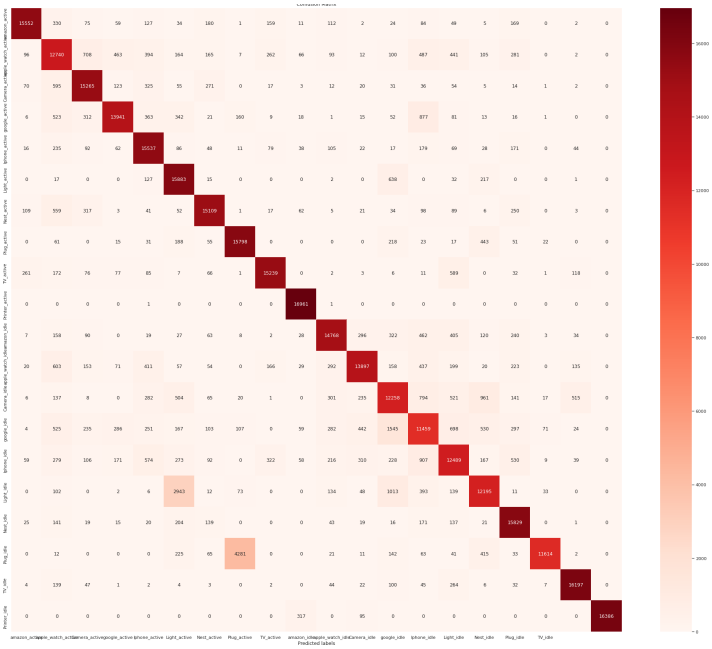


Figure 8: Predicted labels of each IoT device

6.2. Captured Data

Before going to the evaluation, we first show the captured data in terms of packet length and time. In Figure 6, we show the packet length and time taken in transmission or reception of 10 IoT devices, respectively. In particular, in Figure 6 (a), we show a box plot for the packet length of 10 IoT devices. The packet length of each device is captured in both active and idle states. Similarly, in Figure 6 (b), we show a box plot for the time taken by packets to transmit or receive by each IoT device in both active and idle states. To further show the significance of captured data, in Figure 7, we show the packet transmission in both directions (IoT device the packets transmitted from the access point to the IoT device. The above-mentioned box plots prove that a significant amount of data is captured in the proposed investigations, where an attacker can easily access the useful information of the IoT devices and can significantly harm the user. We also prove that the IoT devices can reveal the potential vulnerabilities to hardware/software status as the attacker can change the status of IoT devices.

Considering this as a potential achievement in our investigation, below we present the results of model accuracy, which can further support our claim.

	precision	recall	f1-score	support
1	0.98	0.91	0.95	1841
2	0.75	0.18	0.28	137
3	0.96	0.98	0.97	8025
4	0.85	0.94	0.89	2629
5	0.78	0.65	0.71	202
6	0.61	0.34	0.44	32
7	0.80	0.53	0.63	173
8	0.85	0.23	0.37	47
9	0.98	0.99	0.98	21158
10	0.95	1.00	0.97	483
11	0.87	0.72	0.79	92
12	0.88	0.71	0.79	119
13	0.70	0.51	0.59	162
14	0.77	0.49	0.60	495
15	0.78	0.66	0.72	937
16	0.78	0.63	0.69	105
17	0.00	0.00	0.00	8
18	0.76	0.89	0.82	132
19	0.85	0.94	0.89	538
20	1.00	0.95	0.97	479
accuracy			0.95	37794
macro avg	0.79	0.66	0.70	37794
weighted avg	0.95	0.95	0.95	37794

Figure 9: Machine learning model accuracy on Precision, Recall, and F1-Score.

6.3. Results

We chose a 30-minute time window size for evaluation. We believe that long-term observed traces can teach us more about time-series patterns than short-term ones, which call for much longer time ob- servations to carry out the attack. As we were facing a challenge in working with imbalanced data, for example, data captured from some devices are extremely high such as TV or Camera, whereas other devices are barely showing any record for example Nest, Light as shown in Figure 3. To balance such data, we use SMOTE anal- ysis to prove the significance of the results. Figure 8 shows the confusion matrix of prediction accuracy using a SMOTE analysis on the XGBoost model. The Figure shows that the IoT device ”Printer active” captures maximum true labels, whereas the IoT device ”google idle” captures minimum true labels.

Finally, we show the accuracy of all 10 IoT devices in each active and idle state with respect to Precision, Recall, and F1-Score. In order to provide a significance of the results, here we consider the captured data without balancing and show the result without modify- ing any values. In Figure 9, we show that the model achieves 95% of accuracy. We attribute this performance to the XGBoost model, as it performs significantly well over imbalanced classification datasets.

7. Discussion and Future Work

Our findings support the hypothesis that an outside-of-network at- tacker can successfully identify IoT devices without connecting to a WiFi network. The type of devices and their operating modes may be determined by characteristics like the number of packets, inter-arrival time, packet sizes, and distributions. The attack is straightforward to execute without leaving any traces or fingerprints.

The profiling attack causes serious privacy issues. A potential attacker could drive close to a business to evaluate the volume of economic activity, the socioeconomic background of the clients, expected revenues, revenue trends, or even find possible weak targets for future attacks. It can provide environmental awareness that

can be utilized to track mobile devices in more complicated circum- stances (e.g., cars, drones, phones, etc.). For instance, a swarm of drones can be deployed over a sizable area to classify, identify, and monitor the movement of signal-emitting devices in the covered area. This might show how gadgets communicate with one another and show where each device travels.

As we have proved that the experimental testbed is reliable and has significant importance in the real-world, the implementation in real-world scenarios is of utmost importance to secure the privacy of the individual. To this end, we aim to further extend the proposed implementation in real-world scenarios, where we would be able to show the further importance of such attacks.

8. Conclusion

This paper investigates a privacy leakage from an out-of-network eavesdropper on encrypted WiFi traffic. To this end, we consider 10 IoT devices and capture their data from outside the network without joining the WiFi network. During the investigation, we prove that IoT device eavesdropping is not only possible but also a straightforward process. To this end, we exploit the WiFi frame timing and header information and conduct a detailed evaluation using a machine learning technique for inferring and fingerprinting which IoT device exists in the network and what working status each device is. The models we found had exceptional high accu- racy, they are most likely approaching the point where subsequent improvement becomes more difficult and might even come with the loss of generalizeability [27]. Our evaluation achieves high accuracy, up to 95%, in identifying the devices and their working status. The experimental results show that outside intruders can significantly harm the IoT devices without joining a WiFi network and can launch the attack within a minimum time without leaving any detectable footprints.

References

[1] M. Alyami, I. Alharbi, C. Zou, Y. Solihin, K. Ackerman, "WiFi-based IoT Devices Profiling Attack based on Eavesdropping of Encrypted WiFi Traffic," in 2022 IEEE 19th Annual Consumer Communications & Networking Confer- ence (CCNC), 385–392, IEEE, 2022, doi:10.1109/CCNC49033.2022.9700674.

[2] P. Dudhe, N. Kadam, R. Hushangabade, M. Deshmukh, "Internet of Things (IOT): An overview and its applications," in 2017 International conference on energy, communication, data analytics and soft computing (ICECDS), 2650– 2653, 2017, doi:10.1109/ICECDS.2017.8389935.

[3] G. R. Hiertz, D. Denteneer, L. Stibor, Y. Zang, X. P. Costa, B. Walke, "The IEEE 802.11 universe," IEEE Communications Magazine, **48**(1), 62–70, 2010, doi:10.1109/MCOM.2010.5394032.

[4] K. Boeckl, K. Boeckl, M. Fagan, W. Fisher, N. Lefkovitz, K. N. Megas,E. Nadeau, D. G. O'Rourke, B. Piccarreta, K. Scarfone, Considerations for managing Internet of Things (IoT) cybersecurity and privacy risks, US Department of Commerce, National Institute of Standards and Technology, 2019.

[5] K. Vengatesan, A. Kumar, M. Parthibhan, A. Singhal, R. Rajesh, "Analysis of Mirai botnet malware issues and its prediction methods in internet of things," in International conference on Computer Networks, Big data and IoT, 120–126, Springer, 2018, doi:10.1007/978-3-030-24643-3 13.

[6] E. Ronen, A. Shamir, "Extended functionality attacks on IoT devices: The case of smart lights," in 2016 IEEE European Symposium on Security and Privacy (EuroS&P), 3–12, IEEE, 2016, doi:10.1109/EuroSP.2016.13.

[7] A. Acar, H. Fereidooni, T. Abera, A. K. Sikder, M. Miettinen, H. Aksu, M. Conti, A.-R. Sadeghi, S. Uluagac, "Peek-a-boo: I see your smart home activities, even encrypted!" in Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks, 207–218, doi: 10.1145/3395351.3399421.

[8] S. Aneja, N. Aneja, M. S. Islam, "IoT device fingerprint using deep learning," in 2018 IEEE international conference on internet of things and intelligence system (IOTAIS), 174–179, IEEE, 2018, doi: 10.1109/IOTAIS.2018.8600824

[9] N. Apthorpe, D. Reisman, S. Sundaresan, A. Narayanan, N. Feamster, "Spying on the smart home: Privacy attacks and defenses on encrypted iot traffic," arXiv preprint arXiv:1708.05044, 2017.

[10] M. W. Nadeem, H. G. Goh, M. Hussain, M. Hussain, M. A. Khan, et al., "Internet of Things for Green Building Management: A Survey," in Role of IoT in Green Energy Systems, 156–170, IGI Global, 2021.

[11] F. Meneghello, M. Calore, D. Zucchetto, M. Polese, A. Zanella, "IoT: Internet of threats? A survey of practical security vulnerabilities in real IoT devices," IEEE Internet of Things Journal, **6**(5), 8182–8201, 2019, doi: 10.1109/JIOT.2019.2935189

[12] A. Sujanani, S. Pai, "802.11 Frame-level Network IDS for Public Wireless Networks," in ICT Systems and Sustainability, 453–461, Springer, 2021.

[13] O. Salman, I. H. Elhajj, A. Chehab, A. Kayssi, "A machine learning based framework for IoT device identification and abnormal traffic detection," Trans- actions on Emerging Telecommunications Technologies, **33**(3), e3743, 2022, doi:10.1002/ett.3743.

[14] J. S. Atkinson, J. E. Mitchell, M. Rio, G. Matich, "Your WiFi is leaking: What do your mobile apps gossip about you?" Future Generation Computer Systems, 80, 546–557, 2018, doi:10.1016/j.future.2016.05.030.

[15] A. Sivanathan, H. H. Gharakheili, F. Loi, A. Radford, C. Wijenayake, A. Vish- wanath, V. Sivaraman, "Classifying IoT devices in smart environments using network traffic characteristics," IEEE Transactions on Mobile Computing, **18**(8), 1745–1759, 2018, doi:10.1109/TMC.2018.2866249.

[16] I. Sanchez-Rola, I. Santos, D. Balzarotti, "Clock around the clock: Time-based device fingerprinting," in Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 1502–1514, 2018, doi: 10.1145/3243734.3243796.

[17] R. VanSickle, T. Abegaz, B. Payne, "Effectiveness of tools in identifying rogue access points on a wireless network," 2019.

[18] P. Serrano, M. Zink, J. Kurose, "Assessing the Fidelity of COTS 802.11 Snif- fers," in IEEE INFOCOM 2009, 1089–1097, 2009, doi:10.1109/INFCOM.2009.5062021.

[19] A. Abusnaina, R. Jang, A. Khormali, D. Nyang, D. Mohaisen, "DFD: Adversarial Learning-based Approach to Defend Against Website Fingerprinting," in IEEE INFOCOM 2020 - IEEE Conference on Computer Communications, 2459–2468, 2020, doi:10.1109/INFOCOM41043.2020.9155465.

[20] W. De la Cadena, A. Mitseva, J. Hiller, J. Pennekamp, S. Reuter, J. Filter, T. Engel, K. Wehrle, A. Panchenko, "Trafficsliver: Fighting website fingerprint- ing attacks with traffic splitting," in Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, 1971–1985, 2020, doi:10.1145/3372297.3423351.

[21] F. Zhang, W. He, X. Liu, "Defending against traffic analysis in wireless networks through traffic reshaping," in 2011 31st International Conference on Distributed Computing Systems, 593–602, IEEE, 2011, doi: 10.1109/ICDCS.2011.77.

[22] Z. Jiang, K. Zhao, R. Li, J. Zhao, J. Du, "PHYAlert: identity spoofing attack detection and prevention for a wireless edge network," Journal of Cloud Computing, **9**(1), 1–13, 2020, doi:10.1186/s13677-020-0154-7.

[23] Q. Xu, R. Zheng, W. Saad, Z. Han, "Device fingerprinting in wireless networks: Challenges and opportunities," IEEE Communications Surveys & Tutorials, 18(1), 94–104, 2015, doi: 10.1109/COMST.2015.2476338.

[24] I. I. A. Sulayman, R. He, M. Manka, A. Ning, A. Ouda, "LiFi/WiFi Authentication and Handover Protocols: Survey, Evaluation, and Recommendation," in 2021 International Symposium on Networks, Computers and Communications (ISNCC), 1–6, IEEE, 2021, doi: 10.1109/ISNCC52172.2021.9615853.

[25] K.-C. Lan, J. Heidemann, "On the correlation of internet flow characteristics," Technical report, Citeseer, 2003.

[26] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, et al., "Xgboost: extreme gradient boosting," R package version 0.4-2, 1(4), 1–4, 2015.

[27] I. A. Alharbi, A. J. Almalki, C. C. Zou, "Hyperparameter Optimization and Comparison of Student Performance Prediction Algorithms," in 2021 Interna- tional Conference on Computational Science and Computational Intelligence (CSCI), 889–894, 2021, doi:10.1109/CSCI54926.2021.00207.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 3, 51-61 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Generalized Linear Model for Predicting the Credit Card Default Payment Risk

Lu Xiong[*,1,2], Spendylove Duncan-Williams[3]

[1]*Department of Mathematical Sciences, Middle Tennessee State University, Murfreesboro, TN 37132, USA*

[2] *Computational and Data Science Doctoral Program, Middle Tennessee State University, Murfreesboro, TN 37132, USA*

[3] *Master of Science in Professional Science Program, Middle Tennessee State University, Murfreesboro, TN 37132, USA*

ARTICLE INFO

Article history:
Received: 13 February, 2022
Accepted: 04 May, 2022
Online: 25 May, 2022

Keywords:
Credit Card Default Risk
Generalized Linear Model
Decision Tree
Random Forest

ABSTRACT

Predicting the credit card default is important to the bank and other lenders. The credit default risk directly affects the interest charged to the borrower and the business decision of the lenders. However, very little research about this problem used the Generalized Linear Model (GLM). In this paper, we apply the GLM to predict the risk of the credit card default payment and compare it with a decision tree, a random forest algorithm. The AUC, advantages, and disadvantages of each of the three algorithms are discussed. We explain why the GLM is a better algorithm than the other two algorithms owing to its high accuracy, easy interpretability, and implementation.

1. Introduction

The credit card debt crisis has been a major concern in the capital market and among card-issuing institutions for many years. Credit cards and cash-card debts are abused by most card users, regardless of their payment capabilities. This crisis poses a great threat to both cardholders and banks. The payment default means the failure to pay the credit card bills. Researchers have attempted to forecast the credit card customers' default payments using machine learning techniques [1]. The public's lack of understanding of basic financial principles, as seen by the recent financial crisis, has proven that they are unable to make sound financial judgments.

Individuals' minimum monthly credit card payments should be researched in depth to discover the link between consumer income, payment history, and future default payments [2]. Increasing consumer finance confidence, to avoid delinquency is a big challenge for cardholders and banks as well. In a well-established financial system, risk assessment is more important than crisis management [2]. With historical financial data, for example, business financial statements, client transaction and reimbursement records, etc., we can predict business execution or individual clients' default risk and lessen the harm and instability. To make accurate customer risk assessments for their credit services department, banks are required by having sophisticated credit scoring systems to automate the credit risk scoring tasks [3]. Management of the credit risk for the banking sector and financial organizations have extensively started to gain importance. Developing an automated system to accurately forecast the probability of cardholder's future default, will help not only to manage the efficiency of consumer finance but also effectively handle the credit risk issues encountered in the banking sector [3], [4].

Individual firms have been gathering massive amounts of data every day in the era of big data. Finding the relevant information from data and turning that information into meaningful outcomes is a big issue for businesses. As a result, in this article, we investigate the risk of failure to pay the minimum credit card amount, which is a transaction that should be done monthly. The Generalized Linear Model (GLM), single classification tree, and random forest are compared to predict the credit default risk. Prediction accuracy and interpretability are the two main factors we consider when selecting the final model. The principal component analysis (PCA) is used to reduce the data dimensionality.

2. Literature Review

Machine learning is the act of automatically or semi-automatically exploring and analyzing massive amounts of data to uncover significant patterns and rules [5]. It has been utilized in different types of financial analysis such as predicting money laundering, stock analysis, detection of bankruptcy, the decision of loan approval, etc. [6, 7]. Machine learning algorithms are

[*]Corresponding Author: Lu Xiong, Email: lu.xiong@mtsu.edu

inadequately used for detecting default payments of credit card users. Different kinds of research are ongoing to improve the accuracy of machine learning algorithms in predicting default payment of credit card users, and a small portion of improvement plays a vital role in the economic developments of the related organizations [8, 9]. Before the 1980s, some statistical methods such as Linear Discriminant Analysis (LDA) [10], and Logistic Regression (LR) [11] were used to estimate the credit default probability. Starting from the 1990s, machine learning methods such as K-nearest neighbor (KNN) [12], neural network (NN) [13], genetic algorithm [14], and support vector machine (SVM) [15] were used to assess the default risk of credit card users. In 2018, researchers compared 5 data mining methods on credit card default prediction: logistic regression, SVM, neural network, Xgboost, and LightGBM [16]. In 2020, AdaBoost was employed to build credit default prediction models [17]. In 2020, researchers investigated the credit card default prediction in the imbalanced data sets [18].

However, the GLM is rarely used in credit card default prediction. GLM has the advantage of easy interpretability and implementation. We believe it's worthwhile to apply the GLM in the research of credit card payment default prediction. In this paper, we will compare the GLM with the decision tree and random forest and explain its advantages.

3. Data Description and Exploratory Analysis

3.1. Data description

The data set came from the credit cardholders from a major bank in Taiwan in October 2005, with a total of 25000 records. The target variable to be predicted in this data is a binary variable – default payment (Yes = 1, No = 0). Table 1 lists the 25 variables in the data including all the predictors and the target variable (default status). The first 5 of the predictors are demographic characteristics. The next 6 variables are about the status of past payments. The further 6 variables are about the amount of the past bill statement. The next 6 variables are about the amount of paid bills and the last 2 predictors are the limit balance of individuals and the default status of the client. This credit payment data has a typical characterization of imbalanced datasets in terms of the target variable, for 5529 records (22.12%) are the records of default and 19471 records (77.88%) are non-default payments.

Table 1: Data dictionary

VARIABLE NAME	DESCRIPTION	FACTOR LEVELS
ID	The ID of each client	Values
SEX	Gender	1=male 2=female
EDUCATION	The level of education of each client	0= Unknown 1=graduate school 2=university 3=high school 4=others
MARRIAGE	Marital Status	1=married 2=single, 3=others
AGE	Age in years	Ages
LIMIT_BAL	The amount of given credit in NT dollars includes individual and family credit	Numerical NT Dollars
PAY_0	Repayment status in September 2005	-2=payment made two months earlier, -1=pay duly, 0=paid right before due 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months,)
PAY_2	Repayment status in August 2005	scale same as above
PAY_3	Repayment status in July 2005	scale same as above
PAY_4	Repayment status in June 2005	scale same as above
PAY_5	Repayment status in May 2005	scale same as above
PAY_6	Repayment status in April 2005	scale same as above
BILL_AMT1:	amount of bill statement in September 2005	Numerical NT dollars
BILL_AMT2:	Amount of bill statement in August 2005	Numerical NT dollars
BILL_AMT3	Amount of bill statement in July 2005	Numerical NT dollars
BILL_AMT4	Amount of bill statement in June 2005	Numerical NT dollars
BILL_AMT5	Amount of bill statement in May 2005	Numerical NT dollars
BILL_AMT6	Amount of bill statement in April 2005	Numerical NT dollars
PAY_AMT1	Amount of previous payment in September 2005	Numerical NT dollars
PAY_AMT2	Amount of previous payment in August 2005	Numerical NT dollars
PAY_AMT3	Amount of previous payment in July 2005	Numerical NT dollars
PAY_AMT4	Amount of previous payment in June 2005	Numerical NT dollars
PAY_AMT5	Amount of previous payment in May 2005	Numerical NT dollars
PAY_AMT6	Amount of previous payment in April 2005	Numerical NT dollars
default.payment.next.month	Default Status	1=Yes 0=No

3.2. Explore the relationship between the predictors and the target variable

In this section, relations between the predictors vs the target variable will be explored using graphs.

Figure 1 is the stacked histogram that shows the distribution of the credit limit (limit balance) among defaulting records (blue bars) and non-defaulting records (pink bars).

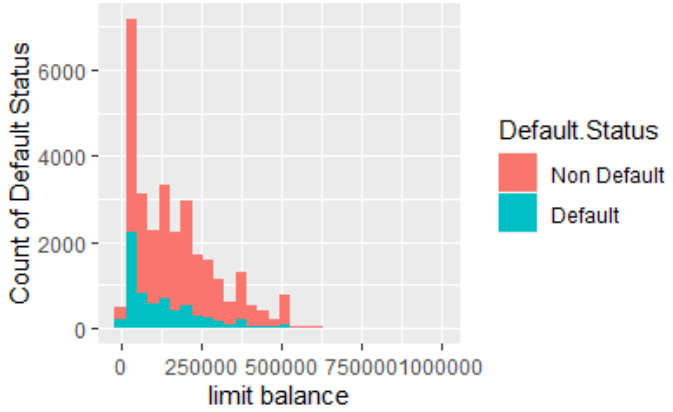


Figure 1: The stacked histogram of the credit limit in default records and non-default records.

From Figure 1, we can observe the default records have a higher proportion of Lower LIMIT_BAL values than non-defaulting records. To confirm this observation, we did a two-sample t-test with a null hypothesis mean of LIMIT_BAL in the non-default group is less than in the default group. The p-value is less than 10e-15 which means we can confidently accept the alternative hypothesis: the mean of LIMIT_BAL in the non-default group is higher than in the default group. This confirms our observation in Figure 1. This conclusion matches the reality that lower credit balances are usually issued to credit users with higher default risk. This distinctive characteristic of the predictor LIMIT_BAL indicates it's a good variable to predict default.

Table 2: Some details of hypothesis tests.

Alternative Hypothesis	Type of Hypothesis Test	P-value
The mean of LIMIT_BAL in the non-default group is greater than in the default group	T-test	$<2.2\times 10^{-16}$
The chance of default at age 25-40 is lower than at other ages.	Two-proportions z-test	$<2.2\times 10^{-16}$

Figure 2 shows the distribution of the age among default credit users and non-default users, using a density histogram. We observe the non-default group has a higher proportion of age 25-40, but a lower proportion of other ages. This makes us infer the age 25-40 has a lower chance of default. To confirm this as the alternative hypothesis, we did a two-proportions z-test. The P-value as shown in Table 2 is very significant and confirms our hypothesis. The reason age 25-40 has a lower default rate could be people in this age range are in the working-age, healthy, and don't have much financial burden from their family yet. In contrast, people younger than 25 may have not started their careers yet, therefore with little to no income. People older than 40 could be less healthy, retired, or need financial support from their families such as children's college education. The above discussion indicates the variable Age can be a good predictor.

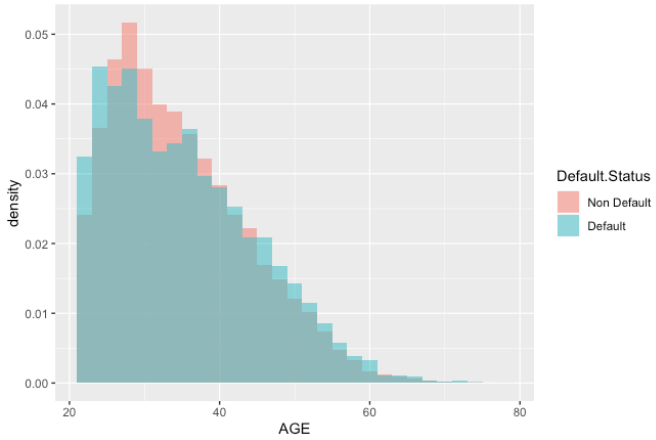


Figure 2: The stacked histogram of age in default records and non-default records.

Figure 3 represents the default status distribution within different gender. The percentage of defaults in males (24%) is slightly higher than in females (20%). To confirm this difference is statistically significant, we did a two-proportion z-test with the alternative hypothesis that male has a higher default chance than female. The p-value of this z-test is 2.47×10^{-12}, which means we should accept the alternative hypothesis. This confirms our observation. This difference is probably because females are more conservative when managing their personal finance than males, which results in lower default risk. This difference indicates the variable Sex is a good predictor.

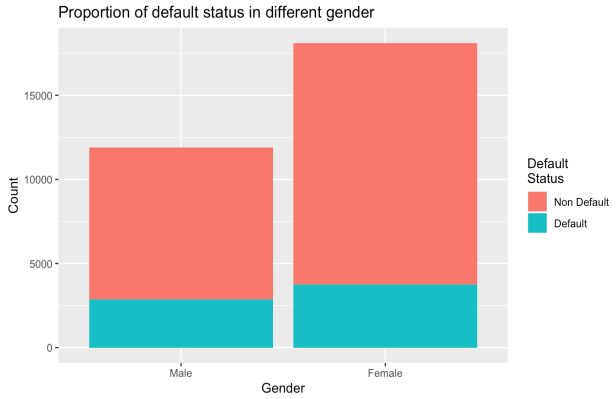


Figure 3: Distribution of default and non-default records in males and females.

Figure 4 displays the proportion of education levels in non-default and default records. It shows that the non-default records have a larger proportion of higher educated individuals from graduate school and university levels. This also matches our intuition that higher education levels associate with higher income

and lower default risk. To statistically confirm this observation, we did a two-proportion z-test with an alternative hypothesis that non-default users have a higher proportion of education level of university or graduate school. The p-value of this test is 0.00226, which confirms our observation.

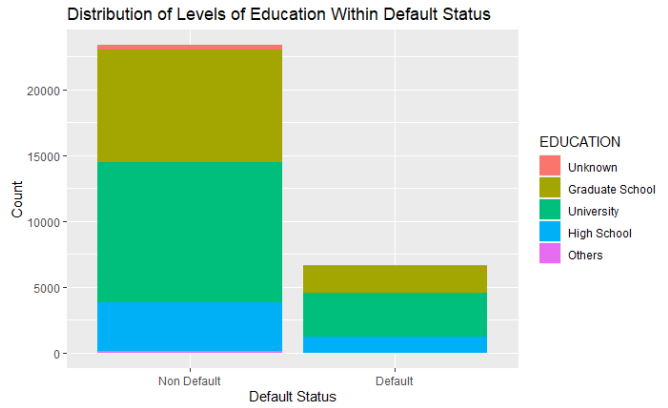


Figure 4: The distribution of education level in default and non-default.

Figure 5 displays the proportion of different lateness in the last statement payment (PAY_0) in default records and non-default records. The variable PAY_0 has more proportion of "On Time" or "Early" values in the non-default records than in default values. We also did a two-proportions z-test to confirm this observation as the alternative hypothesis in the z-test. The p-value of the test is less than 2.2×10^{-16}, which confirms the alternative hypothesis. This means that being current or ahead of payments is associated with non-defaulting in the following month. Therefore the lateness in the last statement payment (PAY_0) is a good predictor to predict the default. For the lateness of the last 2nd payment to the last 7th payment (PAY_1 to PAY_6), we observed the same pattern. That is, non-default records have a higher proportion of "On Time" or "Early" payments in each of the past 7 months than default records. This suggests the payment statuses in the past months have potentially strong predictive power on the default risk.

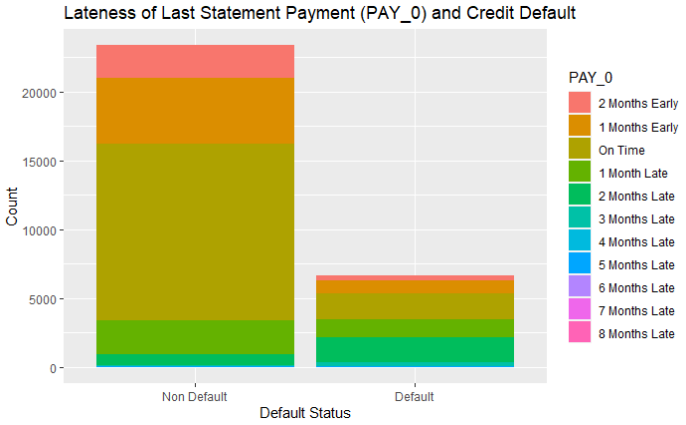


Figure 5: The distribution of PAY_0 values in non-default and default values. Variable PAY_0 is the lateness of the last statement payment

3.3. Reduce the levels of factor variables to increase the predictive power

If a factor variable in the data has too many different values (levels), its predictive power is usually lower. This is because the factor variable is often expanded as the combination of multiple dummy variables when predicting the target variable, where each dummy column is one possible value of the factor variable. If there are many possible values in the variable, then this expanded data will have too many columns versus the number of rows. The data rows will not be long enough compared with its dimensionality to train a highly accurate and robust algorithm. To reduce the data dimensionality, we need to reduce the levels of the factor variables by combining similar levels, so the prediction accuracy can be improved.

Table 3 shows the number of non-default and default records in the factor variable Education. According to this table, the university and high school education have higher proportions of defaults than the other values. We combine these 2 levels as one level and name it "HighSchoolUniversity". Since "others" and "unknown" have a lower proportion of default, they will be combined and named 'others'. The factors are then reduced to HighSchoolUniversity, Graduate School, and Others.

Table 3: Data dictionary

EDUCATION	Non-default	Default	Total	Proportion of default
Unknown	319	26	345	8%
Graduate School	8549	2036	10585	19%
University	10700	3330	14030	24%
High School	3680	1237	4917	25%
Others	116	7	123	6%

Table 4: The proportion of non-default in each level of predictor PAY-0 before its levels combined.

PAY _0	Non Default	Default	Total	Proportion of default
2 Months Early	2394	365	2759	13%
1 Month Early	4732	954	5686	17%
On Time	12849	1888	14737	13%
1 Month Late	2436	1252	3688	34%
2 Months Late	823	1844	2667	69%
3 Months Late	78	244	322	76%
4 Months Late	24	52	76	68%
5 Months Late	13	13	26	50%
6 Months Late	5	6	11	55%
7 Months Late	2	7	9	78%
8 Months Late	8	11	19	58%

Table 4 shows the number of non-default and default records in the factor variable PAY_0. The values of non-default in PAY_0

= "2 Months Early", "1 Month Early", and "On Time" have a similar lower proportion of defaults than the other PAY_0 values. We combine these 3 levels as one level and name it "Pay duly". Also, the PAY_0 = "2 Months Late" or more have a similarly high proportion of default and it will be reasonable to combine them as one level. PAY_0 = "1 Month Late" stands as the other level. Therefore the number of levels in predictor PAY_0 is reduced to 3.

3.4. Using Kernel Principal Component Analysis to generate a new feature

The predictors in this data may not be independent of each other. The predictors SEX, EDUCATION, and MARRIAGE might be related to each other. For instance, the male may have higher education, but a lower percentage of getting married compared with female. We can use the kernel principal component analysis (Kernel-PCA) [19] to generate new variable(s) to represent the related variables so that the dimensionality of the data can be further reduced. The kernel-PCA works better than PCA when nonlinear relation exists between initial variables in the data. Since there is an interaction between the variables according to our later discussion in session 4.2, such nonlinear relation exists. We run the kernel-PCA on the data that only contains the three variables SEX, EDUCATION, and MARRIAGE to find out the principle components among them. Table 5 is the kernel-PCA results showing the first 4 principal components (PC). The first 4 principal components combined can describe 87% of the variation.

Table 5: The importance of the first 2 principal components.

	PC1	PC2	PC3	PC4
Proportion of variance	0.270	0.240	0.207	0.153
Cumulative Proportion	0.270	0.511	0.718	0.871

The new features PC1, …, PC4 will replace the total 8 levels in the variables SEX, EDUCATION, and MARRIAGE in the later predictive modeling to predict the target variable.

4. Methodology

In this session, we will introduce the main idea of GLM and apply it to the data pre-processed in previous sessions. The decision tree and random forest algorithms will also be applied to compare the GLM. The data was split into training (70%) and test (30%) set for each of the algorithms.

4.1. Introduction of GLM

The ordinary linear regression assumes the target variable follows the normal distribution. However, the target variables in many real-world data don't follow the normal distribution. For instance, personal income follows a lognormal distribution with the majority of people in middle income and few are super-rich. In the credit data used in this paper, the target variable is a binary variable that obviously doesn't follow the normal distribution. For the non-normal distributed target variable, the ordinary linear regression no longer works. The GLM is the generalized version of the linear regression that works for a broader range of target variables including non-normal and normal distributed. The GLM uses a link function to transform the non-normal target variable to normal distributed, then the ordinary regression method can be used.

4.2. Select an Interaction

Even though the generalized linear model (GLM) is a linear model, it can still deal with the non-linear relation. That's done by including the interaction term. The interaction means the relationship between one predictor and the target changes when the value of another predictor changes. This suggests the two predictors are not independent of each other but exist in interaction between them.

In Figure 6, we look at how education and marriage interact. Individuals who are married and are in high school have a larger proportion of non-default than those who are single. However, individuals who are single and are in higher education have a higher proportion of non-default than married people. This might be because single people spend less than married people and have a higher proportion of non-default. It is also worth noting that there are more singles in higher education than married people who are in higher education, as well as a large number of married high school graduates. We observe that at each level of education the number of people who are married and will default varies with no pattern hence we will select this interaction between marriage and education.

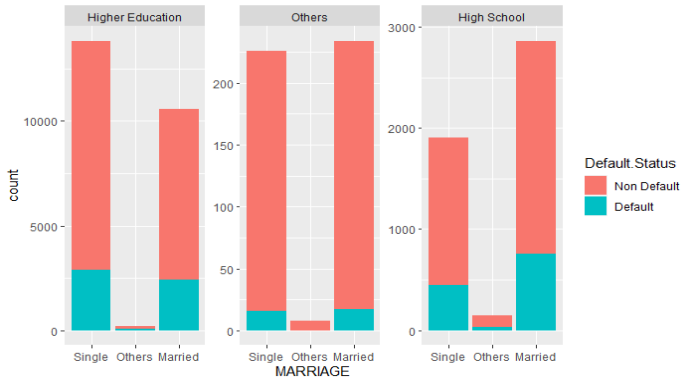


Figure 6: The distribution of default status with Education for each marital level.

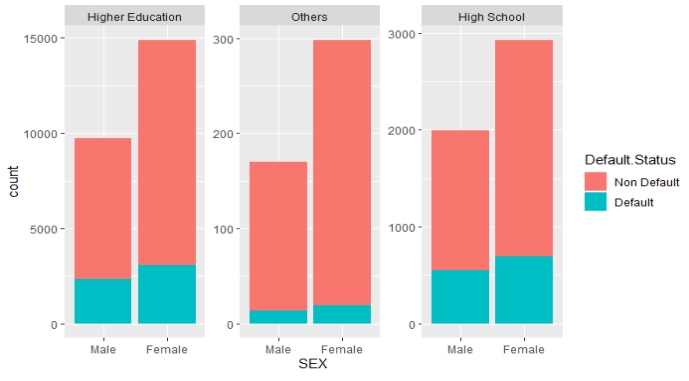


Figure 7: The distribution of default status with Education for each gender.

The interaction between education and sex is explored in Figure 7. There are much more females in all three levels of education who will default than males. Females, on average, have a greater degree of education than males and are less likely to

default. There is an interaction between sex and education because of the pattern it has in each educational level.

We do the same thing to explore the interaction between MARRIAGE and SEX using Figure 8. When SEX = "male", the married male has a significantly lower default risk than the single males. However, when SEX = "female", the default risk is not significantly different between the single females and married females. To explain this, we can regard it's often the men who pay more bills in the marriage. Married men should take more financial responsibility after getting married than single men. Therefore they have higher default risk. While for females, such financial responsibility in marriage is not so significant, thus there is not much difference in terms of default risk between married females and single females. Therefore, the SEX value impacts.

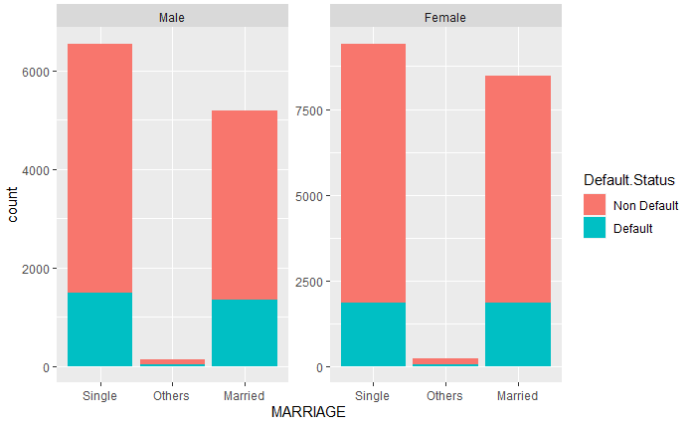


Figure 8: The distribution of default status with Gender for Marital status

4.3. Link Function Selection

We need to specify the distribution of the target variable and what link function to be used. Our target variable is binary, which is either "default" or "not default", the natural choice of the target distribution is binomial.

For the link function, it must map the prediction to the zero to one range because we will first predict the probability of default, then classify it. The binomial distributed GLM has four candidates for the link functions: logit, probit, Cauchit, and cloglog. They will all map the regression result to a value between 0 and 1, which can be regarded as the payment default probability. Table 6 provided details about these 4 link functions.

In Table 6, the $\boldsymbol{X^T\beta}$ is the regression equation. p is the target variable. With the link function, the target variable is transformed into a normal distributed variable that can be regressed using the regular least squares method. The AIC is used as an important metric when we decide which link function to choose. Each link function is tried on the training data and the AIC of its GLM is listed in the last column of Table 6. The Cauchit link and Cloglog link have higher AIC, so we exclude them. The Logit and Probit have similar small AIC, so either of them can be chosen as the link function. Since the Logit link is the canonical link for binary family and it's more widely used, we decide to choose Logit function as the link function.

Table 6: Four link functions for binary distribution and its AIC on our training data.

Name	Link Function	Response Probability	AIC on training data
Logit	$\log\left(\frac{p}{1-p}\right) = \boldsymbol{X^T\beta}$	$p = \frac{\exp(\boldsymbol{X^T\beta})}{1+\exp(\boldsymbol{X^T\beta})}$	18388
Probit	$\Phi^{-1}(p) = \boldsymbol{X^T\beta}$	$p = \Phi(\boldsymbol{X^T\beta})$	18386
Cauchit	$\tan(\text{pv} - \text{p}/2) = \boldsymbol{X^T\beta}$	$p = \frac{1}{\pi} arctan(\boldsymbol{X^T\beta}) + \frac{1}{2}$	18451
Cloglog	$\log(-\log(1 - \text{p})) = \boldsymbol{X^T\beta}$	$p = 1 - exp(-exp(\boldsymbol{X^T\beta}))$	18468

4.4. Feature selection

Feature selection (also known as variable selection, attribute selection, or variable subset selection) is the technique of selecting a subset of relevant features (predictors and variables) for use in the development of a model. It is the automatic selection of the most significant and relevant qualities contained in the data for predictive modeling [20]. Forward selection and backward selection are the two main types of feature selection methods. Forward selection is an iterative approach in which the model starts with no variables. This method keeps adding up the variable that improves the model the most (measured by AIC) in each iteration until adding a new variable no longer enhances the model's performance. The AIC of forward selection is 18380. The backward selection begins with all the variables (full model) and removes the least significant variable one after the other until its AIC no longer decreases. The AIC of backward selection is 18390. The best model solely depends on the defined evaluation criterion of which the AIC was used. Since forward selection has a lower AIC, it's used as the feature selection method. This choice removed variables Age, PC2, PC3, PC4, BILL_AMT3,4,5, PAY_AMT3,5 by not selecting them. The area under curve (AUC) for the test data is reduced to 0.7637 with an AIC decreasing to 18442.

4.5. Data Sampling and G-K-Fold Cross-validation.

To ensure the distribution is unchanged after data is spitted into the training set and testing set, we use stratified sampling. Both sets contain the same portion of credit default data after the data partition. To analyze the accuracy and stability of different algorithms, the G-K-fold stratified cross-validation is used. In K-fold stratified cross-validation, the data is stratified partitioned into K equal parts, where each part of the data has the same distribution for the target variable. 1 part of this data is defined as the testing data, the remaining K-1 parts of the data are the training data. G-K-folder stratified cross-validation will do the K-folder stratified cross-validation for G times, to generate enough results for algorithms performance evaluation statistically.

4.6. Interpretation of the GLM Results

Table 7 listed the results of the regression coefficients of the GLM. It's the trained GLM model generated from the training data

using the features we selected from the step-AIC method based on the new features built from kernel-PCA and variables interaction considered. The "Estimate" column contains the regression coefficients. According to this table, the following variables, levels or interactions are significantly important in predicting the default payment, due to the small p-value and relatively large estimated coefficients:

- PAY_0 = 1 month late
- PAY_0 = More than a month late
- PAY_2 = More than a month late
- PAY_3 = More than a month late
- PAY_4 = More than a month late
- PAY_5 = More than a month late
- PAY_6 = More than a month late
- EDUCATIONOthers:MARRIAGESingle

This makes sense. The PAY_0 is more important than the other payment status variables because it's the most recent payment status. The PAY_0=More_than_a_month_late has a larger estimated coefficient than PAY_0=1_month_late because the former level is associated with a higher probability of default payment, and both levels result in a higher chance of default compared with PAY_0 paid in time. The negative coefficients of interaction levels between EDUCATION and MARRIAGE mean they indicate a lower probability of default payment.

Table 7: Summary of the GLM results.

	Estimate	Std. Error	z value	Pr(>|z|)	Signif. Codes
(Intercept)	-10.51	22.62	-0.47	0.64	
BILL_AMT1	<1E-4	<1E-4	-2.69	0.01	**
BILL_AMT2	<1E-4	<1E-4	3.78	<1E-3	***
BILL_AMT6	<1E-4	<1E-4	-3.10	<1E-2	**
PAY_AMT1	<1E-4	<1E-4	-5.32	<1E-4	***
PAY_AMT2	<1E-4	<1E-4	-2.55	0.01	*
PAY_AMT4	<1E-4	<1E-4	-0.82	0.41	
Limit.Balance	<1E-4	<1E-4	-6.86	<1E-4	***
PAY_AMT6	<1E-4	<1E-4	-2.93	0.00	**
PAY_01 month late	0.78	0.06	13.10	< 2E-16	***
PAY_0More than a month late	2.00	0.06	30.92	< 2E-16	***
PAY_21 month late	-1.26	1.05	-1.20	0.23	
PAY_2More than a month late	0.23	0.07	3.17	<1E-3	**
PAY_31 month late	-10.33	228.91	-0.05	0.96	
PAY_3More than a month late	0.27	0.07	3.79	<1E-4	***
PAY_41 month late	0.87	397.31	0.00	1.00	
PAY_4More than a month late	0.21	0.08	2.76	0.01	**
PAY_5More than a month late	0.27	0.08	3.20	<1E-3	**
PAY_6More than a month late	0.33	0.07	4.49	<1E-4	***
PC1	0.72	2.04	0.35	0.72	
PC2	-2.57	2.57	-1.00	0.32	
PC3	-2.66	4.09	-0.65	0.52	
PC4	-4.28	4.00	-1.07	0.28	
EDUCATIONHigher Education:MARRIAGESingle	11.16	22.68	0.49	0.62	
EDUCATIONOthers:MARRIAGESingle	-51.98	35.15	-1.48	0.14	
EDUCATIONHigh School:MARRIAGESingle	-4.97	13.57	-0.37	0.71	
EDUCATIONHigher Education:MARRIAGEOthers	-45.27	55.01	-0.82	0.41	
EDUCATIONOthers:MARRIAGEOthers	-88.72	127.20	-0.70	0.49	
EDUCATIONHigh School:MARRIAGEOthers	-53.17	38.30	-1.39	0.17	
EDUCATIONHigher Education:MARRIAGEMarried	16.43	34.10	0.48	0.63	
EDUCATIONOthers:MARRIAGEMarried	-48.24	41.72	-1.16	0.25	
EDUCATIONHigh School:MARRIAGEMarried	NA	NA	NA	NA	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The trade-off between sensitivity (or TPR) and specificity (1 – FPR) is depicted by the ROC curve. Classifiers with curves that are closer to the top-left corner perform better. A random classifier is supposed to offer diagonal points (FPR = TPR) as a baseline. The test becomes less accurate as the curve approaches the 45-degree diagonal of the ROC space to the left. The area beneath the ROC curve is referred to as the AUC [21]. To generate the ROC curve, we use the roc() function in R's pROC library by comparing the model prediction probability vs the real target value. The ROC curve of this GLM method on the test data is provided in figure 9 and it has an AUC of 0.757.

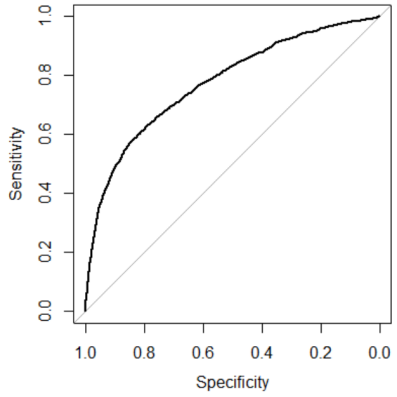


Figure 9: ROC curve of the GLM model

4.7. Classification Tree

Classification tree methods (i.e., decision tree methods) are recommended when the machine learning task contains

classifications or predictions of the outcome. A Classification tree labels, records, and assigns variables to discrete classes. A Classification tree is built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions and then splitting it up further on each of the branches. Before constructing a tree, the data was split into training (70%) and testing (30%) sets.

We use the rpart package in R to implement the decision tree. With its tree structure, the decision tree can automatically incorporate the non-linear relation between variables. Therefore, we don't need to add the interaction term (SexEduMarriage) as the input variable. All the variables in the original data are included as the input variables to build the decision tree. Gini impurity is used to split the tree. Figure 10 is the classification tree produced.

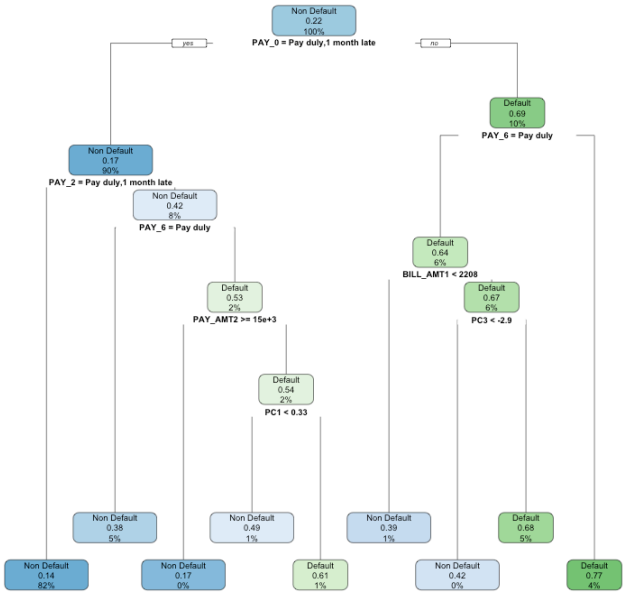


Figure 10: Classification tree result.

A confusion matrix is used to check the performance of a classification model on a set of test data for which the true values are known. Most performance measures such as precision, and recall are calculated from the confusion matrix. We can observe in Table 8 that the confusion matrix of our test data has the true positive to be 6633 with a small false positive of 1253 and there is a false negative of 376 and a true negative of 737, with a classification accuracy of 0.8188.

Table 8: Confusion Matrix of decision tree on the test data

		Reference	
		Non Default	Default
Prediction	Non Default	6633	1253
	Default	376	737

The classification tree in figure 10 indicates that the variable PAY_0 is the most important variable to predict the default risk because it's the variable used to split the first node in the tree. This classification tree algorithm has an accuracy rate of 0.819 and AUC of 0.6966 which is lower than the GLM. Figure 11 provides the ROC curve of this algorithm.

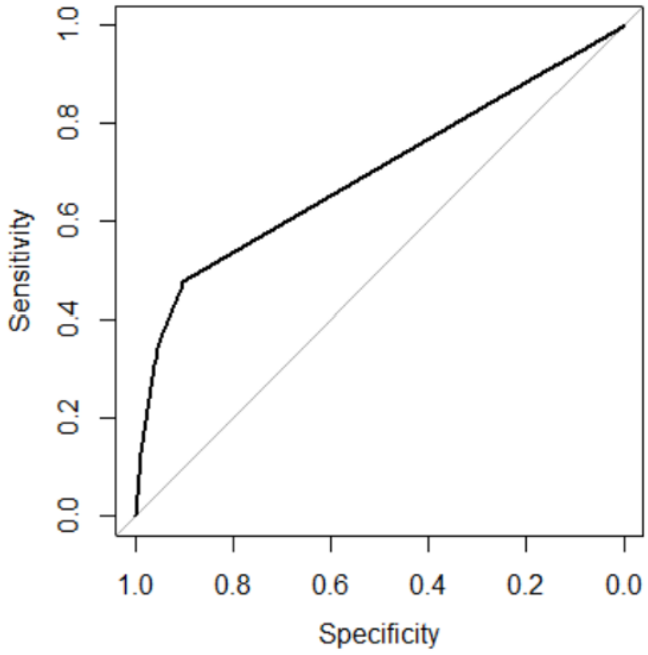


Figure 11: ROC curve for classification tree

4.8. Random Forest

Random forest, as its name implies, consists of many individual decision trees that operate as an ensemble. Each tree in the random forest splits out a class prediction and the class with the most votes becomes our model's final prediction. The random forest can also return the importance of each predictor as shown in Table 9. The results are scaled, so the numbers indicate relative importance. Generally, variables that are used to make the split more frequently and earlier in the trees in the random forest are determined to be more important. According to Table 9, Pay_0 is the most important variable to predict the default payment. This matches the GLM results in Table 7, where the coefficients of the two levels of Pay_0 are highly significant and have large values. The new variable SexEduMarriage that was created by us has an importance of 2.263, which means it was not used frequently in any of the trees. This agrees with the single tree built previously, where the variable SexEduMarriage is not used in any of the split nodes.

Table 9: The relative importance of each predictor variable

	Relative Importance
AGE	392.04
PAY_0	651.78
PAY_2	238.91
PAY_3	189.40
PAY_4	116.50
PAY_5	115.35
PAY_6	77.57
BILL_AMT1	424.41

BILL_AMT2	391.97
BILL_AMT3	376.45
BILL_AMT4	354.90
BILL_AMT5	353.91
BILL_AMT6	356.20
PAY_AMT1	378.61
PAY_AMT2	343.74
PAY_AMT3	330.01
PAY_AMT4	318.30
PAY_AMT5	313.88
PAY_AMT6	328.04
Limit.Balance	371.19
PC1	139.72
PC2	132.30
PC3	149.37
PC4	142.92

The ROC curve of this random forest algorithm is given in figure 12. Its AUC is 0.7639.

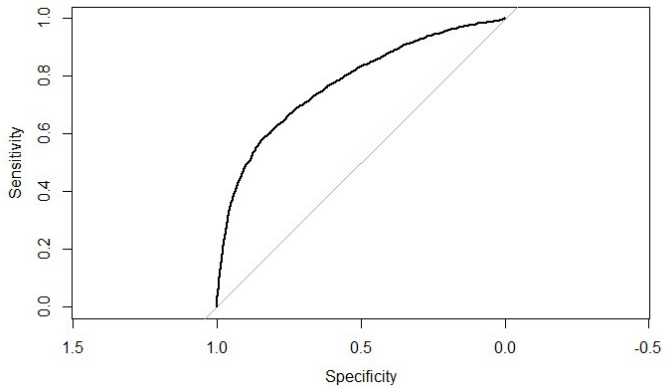


Figure 12: ROC curve for Random Forest

5. Results Comparison and Conclusion

As we mentioned in session 4.5, the G-K-fold stratified cross-validation is used to collect the AUCs and other metrics of each algorithm. We set G=10, K=10. That is in total 100 runs for each algorithm based on different training and testing data partitions. To compare the 3 algorithms, we consider the mean of AUCs, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), run time with the corresponding standard deviation, and 95% confidence interval for each classifier. They are listed in Table 10.1-10.3. The metrics of GLM are significantly better than the decision tree and comparable to the random forest.

All 3 algorithms are implemented in R and run on a MacBook Pro 2018 with 2.2 GHz 6-Core Intel Core i7, 16 GB 2400 MHz DDR4 memory, and macOS Catalina. We found the GLM runs 7.7 times faster than the decision tree, and 45.7 times faster than the random forest. This is also a big advantage of GLM versus the other 2 algorithms. The details of run time are also listed in tables 10.1-10.3.

Table 10.1: Mean of metrics for 3 algorithms

	Mean		
	GLM	Decision Tree	Random Forest
AUC	0.764	0.693	0.762
Accuracy	0.820	0.820	0.817
Sensitivity	0.357	0.350	0.361
Specificity	0.952	0.954	0.946
PPV	0.678	0.683	0.656
NPV	0.839	0.838	0.839
Run time	0.108	0.830	4.931

Table 10.2: Standard deviation of metrics for 3 algorithms

	Standard Deviation		
	GLM	Decision Tree	Random Forest
AUC	0.0051	0.0047	0.0057
Accuracy	0.0027	0.0027	0.0028
Sensitivity	0.0092	0.0191	0.0104
Specificity	0.0022	0.0052	0.0029
PPV	0.0116	0.0165	0.0123
NPV	0.0020	0.0034	0.0021
Run time	0.0245	0.0649	0.2310

Table 10.3: Standard deviation of metrics for 3 algorithms

	95% Confidence Interval		
	GLM	Decision Tree	Random Forest
AUC	[0.756, 0.776]	[0.685, 0.703]	[0.75, 0.77]
Accuracy	[0.814, 0.826]	[0.815, 0.826]	[0.812, 0.822]
Sensitivity	[0.342, 0.376]	[0.312, 0.381]	[0.343, 0.38]
Specificity	[0.948, 0.956]	[0.944, 0.963]	[0.941, 0.952]
PPV	[0.655, 0.7]	[0.65, 0.716]	[0.637, 0.678]
NPV	[0.836, 0.843]	[0.832, 0.844]	[0.836, 0.843]
Run time	[0.088, 0.136]	[0.734, 0.977]	[4.658, 5.682]

DeLong test is used to compare the GLM, decision tree, and random forest classifiers AUCs. We did the DeLong test for GLM vs decision tree with alternative hypnosis "GLM has higher AUCs than decision tree", another DeLong test for GLM vs random forest with alternative hypnosis "GLM has lower AUCs than random forest". The ROC curves for each pair of algorithms obtained in each run of the G-K-fold crossed validation are used as the input for the DeLong test. Therefore, for each pair of algorithms, G×K DeLong test is computed. The p-values of these tests are recorded, and their mean, standard deviation are reported in Table 11. The p-value of the DeLong test of GLM vs decision tree is extremely small, so we should accept its alternative hypothesis that GLM has higher AUCs than the decision tree. For the GLM vs random forest

DeLong test, its p-value is 0.8628, so we should accept its null hypothesis that GLM has higher AUCs than random forest. A boxplot is given in Figure 13 to summarize the comparison of AUCs obtained from the cross-validation for all 3 algorithms. Both DeLong tests and the boxplot show the superiority of GLM vs the other 2 algorithms.

Table 11: DeLong test to compare AUCs.

	GLM vs Decision Tree	GLM vs Random Forest
Alternative hypothesis	GLM has higher AUCs than the decision tree	GLM has lower AUCs than random forest
Mean of p-value	4.71E-38	0.6617
SD of p-value	4.70E-37	0.2530

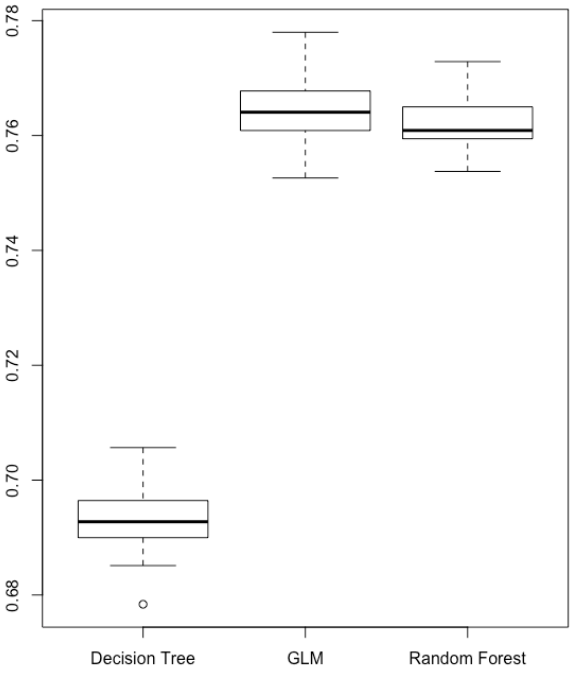


Figure 13: Boxplot to summarize the AUCs comparison.

To discuss the stability of different classifiers, we use the coefficient of variation (CV) of AUC. The CV is defined as the standard deviation divided by the mean. A better classifier is the one with a smaller standard deviation of AUCs and higher average AUCs. The smaller CV of AUC means the algorithm is more stably accurate. We collected the AUCs for 3 algorithms from the G-K-fold cross-validation and calculate the CV of each algorithm's AUCs. The result is listed in Table 12. It shows the 3 algorithms have the similar CV of AUCs.

Table 12: Coefficient of variation (CV) of AUCs.

	GLM	Decision Tree	Random Forest
CV of AUCs	0.00671	0.00676	0.00749

To summarize, the average AUCs for the single decision tree was 0.693 on the test data set, whereas the GLM had 0.764 and the random forest had 0.762. We don't recommend the single decision tree as the final model because of its low AUC. For the remaining two algorithms, the GLM has several advantages over the random forest, one of which is its ease of implementation. The GLM regression formula can be clearly written using an algebra expression. The probability of a default payment can be easily calculated even with simple tools like a calculator or spreadsheet. However, for the random forest, it requires computer programming to obtain the results. Also, the GLM is easier to comprehend and interpret than the random forest, making it easier for credit card companies to grasp and make further business or management decisions upon it. One disadvantage of a GLM over the random forest is that it does not automatically incorporate variable interactions. In GLM, we need first to explore which variables could interact with each other and then manually add those interactions into the regression. The random forest can achieve this automatically. However, the random forest is prone to overfitting, especially when a tree is particularly deep. A small change in the data value could lead to totally different results if using the tree algorithms. Furthermore, the GLM uses fewer features than random forest because of the stepAIC forward selection applied for GLM. Last but not least, the GLM runs much faster than the random forest. Table 13 summarizes the comparison. Based on the above reasons, we conclude GLM model is a better model to predict the credit card default than the decision tree and random forest.

Table 13: Comparison of GLM with other two algorithms on credit card default prediction.

	GLM	**Decision Tree**	**Random Forest**
Average AUC	0.764	0.693	0.762
Interpretability	High	Medium	Low
Implementation	Easiest	Easy	Difficult
Overfitting risk	Low	Medium	High
Algorithm speed	Fast	Slower	Slowest

Conflict of Interest

The author declares no conflict of interest.

Acknowledgment

The author would like to thank the reviews for their valuable time spent on reviewing this paper, and for helpful suggestions to improve this paper.

References

[1] M. Chou, "Cash and credit card crisis in Taiwan." Business weekly: 24-27, 2006.

[2] A. Subasi, S. Cankurt, "Prediction of default payment of credit card clients using Data Mining Techniques," In 2019 International Engineering Conference (IEC), 115-120, IEEE, 2019, doi: 10.1109/IEC47844.2019.8950597.

[3] A. Verikas, Z. Kalsyte, M. Bacauskiene, A. Gelzinis, "Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey," Soft Computing, **14**(9), 995-1010, 2010, doi: 10.1007/s00500-009-0490-5.

[4] F.N. Koutanaei, H. Sajedi, M. Khanbabaei, "A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring," Journal of Retailing and Consumer Services, **27**, 11-23, 2015, doi: 10.1016/j.jretconser.2015.07.003.
[5] M. Berry, G. Linoff, Mastering data mining: The art and science of customer relationship management. New York: John Wiley & Sons, Inc, 2000.
[6] P. Giudici, "Bayesian data mining, with application to benchmarking and credit scoring," Applied Stochastic Models in Business and Industry, **17**(1), 69-81, 2001, doi:,10.1002/asmb.425.
[7] L. C. Thomas, "A survey of credit and behavioral scoring: Forecasting financial risk of lending to consumers," International Journal of Forecasting, **16**, 149–172, 2000, doi: 10.1016/S0169-2070(00)00034-0.
[8] J. N. Crook, D. B. Edelman, L. C. Thomas, "Recent developments in consumer credit risk assessment," European Journal of Operational Research, **183**(3), 1447– 1465, 2007, doi: 10.1016/j.ejor.2006.09.100.
[9] G. Derelioğlu, F. G. Gürgen, "Knowledge discovery using neural approach for SME's credit risk analysis problem in turkey," Expert Systems with Applications, **38**(8), 9313– 9318, 2011, doi: 10.1016/j.eswa.2011.01.012.
[10] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," The journal of finance, **23**(4), 589–609, 1968, doi: 10.2307/2978933.
[11] C. Verma, V. Stoffová, Z. Illés, S. Dahiya, "Binary logistic regression classifying the gender of student towards computer learning in European schools," In The 11th Conference Of Phd Students In Computer Science (p. 45), 2018.
[12] W. Henley, D. J. Hand, "A K-Nearest-Neighbour Classifier for Assessing Consumer Credit Risk," Journal of the Royal Statistical Society: Series D (The Statistician), **45**(1), 77-95, 1996, doi: 10.2307/2348414.
[13] D. West, "Neural network credit scoring models," Computers & Operations Research, **27**(11), 1131–1152, 2000, doi: 10.1016/S0305-0548(99)00149-5.
[14] S. Oreski, G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," Expert systems with applications, **41**(4), 2052-2064, 2014, doi: 10.1016/j.eswa.2013.09.004.
[15] T. Harris, "Credit scoring using the clustered support vector machine," Expert Systems with Applications, **42**(2), 741–750, 2015, doi: 10.1016/j.eswa.2014.08.029.
[16] S. Yang, H. Zhang, "Comparison of several data mining methods in credit card default prediction," Intelligent Information Management, **10**(5), 115-122, 2018, doi: 10.4236/iim.2018.105010.
[17] Y. Yu, "The application of machine learning algorithms in credit card default prediction," In 2020 International Conference on Computing and Data Science (CDS), 212-218, IEEE, 2020, doi: 10.1109/CDS49703.2020.00050.
[18] T. M. Alam, K. Shaukat, I. A. Hameed, S. Luo, M. U. Sarwar, S. Shabbir, J. Li, M. Khushi, "An investigation of credit card default prediction in the imbalanced datasets," IEEE Access, 8, 201173-201198, 2020, doi: 10.1109/ACCESS.2020.3033784.
[19] H. Hoffmann, "Kernel PCA for novelty detection," Pattern recognition, **40**(3), 863-874, 2007, doi: 10.1016/j.patcog.2006.07.009.
[20] V. Kumar, S. Minz. "Feature selection: a literature review," SmartCR, **4**(3), 211-229, 2014, doi: 10.6029/smartcr.2014.03.007.
[21] J. Huang, C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," IEEE Transactions on knowledge and Data Engineering, 17(3), 299-310, 2005, doi: 10.1109/TKDE.2005.50.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 5, 53-61 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

A Proposal of Code Modification Problem for Self-study of Web Client Programming Using JavaScript

Khaing Hsu Wai[1], Nobuo Funabiki[*,1], Khin Thet Mon[1], May Zin Htun[1], San Hay Mar Shwe[1], Htoo Htoo Sandi Kyaw[2], Wen-Chung Kao[3]

[1]*Department of Information and Communication Systems, Okayama University, Okayama, Japan*

[2]*Department of Computer and Information Science, Tokyo University of Agriculture and Technology, Tokyo, Japan*

[3]*Department of Electrical Engineering, National Taiwan Normal University, Taipei, Taiwan*

ARTICLE INFO

Article history:
Received: 16 May, 2022
Accepted: 05 September, 2022
Online: 24 September, 2022

Keywords:
web client programming
JavaScript
HTML
CSS
code modification
coding reading
self-study

ABSTRACT

In current societies, *web application systems* take central roles in computer systems. Thus, *web client programming using JavaScript* has increased values to add dynamic features and functions in web pages by well working with *HTML* and *CSS*. In this paper, as a new type of exercise problem for its self-study, we propose a *code modification problem (CMP)*. In CMP, a source code with *HTML/CSS* elements and *JavaScript* functions for study, and the screenshots of both the original and the slightly altered web pages are provided to the students, who will need to edit the source code to generate the modified page. The goal of CMP is for students to carefully read the source code and comprehend how to use the components and functions through modifying parameters, values, or messages there. *String matching* is used to check the correctness of any answer. Through solving CMP instances, the students are expected to master the basic concepts of *web client programming*. To evaluate the proposal, we generated and assigned 25 CMP instances to 37 students in Okayama University. In addition, we offered project assignments of freely implementing source codes by referring to solved CMP instances to evaluate their learning effects. With the solution results, the validity of the proposal has been confirmed.

1 Introduction

In current societies, computer systems using software and hardware have big impacts on our daily lives. They are crucial to enhancing the capabilities of communications, calculations, and information services. By executing a lot of effective programs that have been created using various programming languages, computer systems are illustrating their powers. In computer systems, several types of programming languages have been used together, depending on the fields and applications of them. Thus, studying the commonly used plural programming languages is essential to develop effective and efficient computer systems for the students in departments of computer science (CS) or information technology (IT) in academic institutions. Especially, web application systems have taken prominent roles in computer systems with expansions of the Internet. For this reason, learning *web client programming with JavaScript* is crucial for students majoring in IT or CS.

The primary scripting language running on web browsers is *JavaScript*, which is sometimes referred as *JS*. *JavaScript* is crucial to modern *web application systems* [1]. Currently, *JavaScript* is adopted in 97% of websites in the world, making it the most popular client-side scripting language for *web client programming*. By combining the common web technology by the *Hyper Text Markup Language (HTML)* and *Cascading Style Sheets (CSS)* , *JavaScript* can provide dynamic features on a *web page*. The structure and meaning of the page are provided in *HTML*. The layout, background colors, and fonts used in *HTML* content are described in *CSS*. Then, the *document object model (DOM)* is used for *JavaScript* interactions with them.

In order to learn a new programming language, novice students should start their study by solving simple and short exercise problems to gain the basic understanding of the grammar and key concepts for *code reading study*. If they cannot read and understand simple source codes, they are unable to properly write them by themselves for *code writing study*. This step-by-step programming learning is important for novice students. After well completing *code reading study*, novice students should move to *code writing study*. To offer step-by-step self-study environments of programming languages to novice students, the project of developing the web-based *programming learning assistant system (PLAS)* has been

[*]Corresponding Author: Nobuo Funabiki, 3-1-1 Tsushimanaka, Okayama University, funabiki@okayama-u.ac.jp

established. Currently, well-known programming languages of C, C++, Java, Python, and JavaScript are supported in PLAS. PLAS offers different types of exercise problems with different learning goals and difficulty levels for the step-by-step studies, as the learning materials or contents on the common system platform.

In PLAS, there exist five types of exercise problems for code reading/writing studies, namely, *grammar-concept understanding problem (GUP)*, *value trace problem (VTP)*, *element fill-in-blank problem (EFP)*, *code completion problem (CCP)*, and *code writing problem (CWP)*.

The *grammar-concept understanding problem (GUP)* [2] offers questions about the source code's keywords by defining terms like reserved words and common libraries in programming languages. The *value trace problem (VTP)* [3], [4] requests the output values of significant variables and output messages in the provided source code. The *element fill-in-blank problem (EFP)* [5], [6] asks the students to complete the missing elements in the provided source code by comprehending its syntax and semantics in order to discover the original source code. The *code completion problem (CCP)* [7], which is intended for debugging study, requests to correct and complete the provided source code when there are blank or incorrect elements in the source code. The *code writing problem (CWP)* [8] requests to complete a new source code from scratch that can pass the given test code. The correctness of the student's answer is automatically checked by *string matching* with the correct one for *GUP*, *VTP*, *EFP*, and *CCP*, and by *software testing* using test codes for *CWP*.

In PLAS, students should first solve the problems related to *code reading study*. They include *GUP* and *VTP* to understand keyword definitions and source code control flows. Next, to learn how to complete source codes by themselves, they should solve *EFP*, *CCP*, and *CWP* in this order. Then, novice students will possibly continue programming studies without dropping out, while solving the problems in PLAS step-by-step, and resulting in improved code reading and writing skills.

However, due to the fact that most web pages are written with the combination of *JavaScript*, *HTML*, and *CSS*, any type of exercise problem currently available in PLAS may not be suitable for studying *web client programming*. After studying each language separately, students must be able to relate them in the source code in *web client programming*.

In this paper, we propose a *code modification problem (CMP)* as the new exercise problem type in PLAS for effective self-study of *web client programming using JavaScript*. Since one web page is usually generated by combining the layout styles of HTML/CSS elements with JavaScript library functions, the effective way to learn them is to read and understand the sample source codes where they are included. A *CMP* instance includes a source code with the HTML/CSS elements and JavaScript functions that should be studied, as well as the two screenshots of the originally produced web page and another modified web page. Then, the original source code is requested to be modified so as to generate the second web page. This CMP is designed for students to carefully study the source code and understand the use of the elements/functions, while changing some parameters, values, or messages. *String matching* is used to check the correctness of any answer. It is expected that they can obtain the basic concepts of *web client programming* through solving the given CMP instances.

To evaluate the proposal, we generated and assigned 25 CMP instances to 37 students in Okayama University. Moreover, we offered project assignments that allow students to implement source codes freely by referring to solved CMP instances to evaluate their learning effects. The validity of the proposal has been confirmed by their solution results.

The outlines of this paper is organized as follows: Section 2 discusses related works in literature. Section 3 proposes the *code modification problem (CMP)*. Section 4 evaluates the analysis results of CMP. Section 5 discusses the learning effects and learning validation after solving the CMP instances. Finally, Section 6 concludes this paper with future works.

2 Related Works in Literature

In this section, we discuss related works to this study in literature.

In [9], the authors proposed an educational game approach called *Reduct* to instruct novice students on fundamental JavaScript programming principles, such as functions, Booleans, equivalence, conditional expressions, and mapping functions onto sets. The designs used theories of progression design and skill learning to scaffold concepts and motivate players to create accurate mental models of the codes. The current objective of this paper is to teach up to the level of advanced and fundamental functional programming in JavaScript.

In [10], the authors described a prototype system to aid students in learning the web language JavaScript. They discussed how portable intelligent exercises activities were implemented and tested them in the web programming course. According to survey assessment data, they demonstrated that the system may assist students in learning of JavaScript, the web-based programming language.

In [11], the authors proposed the *JavaScript development environment (JDE)* for the purpose of supporting programming education. The JDE offers a setting in which JavaScript programming may be done anywhere, at any time. Additionally, the JDE may offer comprehensive snippet features that make it possible to write code using a limited number of actions. It is appropriate for usage on smartphones because it is an environment based on a browser. The JDE can edit HTML, CSS, and JavaScript-enabled web pages.

In [12], the authors presented contributions of an interactive e-learning course for teaching web technologies including JavaScript and HTML as a component of the Moodle platform. This course may have been taught using traditional methods of instruction, or an alternative, utilizing the *Scrum* agile software development methodology and the *EduScrum* teaching methodology.

In [13], the authors proposed an interactive serious programming game for JavaScript programming course at their university. The gamification pattern-based method was used to create this game, together with the *Technology Acceptance Model (TAM)*, and the *Technology-Enhanced Training Effectiveness Model (TETEM)*. Using pre-test and post-test knowledge evaluations, *TAM*, and *TETEM*, they presented the game's evaluation findings.

In [14], the authors made comparative descriptions of several online platforms for teaching programming and chose engaging assignments from the site used to educate students named *hackerrank.com*. They investigated user experiences with *online coding*

platforms (OCP) and contrasted the features of various online platforms that should be utilized to teach programming to aspiring computer scientists and programmers via distance learning. In addition, they also suggested using online programming simulators to enhance computer science instruction, taking into account functionality, as well as students' preparation levels and expected results of learning.

In [15], the authors proposed a game-based learning environment to assist beginner students learning programming. It uses game creation tasks to make basic programming easier to understand and includes idea visualization approaches to let students manipulate game objects to learn important programming concepts.

In [16], in order to investigate students' perceptions of this learning environment, the authors developed a collaborative, learning environment based on the problems powered by the technology for dynamic web. The research was planned as a qualitative study. A interview format that was semi-structured was created to get the opinions of the students about the learning environment, which was supported by technologies for dynamic web and used collaboratively solving issues techniques. Their findings imply that collaborative learning techniques focused on problems and the learning environment at a community college can benefit from technologies for dynamic web pages.

3 Proposal of Code Modification Problem

In this section, the *code modification problem (CMP)* for self-study of *JavaScript* based *web client programming* in PLAS is discussed.

3.1 Definition of Code Modification Problem

In a CMP instance, a source code containing the HTML/CSS elements and the functions to be studied here, and a pair of the screenshot of the original web page generated by the code and the screenshot of the slightly altered web page are provided to students. These web pages could have different parameters, functionalities, or variables. Then, the students are requested to modify the source code to generate the altered web page. By solving the given CMP instances, it is expected that the students can master the basic concepts of *web client programming* and understand the interacted use of HTML, CSS, and JavaScript in the source code. *String matching* is used to check the correctness of any answer in the source code.

CMP is designed with the following goals:

1. Source codes of various kinds are provided along with full forms in order to assist novice students with learning *web client programming using JavaScript*.
2. By answering the questions in CMP instances correctly, students can learn how to generate web pages using various JavaScript functions.
3. As a result of using *string matching*, the student feedback is immediately provided when answers are automatically marked.
4. A novice student who has never studied *web client programming* can answer the questions without encountering serious difficulties.

3.2 Generation Procedure of CMP Instance

The following procedure can be used to generate a new CMP instance:

1. From a website or book that contains the library functions to be examined in this instance, choose a proper source code with HTML, CSS, and JavaScript to develop the web page.
2. Save the screenshot of the web page after running the source code in a web browser.
3. Identify and determine the parts of the source code that students should modify to better understand the intended functionality. Currently, this step is manually carried out. The automatic processing will be studied in future works.
4. Save the screenshot of the altered web page after running the modified source code in a web browser.
5. Due to the fact that HTML tags are not visible on web browsers, replace the HTML tags with the HTML entities according to their numbers and names.
6. Make the input text file for the newly created CMP instance that consists of the problem statement, the original source code, and the modified source code as the correct solution.
7. Using this text file, generate the CMP instance files, which are combined with the HTML, CSS, and JavaScript, for the answer interface on a web browser using the *instance generation program*. It should be noted that this program has already been implemented and used for GUP, VTP, EFP, and CCP.
8. Insert the screenshots of the original and modified web pages into the HTML file for the CMP instance.

3.3 Example of Generating CMP Instance

Next, an example web page source code will be used to discuss the details of the CMP instance generation.

3.3.1 Original Source Code

The original source code should include the fundamental JavaScript library functions for *web client programming* to be studied in this instance. The example source code in Figure 1 shows the web page that displays the *alert box* after clicking *Submit button* from accepting the user name in the input form.

```
01 <html>
02 <body>
03 <p>Submit the Form with alert box</p>
04 <form name="myForm" onsubmit="myFunction()">
05       Enter name: <input type="text" name="fname">
06       <input type="submit" value="Submit">
07 </form>
08 <script>
09 function myFunction() {
10       var x = document.forms["myForm"]["fname"].value;
11       alert("Hi, "+x+" .The form was submitted");
12 }
13 </script>
14 </body>
15 </html>
```

Figure 1: Original source code for CMP instance #19.

3.3.2 *Original Web Page*

The screenshot of the original source code for CMP instance #19 in Figure 2 shows the web page that was created using the source code in Figure 1. This generated web page is offered as a reference for understanding the source code.



Figure 2: Web page by source code for CMP instance #19.

3.3.3 *Modified Source Code*

The original source code can be transformed into the modified source code by changing a few parameters and functions. It is important for students to understand the connections among the corresponding elements of HTML, CSS, and JavaScript when creating web pages. Figure 3 displays the source code for the correct answer. In this example instance, it is necessary to change the text messages in the alert box and the button, as well as the type and the name in the "input" tag.

```
01 <html>
02 <body>
03 <p>Submit the Form with alert box</p>
04 <form name="myForm" onsubmit="myFunction()">
05       Password: <input type="password" name="passwrd">
06       <input type="submit" value="Create Password">
07 </form>
08 <script>
09 function myFunction() {
10       var x = document.forms["myForm"]["passwrd"].value;
11       alert("Your password is created.");
12 }
13 </script>
14 </body>
15 </html>
```

Figure 3: Modified source code for CMP instance #19.

3.3.4 *Modified Web Page*

Figure 4 shows the screenshot of the altered web page by the modified source code.

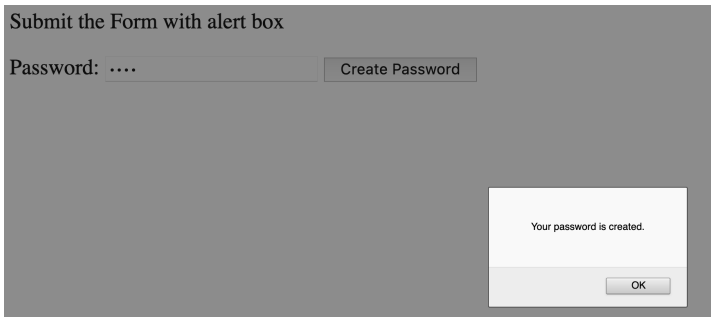


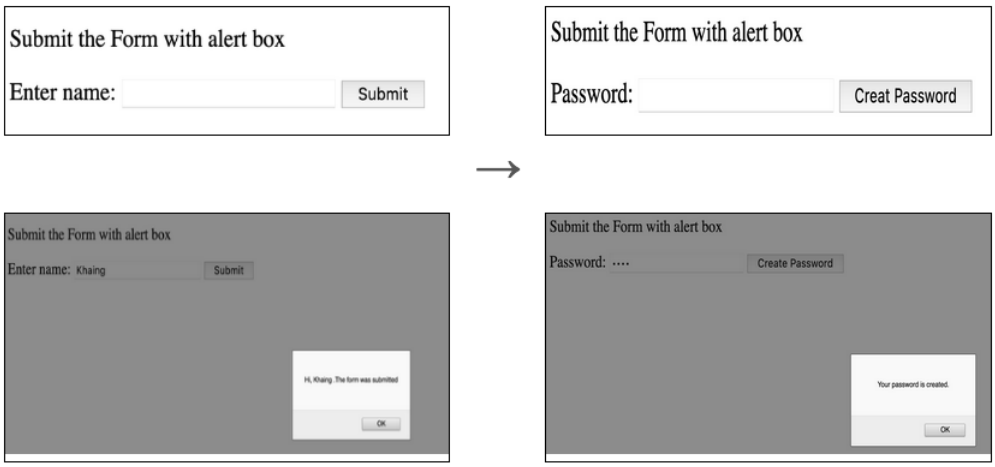
Figure 4: Modified web page for CMP instance #19.

3.4 *Answer Interface*

The answer interface is implemented using HTML, CSS, and JavaScript, where the JavaScript program handles the answer marking for the CMP instance. This allows both the online use and the offline use of the answer interface. In order to prevent students from cheating, the correct answers in any CMP instance are encrypted using *SHA256* [7].

Figure 5 illustrates the response interface for the CMP instance that requests the change of the form for the alert box. The source code lines in the input forms, which are indicated by red backgrounds, must be modified by students to complete this CMP instance. To better understand the source code, students can also download and run the source code, if they want to see how the web page is created and how it works.

After completing all the required changes to the input fields, the student can click the "Answer" button to confirm the validity of the answers. The relevant input form's background color is changed to *red* if the answer on the line is incorrect. Otherwise, it turns *white*. The student can submit their answers again until every answer is verified as correct.



Click Here To Download Source Code

Modify the code to create the password. Use "password" for the type and passwrd" for the name in 'input' tag.

Source Code

```
01:<html>
02:<body>
03:<p>Submit the Form with alert box</p>
04:<form name="myForm" onsubmit="myFunction()">
05:     Enter name: <input type="text" name="fname">
06:     <input type="submit" value="Submit">
07:</form>
08:<script>
09:function myFunction() {
10:     var x = document.forms["myForm"]["fname"].value;
11:     alert("Hi, "+x+" .The form was submitted");
12:}
13:</script>
14:</body>
15:</html>
```

the output

```
01 <html>
02 <body>
03 <p>Submit the Form with alert box</p>
04 <form name="myForm" onsubmit="myFunction()">
05     Enter name: <input type="text" name="fname">
06     <input type="submit" value="Submit">
07 </form>
08 <script>
09 function myFunction() {
10     var x = document.forms["myForm"]["fname"].value;
11     alert("Hi, "+x+" .The form was submitted");
12 }
13 </script>
14 </body>
15 </html>
```

Answer

Answer | File Save

Figure 5: Answer interface for CMP instance #19.

4 Evaluation

This section evaluates the proposed code modification problem (CMP) for *web client programming using JavaScript*.

4.1 Evaluation of CMP Instances

In this evaluation, 25 CMP instances are generated, to examine fundamental concepts and functions in *web client programming* for performing dynamic behaviors of web pages using JavaScript, HTML, and CSS. The topic/function, the number of lines in the source code, and the number of elements that need to be modified for each CMP instance are displayed in Table 1. Other topics like using media devices will be studied in our upcoming works.

4.2 Solution Results

A total of 37 first-year master students in Okayama University, who have not taken any formal course in JavaScript programming, are assigned these 25 CMP instances. Prior to this assignment, we only offered a few websites as references to them.

4.2.1 Results of Individual Students

The results of the 37 individual students are first analyzed in our evaluations. The correct answer percentage (%) and the number of submissions in average among the 25 CMP instances are shown for each student in Figure 6.

According to the results, out of the 37 students, 22 students correctly answered the question and achieved the 100% correct rate, and 10 students scored above 90%. Only five students could not reach 90%. Among these five students, three are considered giving up answering the problems as zero answer submission attempts, and two students had not solved and submitted all the problems. Therefore, in general, the novice students can solve the generated CMP instances through self-study of *web client programming using JavaScript*.

Since each instance requires at least one submission of the answer, 25 is the least number to solve the 25 CMP instances. There are 131.4 submission times on average. These students generally submitted answers $2-14$ times to solve one CMP instance. As a result, they carefully prepared and checked their answers before submitting them.

4.2.2 Results of Individual Instances

Next, we examine the solution results of individual 25 CMP instances. The correct answer percentage (%) and the number of submissions on average are shown for each of the 25 CMP instances among the 37 students in Figure 7.

According to the results, the instance at ID=18 gets the lowest correct rate, which is 90.8%, and the instance at ID=7 achieves the highest correct rate, which is 100%. The maximum number of submissions is 10 for the instances at ID=18 and ID=23 and the minimum ones are 3.3 for the CMP instances of the instances at ID=2, ID=17 and ID=19. Regarding the proportion of the correct answers, 30 elements/functions are needed to be changed in the instance at ID=18. In this case, we can assume that students looked at the screenshots rather than reading the instructions. Some students had not tried to solve this problem from the answer results. In contrast, the source code of the instance at ID=7 was simple for students where they just need to change the colors and the lists.

As for the number of submission, the instances at ID=2, ID=17, and ID=23 are simple to find the necessary modifications. On the other hand, the instances at ID=18 and ID=23 require several modifications. Thus, students submitted their answers 10 times on average. In general, the CMP instances are simple to be solved. However, some instances contain long source codes, where students took time to read and understand them, and made higher submission times.

5 Project Assignment for Learning Effect Evaluation

In this section, we present *project assignments* to evaluate learning effects of the *code modification problem (CMP)* in *JavaScript-based web-client programming*.

5.1 Overview

The proposed CMP is designed for students to read a source code carefully by asking changes of some parameters, values, or messages there, and to understand how to use the HTML/CSS elements and the JavaScript library functions appearing in the code for *web client programming*. However, the CMP does not ask writing source codes by students, which will be a weak point of this approach. Thus, we will evaluate learning effects in code writing to students after solving the CMP instances.

For this purpose, we prepared two *project assignments*, requesting to design web pages and complete the source codes by referring to some given CMP instances. We made the corresponding answer interface to the project assignment that contains the *problem statement* on the requirements, the *input form* for the answer source code, the *web page output area* by running the code, and the *hint* including the sample page layout, the related CMP instances, and the short instruction video. The web page generated by running the code is shown so that students can easily test their source codes. The project assignments were assigned to the same students after solving the 25 CMP instances.

5.2 Project Assignment #1

The first project assignment is *Timer* that requires to implement the text input form, the start, stop, and clear buttons, and the time counting funcion by using the interval function. Figure 8 illustrates the sample web page.

5.2.1 Problem Statement

In this assignment, the problem statement is given by:
"Create and develop simple "Timer" using HTML, CSS and JavaScript. You timer will count the minutes by using *setInterval ( )* function. You can see the hints by clicking the Hint button below. You can also try to write and run codes in the 'Code' area and see how your codes work in the 'Output' area. If your code is OK, you can save your project with your student ID."

Table 1: Generated CMP instances.

ID	topic/function	# of lines		# of modified elements	
		HTML/CSS	JavaScript	HTML/CSS	JavaScript
1	JavaScript object	6	4	1	4
2	JavaScript class	13	10	2	2
3	JavaScript math	7	3	3	1
4	click button	10	3	2	1
5	disable button	12	20	2	3
6	circle drawing	5	14	0	5
7	unordered list	19	7	6	3
8	radio button	11	4	5	1
9	checkbox	8	12	1	1
10	information and color change	15	3	5	1
11	color form type	10	12	2	2
12	element position change	27	10	2	2
13	rotating image	14	7	1	3
14	clickable image map	23	5	3	1
15	image button with counter	22	7	3	3
16	background image position change	21	6	4	2
17	text transformation from text area	37	11	7	1
18	file input type with alert box	25	5	8	1
19	input form with alert box	9	6	2	2
20	numbers addition/subtraction from input form	33	8	3	2
21	progress bar with clicking button	23	23	4	3
22	slider control with range input type	39	9	10	1
23	bar chart drawing from input value	49	14	5	3
24	table row insertion with button	41	9	8	3
25	table column deletion with button	66	16	32	2



Figure 8: Sample *Timer* page for assignment #1.

5.2.2 Hint

In this assignment, the hints are given by:
"For *Timer*, you will need input form to set the minutes and two buttons: start and stop for setting the timer based on the user input value. You can also see Problem #19 and #21 for the reference.
For the *layout interface*, you can create your free design for simple minute count timer. In the following example video, the timer will start as soon as the user put the input minutes value and start the timer button. The timer will be stopped and clear the interval when the user click the stop timer button.

The *setInterval( )* method executes a function or evaluates an expression at specified intervals for *JavaScript library functions* (in milliseconds). Until *clearInterval()* is invoked or the window is closed, the *setInterval( )* method will keep running the procedure. The ID value returned by setInterval( ) is used as the parameter for the *clearInterval( )* method."

5.2.3 Result and Discussion

Among the 37 students, only seven students completed this project assignment by making the source codes satisfying the requirements on the interface and functions. Five students among them achieved the 100% correct rate in solving the 25 CMP instances.

Feedbacks from other students who could not complete this assignment suggested that they found difficulties in combined use of HTML, CSS, and JavaScript, which usually results in long source codes. They also found difficulties in debugging source codes, because the web browser does not show any error message while running them. The use of an advanced IDE returning the error messages should be recommended to edit the source code in *web-client programming*.

Then, it was observed that the web page designs of the submitted source codes are unique from each other. Some pages show the timer with minutes and seconds, while others additionally show hours, and even years, months, and days. It is concluded that these seven students well studied *web-client programming using JavaScript*.

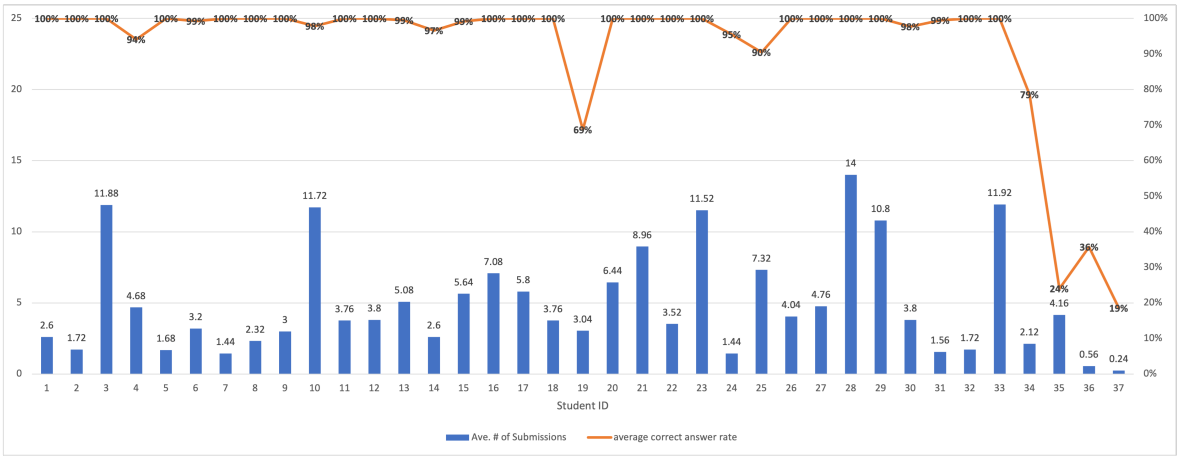


Figure 6: Results for each student.

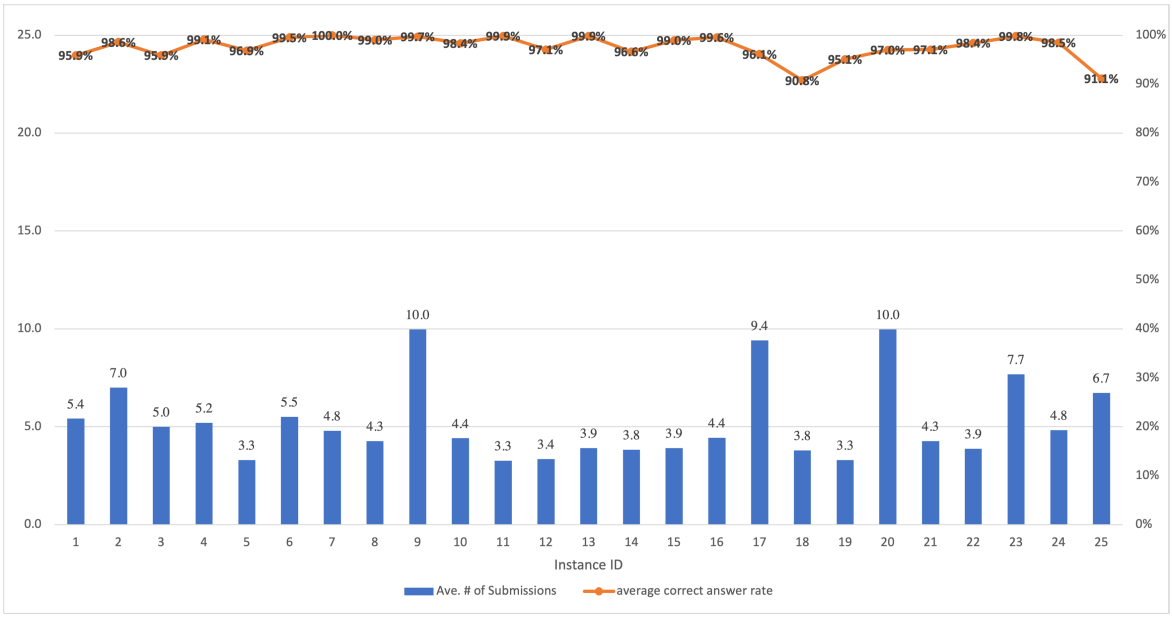


Figure 7: Solution results for individual CMP instances.

5.3 *Project Assignment #2*

The second project assignment is *Calculator* that requires to implement the table layout, the result area, the text input form, the calculation button, and the four arithmetic operations: addition, subtraction, multiplication, and division. Figure 9 illustrates the sample web page. Numbers (0-9), and operators (+, −, ∗, /, %, =) should appear in the table layout and the result area.

5.3.1 *Problem Statement*

In this assignment, the problem statement is given by:
"Create and develop simple "Calculator" using HTML, CSS, and JavaScript. Your calculator may contain numbers (0-9), simple basic calculations (+, −, ∗, /, %, =) and result area to show the calculation answers. You can see the hints by clicking the Hint button below. You can also try to write and run codes in the 'Code' area and see how your codes work in the 'Output' area. If your code is OK, you can save your project with your student ID."

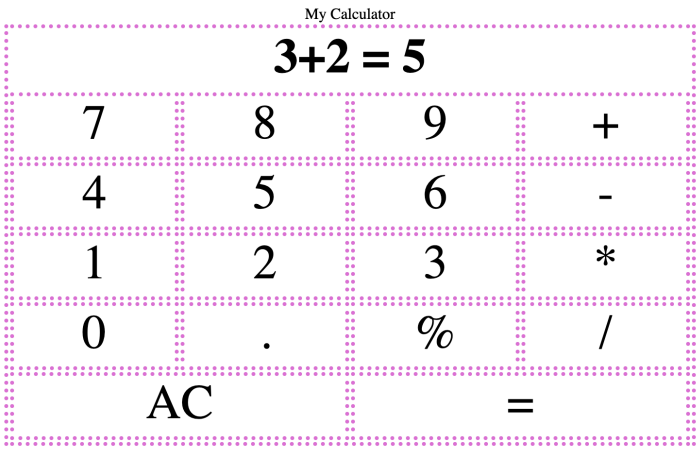


Figure 9: Sample *Calculator* page for assignment #2.

5.3.2 Hint

In this assignment, the hints are given by:
"For *Calculator*, you will need numbers, operators, and display area for the results.
For the *layout interface*, you can create your free design for your calculator. In the following example video, table layout and onclick function is used to get the inputs. You can also see Problem #25 for the reference.
For *JavaScript library functions*, you can use any arithmetic function for calculation in JavaScript. Here, simple *eval( )* function is used in the example. But, using *eval( )* in real-world applications is far more dangerous. We used it here for keeping our project simple."

5.3.3 Result and Discussion

The same seven students completed this project assignment using the table layout design. Their implementations of the arithmetic operations are similar but not the same among them as in the first project assignment.

6 Conclusion

This paper proposed the *code modification problem (CMP)* in *programming learning assistant system (PLAS)* for self-study of *web client programming using JavaScript*. The proposed CMP is designed for students to read a source code carefully by asking changes of parameters, values, or messages there, and to understand how to use the HTML/CSS elements and JavaScript library functions appearing in the code. Students can submit the answers until achieving the correct answers and *string matching* is used to check the correctness of any answer.

For the evaluation of the proposal, 25 CMP instances, which cover the basic functions and topics of *web client programming*, were generated, and 37 first-year master students in Okayama University, who have not taken a formal JavaScript programming course, were assigned to solve them. Based on the results of their solutions, the validity of the proposal has been confirmed. Additionally, two simple project assignments were assigned to the students to evaluate learning effects in code writing abilities.

In future works, we will create CMP instances for web server programming and for other commonly used concepts and functions in web client programming, and provide them to students to evaluate the effectiveness. We will also assist students to complete the project assignments by giving hints on layout designs and code implementations.

References

[1] S.H. Jensen, A. Moller, P. Thiemann, "Type analysis for JavaScript," International Static Analysis Symposium, 238–255, 2009, doi:10.1007/978-3-642-03237-0_17.

[2] S.T. Aung, N. Funabiki, Y.W. Syaifudin, H.H.S. Kyaw, S.L. Aung, N.K. Dim, W.-C. Kao, "A proposal of grammar-concept understanding problem in Java programming learning assistant system," Journal of Advances in Information Technology, **12**(4), 342–350, 2021, doi:10.12720/jait.12.4.342-350.

[3] K. K. Zaw, N. Funabiki, W.-C. Kao, "A proposal of value trace problem for algorithm code reading in Java programming learning assistant system," Information Engineering Express, **1**(3), 9–18, 2015, doi:10.52731/iee.v1.i3.39.

[4] S. H. M. Shwe, N. Funabiki, Y. W. Syaifudin, E. E. Htet, H. H. S. Kyaw, P. P. Tar, N. W. Min, T. Myint, H. A. Thant, W.-C. Kao, "Value trace problems with assisting references for Python programming self-study," International Journal of Web Information Systems, **17**(4), 287–299, 2021, doi:10.1108/IJWIS-03-2021-0025.

[5] N. Funabiki, Tana, K.K. Zaw, N. Ishihara, W.-C. Kao, "A graph-based blank element selection algorithm for fill-in-blank problems in Java programming learning assistant system," IAENG International Journal of Computer Science, **44**(2), 247–260, 2017.

[6] H.H.S. Kyaw, N. Funabiki, S.L. Aung, N.K. Dim, W.-C. Kao, "A study of element fill-in-blank problems for C programming learning assistant system," International Journal of Information and Education Technology, **11**(6), 255–261, 2021, doi:10.18178/ijiet.2021.11.6.1520.

[7] H.H.S. Kyaw, S.S. Wint, N. Funabiki, W.-C. Kao, "A code completion problem in Java programming learning assistant system," IAENG International Journal of Computer Science, **47**(3), 350–359, 2020.

[8] N. Funabiki, Y. Matsushima, T. Nakanishi, N. Amano, "A Java programming learning assistant system using test-driven development method," IAENG International Journal of Computer Science, **40**(1), 38–46, 2013.

[9] I. Arawjo, C.-Y. Wang, A.C. Myers, E. Andersen, F. Guimbretière, "Teaching programming with gamified semantics," in 2017 CHI Conference on Human Factors in Computing, 4911–4923, 2017, doi:10.1145/3025453.3025711.

[10] J. Appleton, "Introducing intelligent exercises to support web application programming students," in 2017 International Conference on Information Communication Technologies in Education (ICICTE), 216–225, 2017.

[11] M. Uehara, "JavaScript development environment for programming education using smartphones," in 2019 International Symposium on Computing and Networking Workshops (CANDARW), 292–297, 2019, doi:10.1109/CANDARW.2019.00058.

[12] P. Vostinar, "Interactive course for JavaScript in LMS Moodle," in 2019 IEEE International Conference on Emerging eLearning Technologies and Applications (ICETA), 810–815, 2019, doi:10.1109/ICETA48886.2019.9039987.

[13] R. Maskeliūnas, A. Kulikajevas, T. Blažauskas, R. Damaševičius, J. Swacha, "An interactive serious mobile game for supporting the learning of programming in JavaScript in the context of eco-friendly city management," Computers, **9**(4), 1–18, 2020, doi:10.3390/computers9040102.

[14] I. S. Zinovieva, V. O. Artemchuk, A. V. Iatsyshyn, O. O. Popov, V. O. Kovach, A. V. Iatsyshyn, Y. O. Romanenko, O. V. Radchenko, "The use of online coding platforms as additional distance tools in programming education," Journal of Physics: Conference Series, **1840**, 2021, doi:10.1088/1742-6596/1840/1/012029.

[15] F. W.-B. Li, C.Watson, "Game-based concept visualization for learning programming," in 2011 International ACM Workshop on Multimedia Technologies for Distance Learning (MTDL), 37–42, 2011, doi:10.1145/2072598.2072607.

[16] E. Ünal, H. Çakir, "Students' views about the problem based collaborative learning environment supported by dynamic web technologies," Malaysian Online Journal of Educational Technology, **5**(2), 1–19, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 6, 58-69 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Field Oriented Control and Commutation Based on Sensorless Methods for High Speed Electrical Motors of Unmanned Multicopters

Chen Zhao*, Weisheng Kong, Federico Percacci, Patrik Gnos

Corporate Center Motion Control, maxon motor ag, Sachseln, 6072, Switzerland

ARTICLE INFO

Article history:
Received: 25 August, 2022
Accepted: 27 October, 2022
Online: 13 November, 2022

Keywords:
Sensorless control
High-speed drive
Field Oriented Control
Permanent magnet synchronous motor
Multicopters and drones

ABSTRACT

In recent years, unmanned aerial vehicles (UAVs), especially small and light multicopters driven by electrical motors and batteries, have experienced a boom in applications. The electrical drive system is a central component of these UAVs. This paper introduces the basics of these drives and presents control methods for them using permanent magnet synchronous motors (PMSMs). Control of these drives is based on field-oriented control (FOC) optimised for high speed (for instance, 200 el. krpm). For the multicopter drives, sensorless control is preferred, i.e., no position or speed sensor on the motors is necessary. Therefore, in this paper, the rotor position is estimated by a sensorless method based on a back electromotive force (back emf) observer combined with a start-up process. The parametrisation methods of the observer and the start-up process are described as well. The observer and the integration of it in multicopter drives are the major innovative parts of this paper. These introduced methods are verified by simulation and experiments. In experiments two motors are considered. One is applied to operate at the maximal speed up to more than 200 el. krpm. The other is a special UAV drive motor and applied for experiments with propeller, under similar operating conditions as UAVs. The results prove the performance and effectiveness of the introduced methods.

1 Introduction

1.1 Multicopters using electric motors

Unmanned aerial vehicles (UAVs), especially multicopters like the one in figure 1, have experienced a surge in popularity in recent years and have found application in various fields. UAVs powered by electrical motors are easy relatively inexpensive to assemble compared to other aircraft structures.

This paper is an extension of work originally presented at the 23rd European Conference on Power Electronics and Applications (EPE21 ECCE Europe), [1]. The technical background and additional experimental results are included, as well as the analysis and discussion of the results[1].

The electric drives and their corresponding control system are one of the core components of UAVs. They are required to be power-efficient, lightweight, robust against a hostile environment, and reliable over many flight hours. Surface mounted permanent magnet synchronous motors (SMPMSMs) are the most suited for UAVs for their compactness, high efficiency, and high power and torque density compared to other motor technologies. While compact multicopters with at least four motors are able to carry a light load, e.g. a camera, large multicopters with six or more motors can carry more equipment and are interesting for tasks such as package delivery in urban areas and the targeted application of fertilisers or pesticides on large farms. To provide the necessary thrust, the SMPMSMs for these large multicopters tend to have a high nominal current and a high number of pole pairs, for instance 20. Since multicopter propellers may be rated for top speeds of even 8000 mechanical revolutions per minute (mech. rpm), the electrical speed experienced by multicopter motors may easily surpass 100'000 electrical rpm (el. rpm).

When it comes to the control of motor current, field-oriented control (FOC) is preferred over e.g. block commutation mainly because of the more even torque generation and the higher power efficiency, especially when used with motors with a long electrical time constant. Either way, the direction of the rotor's magnetic field must be known or measured for the control of motor current. While the most convenient approach consists in measuring the rotor's angle with a position sensor, e.g. an encoder, it is possible

*Corresponding Author: Chen Zhao, Corporate Center Motion Control, maxon motor ag, Brünigstrasse 220, CH-6072 Sachseln, Switzerland, E-Mail: chen.zhao@maxongroup.com

[1]Compared to the conference paper [1], there are following extensions included in this paper: 1) technical background about power electronics and electric motors, 2) technical background about motor drive methods, 3) introduction of modulation methods, 4) the thermal behaviour of motor and motor drive, 5) experimental results of thermal behaviour, 6) additional description of the model and methods, 7) additional results and figures, 8) additional literature papers.

to infer the magnetic field's direction by careful processing of the motor's electrical quantities (voltage and current). This so-called *sensorless* form of position detection is advantageous for UAVs because it renders position sensors superfluous, thus reducing the overall weight and increasing reliability by eliminating a potential source of failure.



Figure 1: A test quadcopter (UAV)

At present, the high electrical speed and high motor current characteristic of large multicopters present a challenge for electrical drives, namely in terms of

- reliable sensorless position detection
- reliable commutation
- reliable current control.

1.2 Related work

There is much research focusing on design and build multicopters, for instance [2]. For a detailed introduction of PMSMs, [3] can be referred.

With consideration of sensorless control of PMSMs, the main approaches for sensorless position detection are two: one is based on the superposition of test signals (*signal injection*) on the voltage vector applied to motor. A number of research papers are published, [4]–[10]. For instance in [4], the Indirect Flux detection by Online Reactance Measurement (INFORM) method is introduced. In [9], the rotor position of SMPMSMs without saliency is estimated by using periodic injection in circumference direction, interleaved with torque generating current. Whereas, the other is based on the estimation of the back electromotive force (back emf).

In this paper, less attention is paid to the signal injection methods, which are computational intensive and may lead to disturbance of motor torque. The Position estimation based on back emf is suitable and sufficient for multicopter drives: as the motors must rotate continuously at high speed during flight, the back emf is large and thus easy to estimate with satisfying accuracy.

The simplest back emf estimation method entails detecting the instants at which back emf changes sign in each motor winding. However, this method cannot be applied in FOC but only in block commutation. A Model-based method using the sliding mode back emf observer is proposed in [11]. It is still only suitable for block commutation. In [12], a load torque estimation method for sensorless control of brushless DC drives is presented, however it is also designed for 6-step commutation.

Further model-based methods compatible with FOC are the object of [9] and [13]. In particular, in [9] a back emf observer combined with signal injection is introduced. The algorithm can be improved and computation load of this method can be reduced for multicopter drives, with consideration of required operation speed. In [13], a sliding mode back emf observer is proposed for sensorless position estimation, in which the position estimation error is compensated by using an integral sliding mode observer. However, for multicopter drives, varying effects have to be considered. Beside the dynamic response and high speed performance, the computation load may also be an issue.

In addition, the performance of sensorless control using pulse width modulation (PWM) and pulse-amplitude modulation (PAM) is analysed in [14]. The focus if this paper is only losses. The control performance is not included, which is very essential for UAV applications. Furthermore, the suggested advantage of PAM is significant at a high speed range compared that of multicopter drives.

1.3 Main contribution of this paper

As aforementioned, sensorless control of PMSMs is considered in a few papers, however, the special requirements and conditions of multicopter drives are taken into account, especially the sensorless drive at high speed range using FOC. With consideration that the multicopters is one of the most interesting application areas of electric drive at present, therefore we propose and describe a control system for electrical drives for drives of multicopters. We achieve sensorless position detection with a Luenberger observer that estimates back emf. We propose a method to parametrise the observer's gains to achieve a desired dynamic response and to guarantee a maximum position estimation error at the maximum operation speed. The position estimate derived from the back emf in this way is updated at the rate of the current control task (25 kHz).

At high electrical speed, the two components of the motor current in the rotor coordinate frame are heavily coupled. While block commutation can obviate this [14], like others we add a decoupling mechanism in the current control loop [15]–[18].

After describing our proposed control system, we evaluate its performance by examining characteristics of interest of a selected multicopter drive (e.g. torque ripple) at its maximum electrical speed, both in simulation and experiment. We show that the multicopter drive with out control system can reach a speed of 200 el. krpm. To the best of our knowledge, other control systems for multicopters with these features do not exist at present.

The rest of the paper is structured as follows. The second section discusses the commutation method and explains the sensorless position detection method and the current controller designed for high speed applications. In the third section, the methods are verified by using simulations and experiments. The last section contains the conclusion.

2 Electrical drive systems

2.1 Power inverter and control

Drive systems using electric motors are very diverse. For PMSMs, the most widely applied motor drive is a voltage source inverter (VSI), [19], which is also applied in this paper. The topology of a VSI is illustrated in the Figure 2, left. The main components are the six metal–oxide–semiconductor field-effect transistors (MOSFET), T. They build three switching bridges for the three motor phases, respectively.

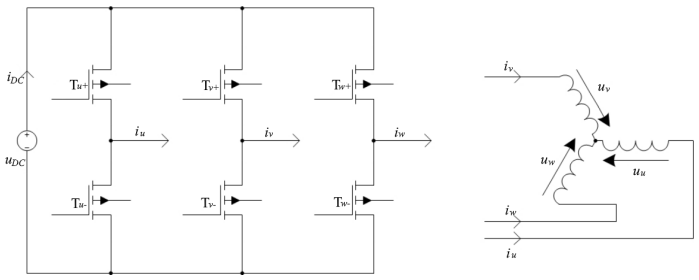


Figure 2: Structure of voltage source inverter and PMSM.

The motor winding terminals are connect to the outputs of the VSI, Figure 2, right, and can be connected to supply voltage or the ground by switching the power transistors in the left figure. In this paper, the switching pattern is determined according to space vector pulse width modulation (SVPWM) methods. Figure 3 shows simulated current in one motor phase and the corresponding PWM duty cycle as an example for generation of current using 5-segment SVPWM. The first plot illustrates the expected reference current. For an easier comparison with the actual current, the reference current is projected to the $u-v-w$ phase coordinate. According to the SVPWM method the duty cycle is calculated and the switching signals are generated, which are the third and fourth plot, respectively. Then, the generated actual current matches the reference current, shown in the second plot, except for the current ripples.

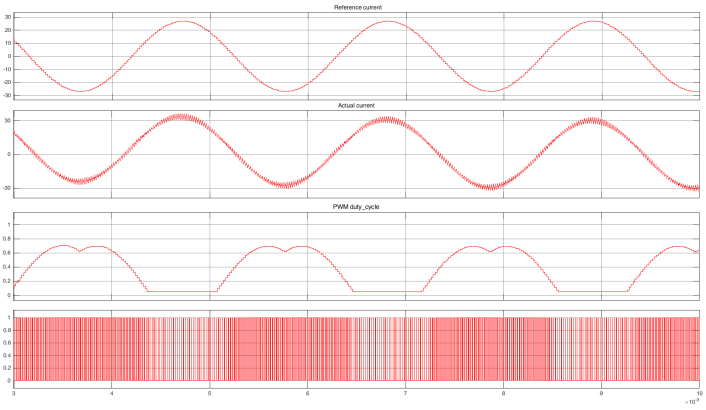


Figure 3: Example of generating current using PWM in one motor phase. The first plot, from above, is the reference current projected in $u-v-w$ phase coordinate, the second is the actual current, the third and fourth are the PWM duty cycle and switching signal, respectively. In the figure of actual current current ripple can be observed, which is caused by PWM switching process.

As mentioned in the introduction, the motor current is regulated with FOC. The principle of FOC is to transform the total motor current flowing in the three phases into two orthogonal components, i.e., the torque generating and magnetic flux generating current by using the Clarke and Park transformations. An electric motor is a complex electrical-magnetic system. This transformations simplify this system significantly and allow the torque and magnetic flux to be controlled separately. Furthermore, the back emf observer mentioned in Section 3 is also implemented according to FOC, so that the estimated position, derived by the observer, can be applied for current control in FOC directly.

The Clarke transformation transforms motor current from phase coordinate, $u-v-w$, to stator coordinate, $\alpha-\beta$, (1). The stator coordinate is fixed with motor stator and the α axis has the same direction as the winding u.

$$\mathbf{i}_{\alpha\beta} = \begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} = \frac{2}{3}\begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix}\begin{bmatrix} i_u \\ i_v \\ i_w \end{bmatrix} \tag{1}$$

The Park transformation transforms current from stator coordinate to rotor coordinate system, $d-q$, (2). The rotor system is fixed to the rotor magnet and rotates with the rotor.

$$i_{dq} = \begin{bmatrix} i_d \\ i_q \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} \tag{2}$$

The aforementioned coordinate systems are shown in Figure 4. The angle, θ, between rotor and stator systems indicates also the electrical rotation angle of rotor, also known as commutation angle.

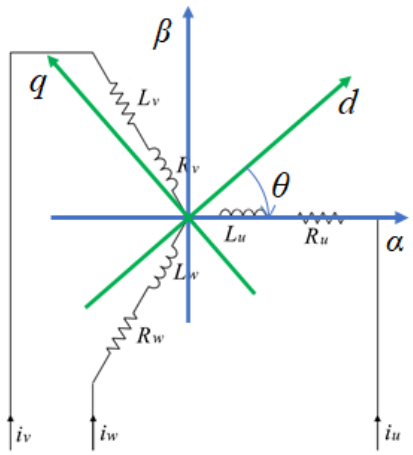


Figure 4: Coordinate systems applied in FOC, including phase, stator and rotor coordinate. θ is the angle between rotor and stator coordinate systems.

In order to control thrust, multicopter drives are operated in speed control mode. Therefore, the control system contains a PI speed controller cascaded with an PI current controller, see Figure 5. The speed feedback are derived by a speed observer or a speed filter. The configurations and gains of both controllers, observer or filter can be determined by using classical methods for electrical drives, for instance plant inversion or pole placement, [20]. As the algorithms and parametrisation methods for PI current and speed controllers are well known, they will not be introduced in this pa-

per. The next subsections describe the features that are specific to operation at high speed.

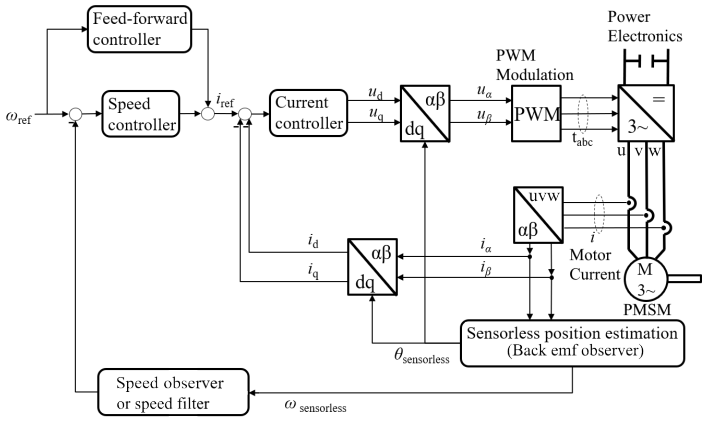


Figure 5: Control structure of multicopter drives, including speed and current control loops.

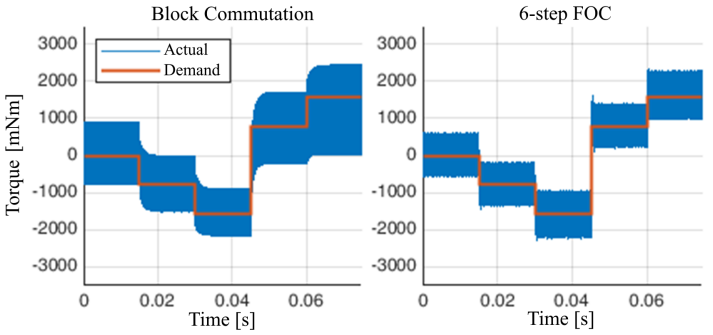


Figure 6: Simulated torque generation of block and 6-step FOC at 100 el. krpm. The FOC has faster response and smaller torque ripple compared to block commutation. The motor data are from PMSM 1 in Table 2 in the "Simulation and experimental results" section.

The output of the current controller, i.e., the demanded voltage, is converted to PWM signals via SVPWM generator. The PWM signals then drive the DC to AC power inverter, which is also included in the top-right corner of Figure 5.

2.2 *Commutation method*

Since multicopter drives are powered by onboard batteries, efficiency is arguably the most important optimisation criterion. In this section, block commutation and FOC are analysed and compared. If the motor speed is high enough relative to the execution frequency of current control, FOC may only generate ten or even less consecutive voltage vectors per electrical revolution, which is comparable to block commutation (six states per electrical revolution). In order to make a fair comparison between the two commutation methods, the position resolution is fixed to 6 per electrical revolution for both commutation methods in the simulation. The motor data from PMSM 1 in Table 2 in the experiment section is used for the simulation. The FOC current controller is explained in the "Current control with decoupling and delay compensation" subsection. Figure 6 and 7 show torque generation performance at two different speeds. Figure 6 demonstrates that FOC has a faster response and lower current (torque) ripple. Besides, FOC is able to generate the demanded torque at a higher speed with the same supply voltage, see Figure 7. Therefore, FOC is chosen.

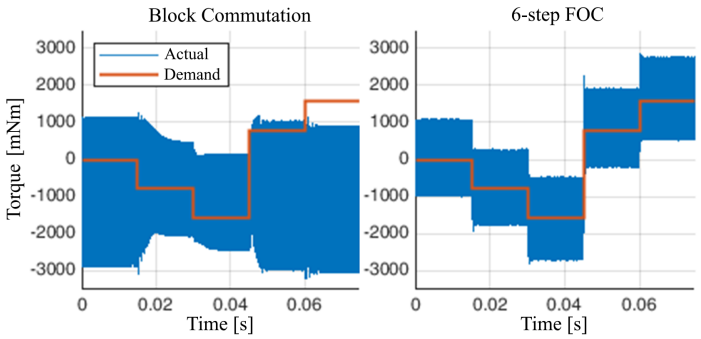


Figure 7: Simulated torque generation of block and 6-step FOC at 150 el. krpm. The FOC can generate the desired torque at 150 el. krpm, whereas block commutation fails. The motor data are from PMSM 1 in Table 2 in the "Simulation and experimental results" section. For block commutation to work at this frequency, a much higher DC supply voltage is needed.

As already mentioned, block commutation has the advantage that the rotor position can be inferred from the instant at which back emf changes sign (i.e., crosses zero) in the disconnected motor phase. However, this advantage is lost for motors with large electrical time constant operating at high speed: under these conditions, the current flowing through the body diodes in the leg of the power stage corresponding to the disconnected motor phase can shift the instant of the back emf zero-crossing.

Furthermore, for FOC, the PWM frequency and modulation methods are important parameters. In this paper, 5-level and 7-level space vector PWM (SVPWM) are considered, as shown in Figure 8. For the PWM frequency, 25 kHz and 50 kHz are considered.

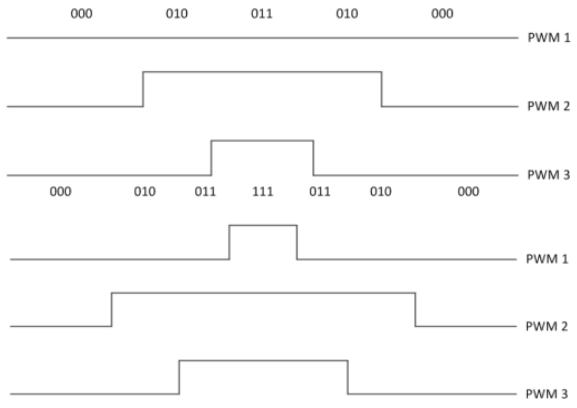


Figure 8: Example of 5-level (upper) and 7-level (lower) PWM duty cycle of a three-phase system. High and low, PWM1, PWM2 and PWM3 are according to the three phases respectively.

7-level SVPWM has lower current ripple, as shown in Figure 9. One should be aware that the difference of current ripple depends on motor characteristics, PWM frequency, and operating conditions. For motors with larger electrical time constant the difference becomes smaller.

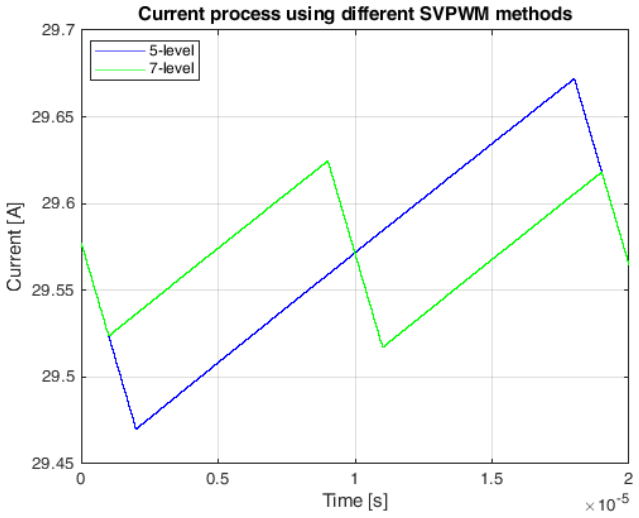


Figure 9: Simulated current process in a single PWM period using different modulation methods, current ripple of 7-level SVPWM is lower than that of 5-level.

On the other side, the switching losses on the motor controller significantly depend on the PWM frequency and modulation method. The comparison results in experiments are included in Section 4.

3 Motion control for high speed applications

3.1 Sensorless position estimation

As mentioned in the introduction, the reliability of the drive system is improved and its weight is reduced by eliminating the position sensor. There are three fundamental requirements on sensorless position detection based on back emf estimation:

1. Stability of the whole control system
2. Accuracy of back emf estimation at low speed, which is essential to use the position estimate for current control
3. Accuracy of position estimation, which affects the efficiency of torque generation.

We choose to estimate back emf with an observer. The observer requires a model of the system. For this purpose, we consider the electrical model of a SMPMSM

$$\mathbf{u}^S = \mathbf{R}^S \cdot \mathbf{i}^S + \mathbf{L}^S \frac{\mathrm{d}\mathbf{i}^S}{\mathrm{d}t} + \mathbf{e}^S \tag{3}$$

where $\mathbf{u}$ and $\mathbf{i}$ are applied voltage and motor current, which are functions of time, $\mathbf{R}$ and $\mathbf{L}$ are the motor resistance and inductance matrices, and $\mathbf{e}$ is the back emf. The superscript S indicates that the quantities are expressed in the stator coordinate system.

As with most control systems in practice, ours is implemented on a microcontroller in discrete form. The discrete implementation of the observer contains two separate steps for prediction and measurement update. The observer differs from a textbook Luenberger observer because the update step is based only on the measurement of current and not of back emf. The prediction (4) and update (6) equations in discrete time are as follows

$$\tilde{\mathbf{i}}_k^S = \mathbf{\Phi}^S \hat{\mathbf{i}}_{k-1}^S + \mathbf{B}_d^S (\mathbf{u}_{k-1}^S - \hat{\mathbf{e}}_{k-1}^S) \tag{4}$$

$$\begin{aligned} \hat{\mathbf{i}}_k^S &= \tilde{\mathbf{i}}_k^S + \mathbf{L}_i (\hat{\mathbf{i}}_k^S - \tilde{\mathbf{i}}_k^S) \\ \hat{\mathbf{e}}_k^S &= \hat{\mathbf{e}}_{k-1}^S + \mathbf{L}_e (\hat{\mathbf{i}}_k^S - \tilde{\mathbf{i}}_k^S). \end{aligned} \tag{5}$$

where subscript k indicates the sampling instant and $\mathbf{\Phi}$ and $\mathbf{B}_d$ are the transition and input matrices of the discrete time system, respectively.

$$\mathbf{\Phi}^S = \exp\left(-\left(\mathbf{L}^S\right)^{-1} \mathbf{R}^S \Delta t\right) \tag{6}$$

$$\mathbf{B}_d^S = \left(\mathbf{I} - \mathbf{\Phi}^S\right)\left(\mathbf{R}^S\right)^{-1} \tag{7}$$

$\mathbf{L}_i = \mathrm{diag}\{l_i\}$ and $\mathbf{L}_e = \mathrm{diag}\{l_e\}$ are the observer feedback gain matrices for current and back emf, respectively. The ˜and ˆ symbols indicate prediction and estimation variables, respectively. In this model, Δt is equal to the sampling time T_s.

In order to determine the values of the observer feedback gains, the electrical control loop, i.e. the linearised innermost feedback loop in Figure 5, is analysed. This loop consists of the current controller, the inverse Park transform (to the right of the "Current controller" block), the motor electrical dynamics, and the back emf observer ("Sensorless position estimation" block). This loop is stable if and only if

$$\gamma = \frac{|\mathbf{u_q}|}{|\mathbf{e}|} < \frac{1 + \phi\,(l_i - 1)}{l_e b_d}. \tag{8}$$

where ϕ and b_d are the diagonal elements of $\mathbf{\Phi}^S$ and $\mathbf{B}_d^S$ in (3). Equation (8) indicates that the operating range of the motor is constrained as a function of the motor's electrical dynamics and the observer gains. The ratio of voltage and back emf in the left hand side of (8) we call *operating point ratio*. Because the constraint also contains the observer gains, the allowed operating range can be manipulated at least to a certain extent.

The feedback gains of the observer are determined by using pole-placement with two conjugate poles. The available measurement is current and the expected output is back emf. For this reason, the transfer function in (9) from applied voltage, $\mathbf{u}$, to back emf, $\mathbf{e}$, has direct meaning in the controller and is applied. This transfer function is derived directly from the observer, according to the observer equations.

$$G(z) = \frac{e(z)}{u(z)} = \frac{l_e z}{z^2 + (l_e - 1 - \phi^S(1 - l_i))z + \phi^S(1 - l_i)} \tag{9}$$

The back emf observer's gains are calculated such that the phase lag of the transfer function $G(z)$ is below an predefined threshold at a certain frequency. This frequency corresponds to the maximum operating speed of the drive system expressed in Hertz. Because the phase lag of $G(z)$ corresponds to the error in the estimation of the back emf angle, and thus rotor position, this criterion guarantees that torque generation efficiency lies above a desired threshold in the whole operating range.

The calculation of the gains entails placing the conjugate poles' frequency a factor ten faster than the maximum intended operating

electrical speed. The corresponding expression for the observer's gains is

$$l_e = 1 - 2e^{-\xi\omega T_s}\cos\left(\omega T_s\sqrt{1-\xi^2}\right) + e^{-2\xi\omega T_s} \quad (10)$$

$$l_i = 1 - \frac{1}{a}\left(e^{-2\xi\omega T_s} + b\gamma l_e\right). \quad (11)$$

where ξ is the conjugate poles' damping coefficient, ω their frequency in rad/s, and T_s the sampling period, which is equal to $40\mu s$ in the experimental system. To sum up, this method requires two inputs to calculate the observer gains

1. Maximum intended operating point ratio for the present gain set
2. Maximum intended operating speed for the present gain set.

With this parametrisation method, the position obtained from the back emf estimate is reliable enough for use in current control from a speed of about 300 to 500 rpm with most motors.

3.2 Operation process

The motor drive operates under two different operating modes, the start-up mode and normal back emf mode, as long as there is not any error in the drive system. The switching between the modes is illustrated in the algorithm below.

Algorithm 1: Operating process and modes

```
Result: Operating process and modes
Initialisation;
while system is enabled and reference speed is suitable do
    if |e_k| > |e_k|_thr then
        Operating in back emf mode;
    else
        Operating in start-up mode;
    end
end
```

We use the magnitude of back emf, $|\mathbf{e_k}|$, as a measure of the reliability of the back emf estimate for the purpose of current control. If it exceeds a predefined threshold, $|\mathbf{e_k}|_{thr}$, the position obtained from the back emf estimate is used for control: we call this operating mode "back emf mode". Compared to the speed range of propellers mentioned in the introduction (2000-8000 rpm), the speed corresponding to this back emf magnitude is low.

Below this speed, the motor operates in what we call the "start-up mode": starting at standstill, a constant-magnitude current vector is applied whose angle rotates with linearly increasing speed, i.e., constant acceleration. This is an open loop control mode, since the actual position and speed are not measured. For propeller drives, the open loop start-up mode and the transition to closed loop back emf mode works reliably, since the load is known and light. In this way, the motor is accelerated until the back emf magnitude is sufficient for closed loop control.

In addition, one must be aware that the position estimated by the back emf observer does not include the polarity of the rotor magnetic field and the rotation direction. To use the position estimate from the back emf observer for commutation, the rotation direction must be known. This is determined by taking the cross-product of the last two consecutive position estimates from the back emf observer. The scalar product of the motor speed with the unit vector z perpendicular to the stator coordinate (α-β) plane yields s_k, which corresponds to the magnitude of rotational speed.

$$s_k = (\mathbf{e}_{k-1} \times \mathbf{e}_k) \cdot \mathbf{z}. \quad (12)$$

In practice, since s_k is very noisy, it is low-pass filtered before evaluation of its sign.

3.3 Current control with decoupling and delay compensation

As mentioned in the introduction, cross-coupling between the current components in the rotor coordinate frame is proportional to the product of the motor's electrical time constant and of the motor's speed. If this cross-coupling is not considered when tuning the current controller, it deteriorates current control performance perceptibly. What's more, the discrete implementation of the current control system inevitably gives rise to sampling delay. To counteract the effect of cross-coupling and sampling delay in the current control loop, we apply d-q decoupling and delay compensation.

The derivation begins from the motor electrical model. Transforming the electrical model of PMSM in (3) to the rotor coordinate frame yields (13). The aforementioned cross-coupling between the d and q components, $\omega_e L$, cannot be neglected anymore

$$\mathbf{u}^R = \begin{bmatrix} R & -\omega_e \cdot L \\ \omega_e \cdot L & R \end{bmatrix}\mathbf{i}^R + \begin{bmatrix} L & 0 \\ 0 & L \end{bmatrix}\frac{\mathrm{d}\mathbf{i}^R}{\mathrm{d}t} + \mathbf{e}^R, \quad (13)$$

where superscript R indicates the rotor coordinate frame and $\mathbf{e}$ is the back emf, whose magnitude is equal to the product of magnetic linkage and electrical speed, $\Psi\omega_e$.

In independent d-q current control, the cross coupling terms are neglected and the current controller is as follows

$$\begin{aligned} \mathbf{u}^R_{c,k+1} &= k_P\Delta\mathbf{i}^R_k + \mathbf{I}^R_k \\ \mathbf{I}^R_{k+1} &= \mathbf{u}^R_{c,k+1} - (k_P - k_I)\Delta\mathbf{i}^R_{k+1}, \end{aligned} \quad (14)$$

where $\mathbf{u}_c$ is the output demand voltage, $\Delta\mathbf{i}$ is the current control error, and $\mathbf{I}$ is the integral term in PI current controller.

When $|\omega_e L|$ is comparable to or larger than R, independent control of the current components of $\mathbf{i}^R$, i.e., i_d and i_q, undermines transient performance and may even lead to instability. It is true that a higher control bandwidth can help. However, in practice the bandwidth is always limited by sampling frequency, which is determined by the computational power of the selected MCU and characteristics of the power electronics components.

We adopt the method in [17] to improve the performance of current control loop without having to increase the sampling frequency. Equation (15) describes the decoupling PI current control in discrete time.

$$\begin{aligned} \mathbf{u}^R_{c,k+1} &= k_P\begin{bmatrix} \cos(T_s\omega_e) & -\sin(T_s\omega_e) \\ \sin(T_s\omega_e) & \cos(T_s\omega_e) \end{bmatrix}\Delta\mathbf{i}^R_k + \mathbf{I}^R_k \\ \mathbf{I}^R_{k+1} &= \mathbf{u}^R_{c,k+1} - (k_P - k_I)\Delta\mathbf{i}^R_{k+1}. \end{aligned} \quad (15)$$

In this current control method, the tuning of current controller should guarantee (16), to ensure the equivalency of discrete time and continuous time controller,

$$\frac{k_I}{k_P} = 1 - e^{-T_s \frac{\hat{R}}{\hat{L}}} \tag{16}$$

where $\hat{R}$ and $\hat{L}$ are the estimated values of phase resistance and inductance, respectively. They can be identified or obtained from the motor's datasheet.

The sampling delay is another factor which affects the high-speed performance. It results in a lower torque than desired causing a worse torque generation efficiency. To address this issue, the demand voltage $\mathbf{u}_c$ is rotated forward by $T_s\omega_e$ yielding $\mathbf{u}_c^*$, which is the final output of current controller

$$\mathbf{u}^*_{c,k+1} = \begin{bmatrix} \cos(T_s\omega_e) & -\sin(T_s\omega_e) \\ \sin(T_s\omega_e) & \cos(T_s\omega_e) \end{bmatrix} \mathbf{u}_{c,k+1}. \tag{17}$$



Figure 11: The test motor PMSM 1, UAV motor. On the back side, left, there are three thick motor power cables and four thin cables for two motor temperature sensors, which are applied for measurement of thermal behaviour later.

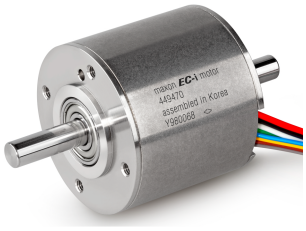


Figure 12: The test motor PMSM 2.

4 Simulation and experimental results

4.1 Simulation and experimental system

The introduced control methods are implemented in an Electronic Speed Controller (ESC), the maxon UAV-ESC 52/30, see Figure 10. The experiments are conducted using this ESC and two test motors, PMSM 1 (Figure 11) is a Ø87 mm outrunner drone motor by maxon designed for multicopters, whereas PMSM 2 (Figure 12) is a general-purpose inrunner motor, also by maxon, which can reach a high electrical speed (210 el. krpm). The MCU of the UAV-ESC 52/30 is an ARM Cortex-M microcontroller from STMicroelectronics. The modelling and simulations are according to the same hardware systems under the Matlab/Simulink environment. Specifications of the controller and test motors are summed up in Table 1 and Table 2, respectively.

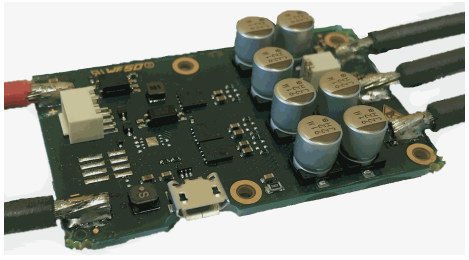


Figure 10: The UAV-ESC 52/30 without housing.

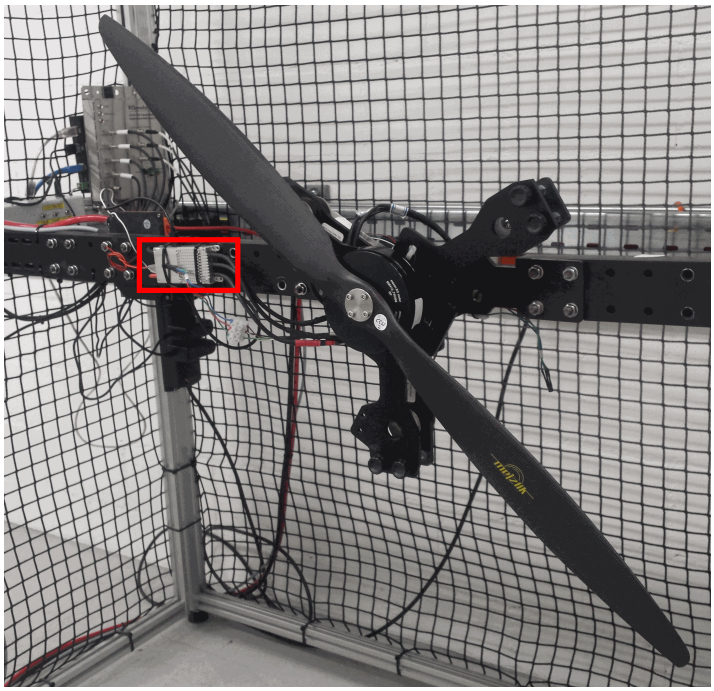


Figure 13: Experimental setup, including the test motor PMSM 1 with propeller and the controller with aluminium housing, indicated by the red rectangle.

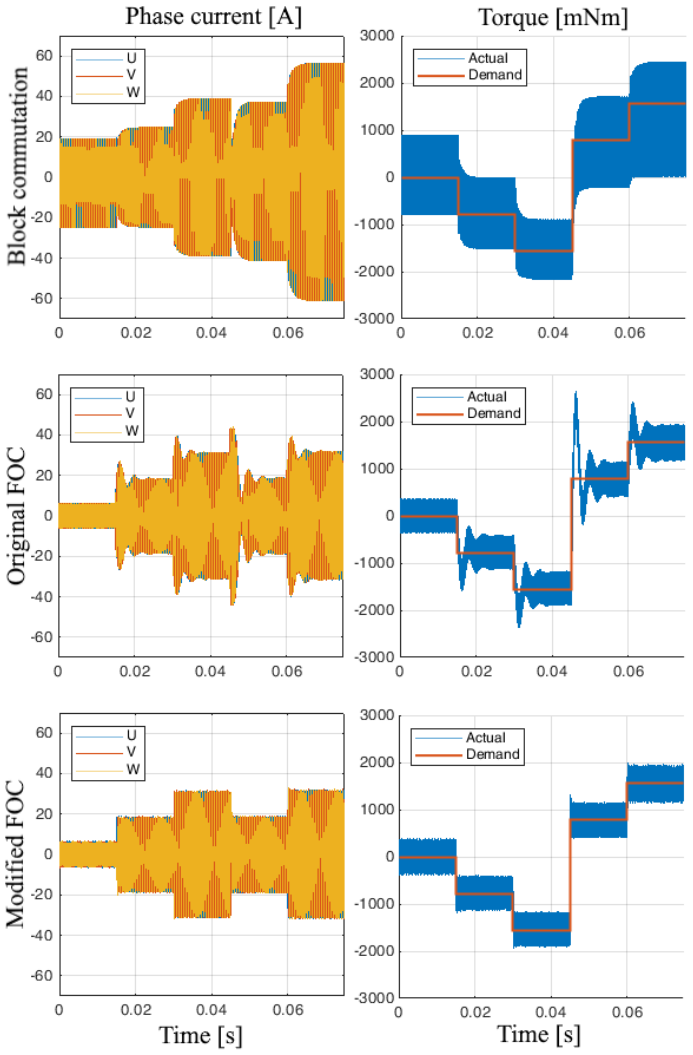


Figure 14: Simulation of motor current and torque during current step response at speed 100 el. krpm, using PMSM 1. The modified FOC has better transient performance and lower current/torque ripple compared to both block commutation and the original FOC.

Table 1: Controller specification

Type of controller	maxon UAV-ESC 52/30
Mechanical dimension:	
length	58 mm
width	36 mm
thickness	11.6 mm
Electric characteristics:	
PWM frequency	25 kHz
Current control sampling frequency	25 kHz
Position estimation frequency	25 kHz
Speed control sampling frequency	2.5 kHz
maximum continuous operating current	30 A
maximum output current	90 A
Supply voltage	12-52 V DC

Table 2: Maxon motor specification

Test motor	PMSM 1	PMSM 2
Diameter	Ø87 mm	Ø40 mm
Length	20 mm	40 mm
Max. continuous operating current	30 A	5.9 A
Max. operating current	90 A	15 A
Motor pole pair number	21	7
Motor maximum mechanical speed	6500 rpm	30000 rpm
Motor resistance (phase to phase)	0.0891 Ω	0.136 Ω
Motor inductance (phase to phase)	0.05 mH	0.0639 mH
Electrical time constant	5.61×10^{-4} s	4.70×10^{-4} s
Structure	Outer runner	Inner runner

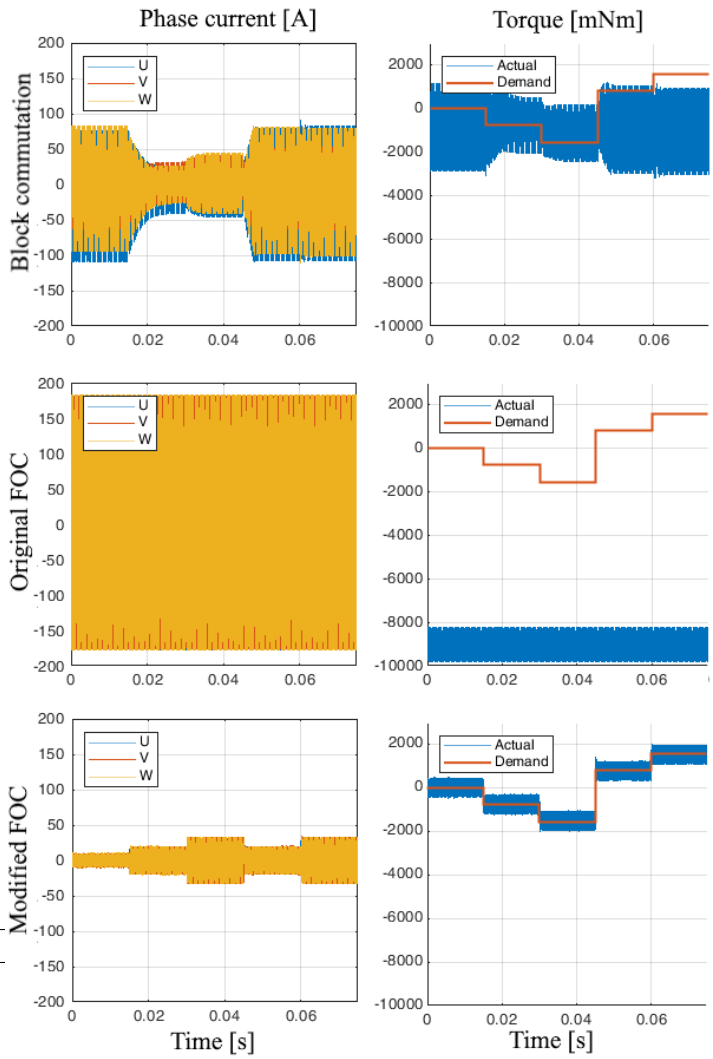


Figure 15: Simulation of motor current and torque during current step response at speed 150 el. krpm, using PMSM 1. The modified FOC performs well, whereas the block commutation cannot output the desired torque, and the original FOC becomes unstable.

The experiments are conducted with different loads. If the test motor PMSM 2 is used, the load is created by a coupled motor generating an opposing torque. On the other hand, the experiments with the test motor PMSM 1 are conducted with a propeller on the

test setup shown in Figure 13. This paper concentrates on motor control.

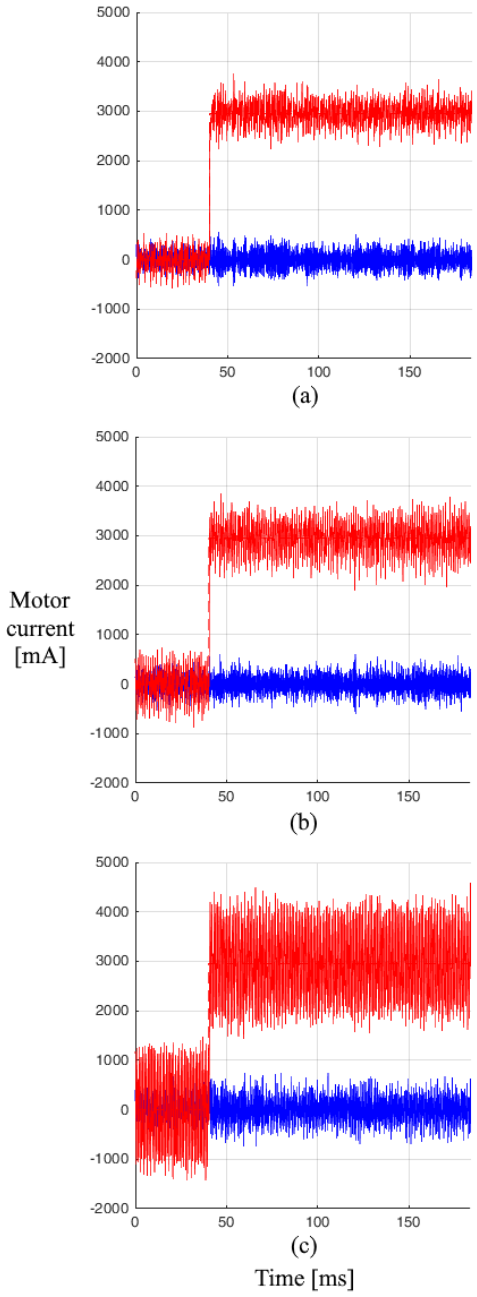


Figure 16: Motor current in experiments of current step response under current control using PMSM 2 at 40, 175, and 210 el. krpm in sub-figures (a), (b) and (c), respectively. The blue curve indicates the d-current and the red curve indicates the q-current. The current controller performs well, with increasing current ripple with operating speed.

4.2 Simulation results of control behaviour

The performance of current/torque control and commutation are simulated for different current step responses. In Figure 14, simulations show that at 100 el. krpm, FOC has a fast response and generally smaller torque/current ripple compared to block commutation. Current control with decoupling and delay compensation ("modified FOC" for short) has an even slightly smaller current and torque ripple compared to the independent d-q current control ("original FOC" for short). In particular, the modified FOC has a significantly better transient performance compared to the original FOC, with consideration of overshoot and settling time.

Figure 15 shows that at 150 el. krpm the original FOC becomes unstable. For block commutation, the DC supply voltage is not enough to correctly apply the desired torque. Instead, the modified FOC still performs well.

4.3 Experimental results of control behaviour

In experiments, different operating points and operating modes are included. Figure 16 shows current in the d-q coordinate frame. There are desired q-current steps at different constant motor speeds (140, 175, and 210 el. krpm), using the modified FOC and PMSM 2. The desired d-current is always zero. At the maximum tested speed, 210 el. krpm, there are 7.5 commutations per electrical revolution. The modified FOC is stable and sufficiently dynamic at all tested speeds, although current ripple increases with operating speed.

In Figure 17, the start-up process using PMSM 1 with propeller is illustrated. The drive starts from standstill in start-up mode. The magnitude of back emf is evaluated in each current period. As soon as the magnitude of back emf exceeds a predefined threshold, 0.35 V in this experiment, the operating mode switches to the back emf operating mode (at about 0.26 seconds in Figure 17). This back emf threshold corresponds to a mechanical speed of about 300 rpm. From then on, the position estimate obtained from the back emf estimate is considered reliable enough for use in closed loop control.

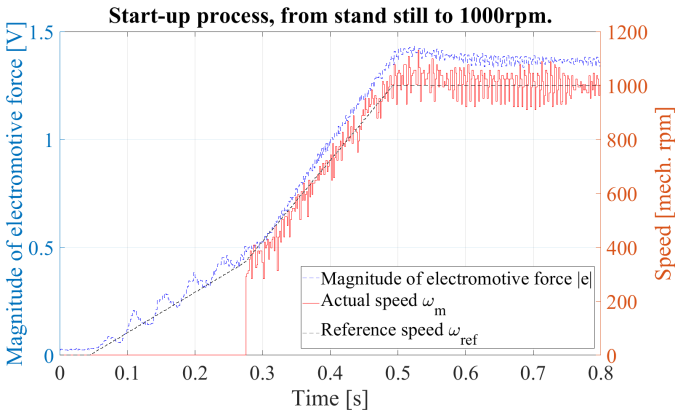


Figure 17: Start-up process from stand still to 1000 rpm , using PMSM 1 with propeller. The black dashed curve and red curve indicate the desired and measured speed, respectively. The speed in the plot is measured based on the back emf position estimation, which is available only after the switch into the back emf operating mode. The blue curve illustrates the magnitude of back emf. In this experiment, the threshold is 0.35 volts.

Further experiments of speed control are conducted. Figure 18 illustrates acceleration under speed control from 600 to 4000 rpm, using the PMSM 1 with propeller. The desired speed is implemented according to a profile acceleration of 5500 rpm/s. The drive operates in back emf mode in this speed range. The speed control and position estimation are reliable.

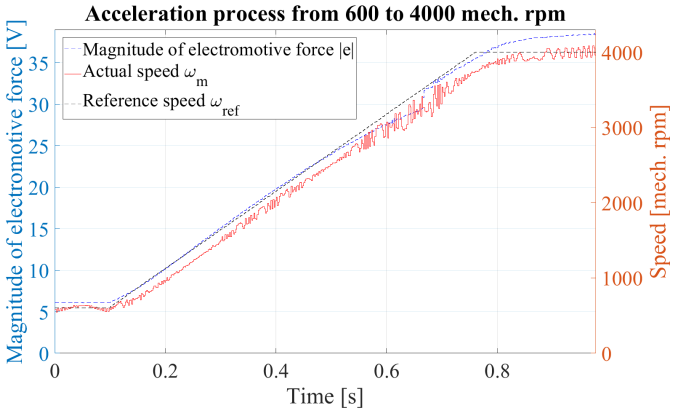


Figure 18: Acceleration using PMSM 1 with propeller, from 600 to 4000 rpm. The black dashed and red curves indicate the desired and measured speed, respectively.

Figure 19 illustrates speed control using fast acceleration from with stand still. The process of motor current is also included. The speed overshoot in this experiment is created with propose in order to illustrate a strong acceleration current, which exceeds the maximal continuous operating current for short period.

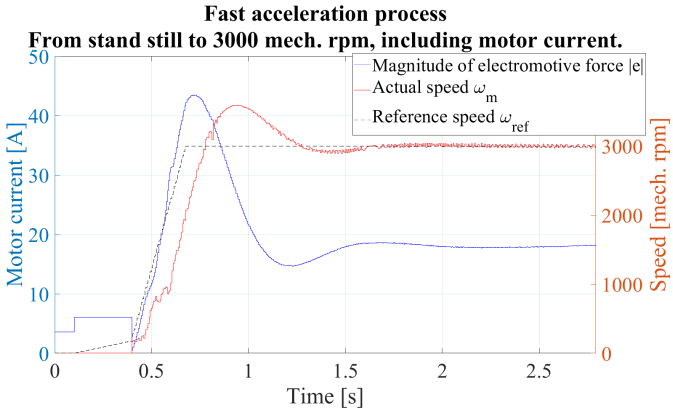


Figure 19: Acceleration using PMSM 1 with propeller, from stand still to 3000 rpm. The black dashed and red curves indicate the desired and measured speed, respectively. The blue curve indicate the motor current.

In the next experiment, true load condition of multicopters is tested. Figure 20 shows the speed control with a constant reference speed, on PMSM 1 with propeller. For propeller drives, the propeller's thrust and drag torque are both monotonic functions with respect to propeller speed. Therefore, at least under laboratory settings, the drive's torque (current) will be a monotonic function of speed. In this experiment, the drive is tested at its maximum continuous operating current, i.e. 30 amperes under similar condition as that of true UAV applications.

To achieve this high fidelity in ground experiment, the motor is under speed control. The load torque is continuously disturbed by an external mechanism on purpose, oscillating between ±10% of the maximum continuous torque. This leads to an effect that the controlled speed oscillates about ±2.5%. The speed controller is not optimised for disturbance rejection. The speed control reference is about 3390 rpm, i.e., about 71.5 el. krpm.

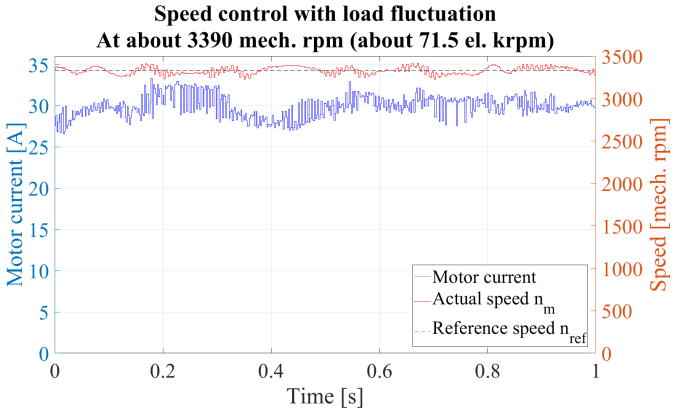


Figure 20: Speed control with maximum output power using PMSM 1. The load torque is continuously disturbed by an external mechanism, between ±10% of the maximum continuous torque. The controlled speed oscillates about ±2.5%. The black dashed and red curves are the desired and measured speed, respectively. The blue curve indicates the motor current.

4.4 *Thermal behaviour*

The thermal behaviour of the drive system is also experimentally investigated. In this experiment, two motors are coupled, one operates as drive (PMSM 1 in Table 2) and the other as load. The temperature of motor and power stage of motor controller is measured by using temperature sensors. A PT100 temperature sensor attached on one MOSFET in the power stage of the UAV-ESC 52/30 as shown in Figure 21. A NTC temperature sensor in the motor, shown in Figure 11, measures the temperature of the motor's windings.

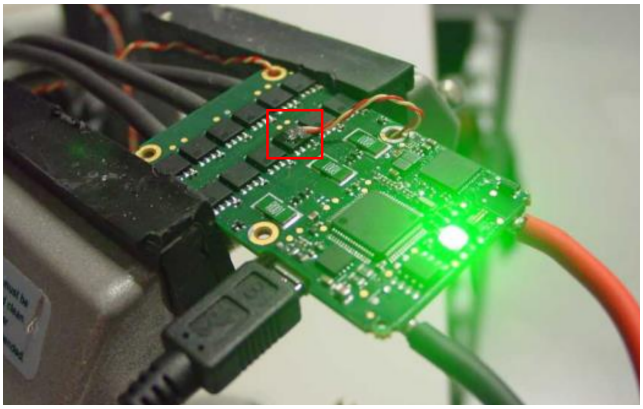


Figure 21: Installation of PT100 temperature sensor on a MOSFET of power stage, the red rectangle. There are totally 12 MOSFETs used in one UAV-ESC 52/30 controller. The red-yellow cables connect the PT100.

The results are summed up in Table 3 and 4, for motor and power stage, respectively, with consideration of using different modulation methods and PWM frequencies. The thermal losses are indicated by using relative temperature. The temperature using 7-level SVPWM with PWM frequency of 25 kHz is considered as reference temperature. The temperature difference is with respect to this reference temperature. It can be recognised that the applied 25 kHz, 7-level SVPWM is optimal for the combination of motor and power stage. According to Table 3, higher PWM frequency leads to lower motor temperature, and the temperature of 7-level SVPWM is lower than

that of 5-level SVPWM using the same PWM frequency. The reason is the smaller current ripple.

Table 3: Thermal losses on motor, indicated by relative temperature difference on motor. The motor temperature using 7-level SVPWM with PWM frequency of 25kHZ is considered as reference.

SVPWM method	f_{PWM}	
	25 kHz	50 kHz
5-level	$3.7°C$	$0.3°C$
7-level	0_{ref}	$-0.8°C$

For motor controller, the relative temperature of power electronics circuit is considered in Table 4. Analogously, the temperature of power stage using 25 kHz 7-level SVPWM is referred. Higher PWM frequency leads to higher power stage temperature, and 5-level SVPWM has lower temperature compared to 7-level SVPWM, due to less switching losses in the MOSFETs.

Table 4: Thermal losses on power stage, indicated by temperature difference of power stage. The power stage temperature using 7-level SVPWM with PWM frequency of 25 kHz is considered as reference.

SVPWM method	f_{PWM}	
	25 kHz	50 kHz
5-level	$-2.8°C$	$6.2°C$
7-level	0_{ref}	$23.2°C$

Comparing both tables, we can sum up:

1. For the tested motor, increasing PWM frequency leads to significantly more losses on power electronics, whereas the reduction of motor losses is minor, because the multicopter motor has relatively large electrical time constant. Therefore, the PWM frequency 25 kHz if preferred.

2. At 25 kHz the 5-level PWM has slightly less losses in the power stage and slightly more losses in the motor compared to 7-level PWM. Because of the better accuracy of the output voltage using 7-level compared to 5-level SVPWM, the former is preferred.

That means the configuration of 7-level SVPWM at 25 kHz, as mentioned in Table 1, is reasonable for the multicopter drive controller.

5 Conclusion

As one of the most rapidly developed application areas of electric motors, UAVs, especially multicopters, using electric motors have gotten a lot of attention. It is still a challenge to fulfil the requirements for professional multicopters, for instance for industry, agriculture and security applications. This paper focuses on these desires. Accurate sensorless control and commutation methods are developed in order to improve efficiency and extend the operation speed range.

In this paper, at first, a motor controller for multicopter drives is introduced. The technical background of electric drive systems for PMSM is introduced briefly, with consideration of power inverter, general control structure, commutation methods, including block commutation and FOC, and different modulation methods of SVPWM. The special requirements and operation conditions of multicopter drives are considered. Then, as the main innovative part of this paper, a control system for electrical drives for multicopters is introduced. The control system relies on sensorless position detection provided by a back emf observer, on a decoupling mechanism of the motor current components in the rotor coordinate frame, and on the compensation of the effect of sampling delay in the current control loop. The parametrisation method of the back emf observer is also included in order to minimise the offset of estimated position, especially at high speed. The proposed methods have taken requirements and application situations of multicopter drives into consideration. The control system is implemented in a compact motor controller. We show in simulation and experiment that the control system can drive two motors at high speed, up to 200 el. krpm. In experiments, a compact motor drive platform, maxon UAV-ESC 52/30, is used. High fidelity validation with propeller and load with disturbance and oscillation are also included in experiments. Furthermore, the thermal behaviour of the multicopter drives is analysed. Measurement of the temperature of both motors and power electronics validate the choice of PWM frequency and space vector modulation method.

Besides multicopters, these methods can also be applied in other systems with similar requirements.

Conflict of Interest The authors declare that there is not any conflict of interest.

Acknowledgement The authors would like to thank their colleagues in various teams of maxon motor ag, specifically, at the Corporate Center Motion Control for their collaboration, the Development Laboratory for access to the measurement installation and the multimedia files, and the Business Unit Aerospace.

References

[1] C. Zhao, W. Kong, F. Percacci, P. Gnos, "Sensorless FOC Drive Control Methods for High Speed Motors of Unmanned Multicopters," in Proc. 23rd Conference on Power Electronics and Applications (EPE'21 ECCE), 2021, doi:10.23919/EPE21ECCEEurope50061.2021.9570534.

[2] N. Michel, P. Wei, Z. Kong, A. K. Sinha, X. Lin, "Modeling and validation of electric multirotor unmanned aerial vehicle system energy dynamics," eTransportation, **12**, 100173, 2022, doi:10.1016/j.etran.2022.100173.

[3] D. Mohanraj, R. Aruldavid, R. Verma, K. Sathiyasekar, A. B. Barnawi, B. Chokkalingam, L. Mihet-Popa, "A Review of BLDC Motor: State of Art, Advanced Control Techniques, and Applications," IEEE Access, **10**, 54833–54869, 2022, doi:10.1109/ACCESS.2022.3175011.

[4] M. Schroedl, "Sensorless control of AC machines at low speed and standstill based on the 'INFORM' method," in Proc. Industry Applications Conference, 270–277, 1996, doi:10.1109/IAS.1996.557028.

[5] A. Consoli, G. Scarcella, A. Testa, "Industry application of zero-speed sensorless control techniques for PM synchronous motors," IEEE Transactions on Industry Applications, **37**(2), 513–521, 2001, doi:10.1109/28.913716.

[6] N. Bianchi, E. Fornasiero, S. Bolognani, "Effect of Stator and Rotor Saturation on Sensorless Rotor Position Detection," IEEE Transactions on Industry Applications, **49**(3), 1333–1342, 2013, doi:10.1109/TIA.2013.2253437.

[7] M. Laumann, C. Weiner, R. Kennel, "Arbitrary injection based sensorless control with a defined high frequency current ripple and reduced current and sound level harmonics," in Proc. IEEE International Symposium on Sensorless Control for Electrical Drives (SLED), 103–108, 2017, doi: 10.1109/SLED.2017.8078438.

[8] T. Slininger, M. Petit, H. Flieh, S. Chien, L. Ku, R. D. Lorenz, "Full Order Discrete-Time Modeling for Accurate and Speed-Independent Pulsating Voltage Injection Self-Sensing," in Proc. IEEE 10th International Symposium on Sensorless Control for Electrical Drives (SLED), 1–6, 2019, doi: 10.1109/SLED.2019.8896346.

[9] C. Zhao, M. Tanaskovic, F. Percacci, S. Mariéthoz, P. Gnos, "Sensorless position estimation for slotless surface mounted permanent magnet synchronous motors in full speed range," IEEE Transactions on Power Electronics, **34**(12), 11566–11579, 2019, doi:10.1109/TPEL.2019.2908408.

[10] K. Schuhmacher, S. Kleen, M. Nienhaus, "Comparison of Anisotropy Signals for Sensorless Control of Star-Connected PMSMs," in Proc. IEEE 10th International Symposium on Sensorless Control for Electrical Drives (SLED), 1–6, 2019, doi:10.1109/SLED.2019.8896351.

[11] H. Fakham, M. Djemai, K. Busawon, "Design and practical implementation of a back-EMF sliding-mode observer for a brushless DC motor," IET Electric Power Applications, **2**, 353–361, 2008, doi:10.1049/iet-epa:20070242.

[12] A. Darba, P. D'haese, F. D. Belie, J.Melkebeek, "Rotor speed, position and load torque estimation using back-EMF sampling for self-sensing brushless DC machine drives," in Proc. IEEE 5th International Symposium on Sensorless Control Elect. Drives, 1–7, 2014, doi:10.1109/SLED.2014.6844968.

[13] Y. Shao, B. Wang, Y. Yu, Q. Dong, M. Tian, D. Xu, "An Integral Sliding Mode Back-EMF Observer for Position-Sensorless Permanent Magnet Synchronous Motor Drives," in Proc. 22nd International Conference on Electrical Machines and Systems (ICEMS), 1–5, 2019, doi:10.1109/ICEMS.2019.8922158.

[14] L. Schwager, A. Tüysüz, C. Zwyssig, J. W. Kolar, "Modeling and Comparison of Machine and Converter Losses for PWM and PAM in High-Speed Drives," IEEE Transactions on Industry Applications, **50**(2), 995–1006, 2014, doi:10.1109/TIA.2013.2272711.

[15] F. Briz, M. W. Degner, R. D. Lorenz, "Analysis and design of current regulators using complex vectors," IEEE Transactions on Industry Applications, **36**(3), 817–825, 2000, doi:10.1109/28.845057.

[16] H. Kim, M. W. Degner, J. M. Guerrero, F. Briz, R. D. Lorenz, "Discrete-time current regulator design for AC machine drives," IEEE Transactions on Industry Applications, **46**(4), 1425–1435, 2010, doi:10.1109/TIA.2010.2049628.

[17] B. Bon-Ho, S.-K. Sul, "A compensation method for time delay of full-digital synchronous frame current regulator of PWM AC drives," IEEE Transactions on Industry Applications, **39**(3), 802–810, 2003, doi:10.1109/TIA.2003.810660.

[18] J.-S. Yim, S.-K. Sul, B.-H. Bae, N. R. Patel, S. Hiti, "Modified current control schemes for high-performance permanent-magnet AC drives with low sampling to operating frequency ratio," IEEE Transactions on Industry Applications, **45**(2), 763–771, 2009, doi:10.1109/TIA.2009.2013600.

[19] D. Schröder, Elektrische Antriebe –RegelungvonAntriebssystemen, Springer-Verlag, 2009.

[20] D. Schröder, Leistungselektronische Schaltungen, Springer-Verlag, 2008.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 4, 59-71 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

μPMU Hardware and Software Design Consideration and Implementation for Distribution Grid Applications

Ahmed Abdelaziz Elsayed*, Mohamed Ahmed Abdellah, Mansour Ahmed Mohamed, Mohamed Abd Elazim Nayel

Faculty of Engineering, Department of Electrical Engineering, Assiut University, Assiut, 71511, Egypt

ARTICLE INFO

Article history:
Received: 05 June, 2022
Accepted: 23 July, 2022
Online: 30 July, 2022

Keywords:
Synchrophasor roadmap
Distribution grid
Real-time monitoring system
µPMU hardware and software design
Error and phasor estimation
Computational time minimization

ABSTRACT

This article presents a roadmap for distribution grid µPMU hardware and software design consideration and implantation to ensure high performance within limited computational time of sampling frequency 512 samples/cycle. A proposed 12 channels, multi-voltage level µPMU hardware and rules of voltage and current transducer, analog filter, analog-to-digital converter, sampling rate definition, and PCB design and selection are presented. From the software view, software minimization procedures are implemented to reduce the estimation time of the proposed µPMU to 18 µsec under high sampling frequency operation. Additionally, error estimation and compensation are used to ensure robust performance, while the computational burden of the error compensation stage is reduced by Taylor series linearization. The proposed µPMU is designed to provide traditional phasor, frequency and harmonics measurements besides a point-on-wave under dynamic operation mode. The proposed device is tested under IEEE Std C37. 118.1 and 118.2 and showed accurate phasor estimation up to 0.03% for the magnitude and angle accuracy up to 0.0036^o*, while the frequency is estimated with maximum variation of 0.032% under dynamic operation.*

1 Introduction

Distribution grids modern structure is more dynamic in nature due to the rapid integration of Distributed Energy Resources (DER) and electric vehicles [1]. In light of the presence of these factors, the traditional uni-directional power flow control and protection infrastructure of the distribution grid required to be integrated with adding smartness features, real-time monitoring, and two-way communications to ensure grid stability. A real-time monitoring synchrophasor measurement is a leading technology to observe the power system with reporting rates up to 2 measurements/cycle and via two-way communications.

Traditional Phasor Measurement Units (PMUs) are suitable for transmission systems application, with Total Vector Error (TVE) up to 1% and angle resolution up to 1°, but could not be applied for distribution grid applications for the following reasons: 1) small distance between buses, 2) high Total Harmonic Distortion (THD), 3) unbalanced conditions, 4) fast load-changing and, 5) large R/X ratio. μPMU is developed as an upgraded version of the PMU with higher sampling frequency and specification up to TVE 0.1% and 0.01^o angle accuracy, respectively [2].

The high cost of the μPMU and the large number of distribution grid buses, which increases the number of installed $\mu PMUs$ to observe the grid, are the main factors constraint the use of synchrophasor technology in the distribution grid applications [3]. However, great investment and consideration are given to the distribution grids due to the rapid and critical changes of their rules over the last years [4]. Motivated to address these challenges, researchers seek to reduce the distribution grid real-time monitoring system cost and enhance the performance of synchrophasor technology. In one track, development and online-learning tools, high-performance frequency and phasor estimation algorithm for low-cost PMUs, and statistical performance and error analysis are studied to allow the researchers to test and improve the synchrophasor technology on one hand, and reduce the hardware specification needed with more accurate phasor estimation abilities on other hand [5]–[6].

In another avenue, the development of low cost μPMU is one of the research tracks to reduce monitoring system costs, by relying on the low Voltage Side (LVS) measurements and downscaling the accuracy and resolution specifications of the designed μPMU [7]–[8]. In [9], a low cost μPMU is proposed by modifying smart meter infrastructure, adding Global Positioning System (GPS) to synchronize the measurements and updating the calculation algorithms to

*Corresponding Author: Ahmed Abdelaziz Elsayed, Contact No +201114040626& Email Ahmed-A.Elaziz@aun.edu.eg

estimate the input signals phasors. The unit used a sampling frequency of 32 samples/cycle and the main measurements unit address magnitude TVE and angle accuracy of 0.2% and 0.3^0, respectively. The work is extended in [10], where the hardware and software are upgraded with a sampling frequency of 64 samples/cycle. The main measurement unit addressed magnitude TVE and frequency estimation accuracy 0.13% and 11mHz, while the angle estimation accuracy was not studied. The low sampling frequency mitigates both units from reaching high angle resolution and only measured harmonics content up to 16-32 order.

In [11], a low-cost PMU is designed with a simple main measurement unit structure comprised of GPS and Analog to Digital Converter (ADC), while the other components are ignored. The unit is designed using a modified Discrete Fourier Transform (DFT) to mitigate the off-nominal operation with a sampling frequency of 256 samples/cycle. The proposed unit is tested for frequency estimation only and showed a maximum variation of 10 mHz. In [8], Field Programmable Gate Array (FPGA) based μPMU using iterative-Interpolated DFT (i-IpDFT) is used to improve the μPMU latency and speed by relying on a parallel operation to estimate the phasor and the frequency within the sampling time of 512 sample/cycle. The μPMU hardware is considered as the GPS, ADC, and microcontroller process where a scaled signal of 1.25 volt is directly injected into the main measurements unit. The i-IpDFT has a TVE of 0.02% compared with DFT under steady-state, but it is faster. The phasor magnitude TVE, angle accuracy, and frequency variation were not tested and only the latency of the device is considered.

From the above literature the following gaps are remarked:

1. Only a few works considered the low cost μPMU design and implementation and studied uncompleted main measurement unit structure comprised of ADC, GPS, and microcontroller without taking into account the overall structure and Voltage and Current Transducers (VT and CT) underperformance, which limits the integration of the previous models in the actual distribution grid.

2. None of the previous works gave a clear roadmap to deciding the μPMU specifications and the main measurement unit design criteria.

3. Error estimation and compensation of each process inside μPMU were not implemented, which limits the previous design to reach the specification of the commercial μPMU [12, 13].

4. Increasing the sampling frequency, on one hand, improves the estimation of the μPMU while on the other hand, it reduces the computational time and limits using complex hardware, which required software to be optimized.

In this article, a proposed roadmap for μPMU design and implementation is discussed. The proposed roadmap describes the μPMU specification selection, structure design, hardware component, and software design to minimize the μPMU computational effort within a sampling time of 512 samples/cycle. In addition, μPMU source of errors are clarified, the impacts of the error on the estimated phasor are explained and the error calibration and compensation to ensure high phasor estimation are mentioned and implemented. The main contribution of this article is to provide a roadmap for the researcher to design and implement μPMU for distribution grid to further improve the distribution grid applications and boost their performance.

The key contribution of this article is summarized in the following points:

1. We propose robust hardware to ensure high performance of the μPMU under steady-state and dynamic operations that provide traditional and synchronized point-on-wave measurements.

2. We propose a μPMU software that has a very small computational time, which estimates all the output information within the sampling time and sends the information per half cycle.

3. We clarify, estimate, and compensate the μPMU error at each stage to ensure high accuracy of the μPMU measurements and ensure light and fast software for the μPMU using Taylor series linearization.

4. We propose a low-cost high-performance μPMU based on Multi-voltage level measurements.

In order to evaluate the performance of the proposed μPMU, the proposed μPMU is tested to check the measurement accuracy and communication performance using IEEE Std C37. 118.1 and 118.2 [14, 15]. Opal-RealTime Simulator-OP4510 is used as a reference and calibration device for the proposed μPMU to evaluate the accuracy and resolution of the estimated phasor and the frequency variation under steady-state and dynamic operation. Three tests are conducted in steady-state, 1) the magnitude changes by 0.1 p.u in the range of 0.1-1 p.u at nominal frequency, 2) the frequency changes by 0.1 Hz in the range of 49.5-50.5 Hz while the magnitude is 1 p.u and 3) the unit is tested to measure a phase difference of 1, 0.1 and 0.01 degrees to find the angle accuracy and resolution. For the dynamic operation, the phasor and frequency estimation performance is tested under 1) ramp frequency change of 0.2 Hz 2) sudden change in magnitude and phase of 0.1 and 10 degrees to estimate the step response of the unit, and 3) harmonics estimation under dynamic operation and point-on-wave recording.

The rest of this article is organized as follows: in Section 2, the μPMU specification selection and the unit structure design are clarified. In Section 3, the μPMU hardware and software design and implementation are discussed, while the error estimation and compensation are presented in Section 4. The test and validation environment of the unit is mentioned in Section 5 and the result and discussion of the proposed μPMU are presented in Section 6. Finally, the article is concluded in Section 7.

2 μPMU Specification Selection and Structure Design

In this section μPMU specifications selection of the main measurement unit is discussed considering accuracy, resolution, reporting rate, and modes of operations. Also, the sampling frequency selection is clarified from the power quality requirements view and the

μPMU structure design, including number of current and voltage channels and their locations, are described.

2.1 μPMU Specification selection

μPMU specifications are selected based on the list of applications to be addressed. Each application has a level of measurements accuracy and resolution limit, TVE, measurements types, and mode of operation (steady-state or dynamic operation mode), which require different performance and reporting rates. Also, other specifications are decided by the designer such as operation range of the voltage and the current, LCD reporting rate and SD card storing rate,...etc. A survey of 75 distribution grid applications is published by Quanta Technology in 2022, which described the existed distribution grid applications, importance, complexity, and the minimum specifications required for each application from the industrial and research perspective [16], which is useful to define the required specifications limit of the μPMU. Four applications are selected to design the proposed μPMU to meet their limits. A7 (Frequency monitoring) and A26 (Distribution state estimation) are selected as steady-state applications. While for dynamic applications, A42 (Faulted circuit identification) and A55 (Islanding detection for distribution generation) are used. Table.1 summarises the specifications of the four applications. Based on these requirements, the designed μPMU has two operation modes (steady-state and dynamic operations), and the accuracy and resolution limits are 1% and 0.1% for magnitude and 0.1^o% and 0.01^o% for angle, respectively with latency below 300 msec. For steady-state, the normal measurements are sent with a reporting rate of 1 measurement/cycle for steady-state application, while for the dynamic mode, the point-on-wave data are also sent to the (Phasor Data concentrates) PDC with a reporting rate of 2 measurement/cycle.

Table 1: List of specifications

Applications		***Steady-state***		***Dynamic***	
		A7	***A26***	***A42***	***A55***
Accuracy	Mag	*1%*	*1%*	*1%*	*1%*
	Ang	-	0.1^o	0.1^o	0.1^o
Resolution	Mag	*0.1%*	*0.1%*	*0.1%*	*0.1%*
	Ang	-	0.01^o	0.01^o	0.01^o
Measurements	**Phasor, Frequency ROFC, Harmonics**			**Point-on-wave**	
Latency	***2000msec***			***300msec***	***500msec***
Reporting rate	1 report/cycle			2 reports/cycle	

2.1.1 μPMU Modes of operation and reporting rate

Modes of operation are designed in the μPMU to adjust its specifications to a certain performance. Different modes required extra measurements such as point-on-wave data and a higher reporting rate for better and faster decisions. Prony analyses are used for the detection of the transient or dynamic events to switch the mode of operation [17].

2.1.2 μPMU Measurements type

$\mu PMU s$ are designed to measure the voltage and current phasors of the fundamental and harmonics, frequency, Rate of Frequency Change (ROFC) and provide the point-on-wave data. The Point-on-wave measurements are provided for dynamic applications to allow off-line investigation of the system state at PDC. Measurements types are identified by the group of applications needed to be addressed by the real-time monitoring system.

2.1.3 μPMU Measurements accuracy and resolution

Accuracy and resolution are indications of the measurement quality and the minimum change that could be detected by the μPMU. Both parameters depend on the sampling frequency, ADC bits number, phasor estimation algorithms, hardware deviation, and error compensation techniques, where the final values are required to meet the application requirements. Since these two parameters depend on different factors, the designed μPMU accuracy and resolution are defined under test. Factors that affect the accuracy and resolution are optimized during the design to increase them within the allowed computational and communication time limits.

2.1.4 μPMU TVE

The TVE is a measure of the overall accuracy reached by the μPMU. It is normalized per unit difference between the ideal samples of measurements and the measured value by the units under the test. Eq.1 describes the mathematical expression of the TVE.

$$TVE = \sqrt{\frac{[(X_r^{'} - X_r)^2 + (X_i^{'} - X_i)^2]}{X_r^2 + X_i^2}} \quad (1)$$

Here, $X_r^{'}$ and $X_i^{'}$ are the real and imaginary parts of estimated measurements by the unit under test.

2.1.5 μPMU Sampling Frequency

μPMU Sampling frequency f_s is one of the main critical factors in the μPMU design. On one hand, high sampling frequency increases the measured phasors, frequency, and harmonics content accuracy and resolution. While on the other hand, high sampling frequency reduces the computational time that existed for calculation, error compensation, and communication, which reduces the ability for stable and reliable operation. From the power quality view, the sampling frequency should be designed to be at least twice the frequency of the maximum harmonic required to be observed to meet the Nyquist theory [18]. This is represented by Eq.2, where h is the order of the maximum harmonic required to be measured and f_o is the nominal frequency of the system.

$$f_s \geq 2h \times f_0 \quad (2)$$

For the distribution grid Low and Medium Voltage sides (LVS and MVS) the range of 35th -50th harmonics content is required to be estimated [19], which required sampling frequency to be at least $f_s \geq 3 \times 50 = 150$ sample/cycle. From the DFT view, $f_s = 256$, 512 and 1024 samples/cycle can be used for the power quality monitoring purpose. For the sampling frequency 256, the estimated phasor

magnitude and frequency showed resolution of 0.2 % and 0.2 Hz using FFT, which did not address the resolution limits required in Table.1. Therefore, f_s=512f_o is selected as the minimum sampling frequency to meet these limits. With a sampling frequency of 512f_o, the accuracy level is improved to 0.1% and the deviation in the estimated frequency is 0.1 Hz. However, working on a sampling frequency 1024f_o gives more accurate results, but it limits the computational time of the μPMU to 19.56/16.276 μsec for a 50/60 Hz system. Designing software to read ADC samples, estimate the phasor, frequency, ROFC, harmonics content, compensate for the error, and sent the information within this limited time is a very complex and requires an expensive set-up. Therefore, the sampling frequency is selected to be 512f_o, which allows a time operation of the μPMU to be 39.06/32.55μsec for a 50/60 Hz system, compared with 416.67/347.22 μsec for the traditional PMU. To meet the time requirements, the designed μPMU hardware and software are required to mitigate the excessive and unnecessary processes that consume high time.

2.2 μPMU structure

The structure of the μPMU is the term that describes the number of the μPMU measurement channels, their operation level, and their location in the power system. These parameters are calculated using Optimal μPMU placement ($O\mu PP$), which estimates the μPMU locations and the maximum number of current channels needed to ensure full observability for the studied grid and location of the voltage and current channels. In [20, 21], μPMU with two current channels one located at MVS and the other at LVS of the distribution transformer load with the voltage channels, is found to be the optimal structure that ensures full observability of IEEE 33 and 69 systems. This configuration minimized the overall cost, considering development and running cost, and the proposed μPMU is designed with the same structure. Each channel is comprised of four ports to measure the three phases and the neutral signal. In this case, 12 signals are measured from the power system and 13 steamers are sent to the PDC, which includes the information of the 12 signals and an extra steamer is sent for unit specifications (IP and time information). Figure.1 shows the proposed μPMU structure.

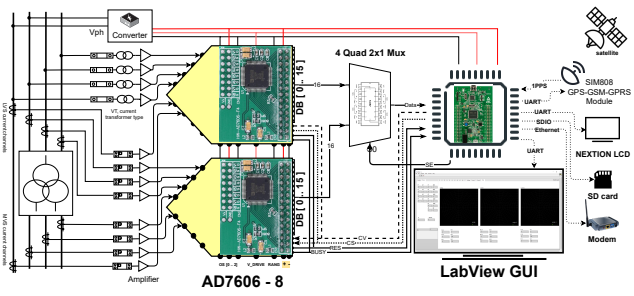


Figure 1: Proposed μPMU structure.

3 μPMU Hardware and Software Design and Implementation

In this section, the μPMU hardware design is clarified to select the appropriate components to meet the computational time and accuracy requirements including VT and CT, ADC, analog filter, GPS, and microcontroller selection. In addition, the μPMU software design to minimize the operation process within the limited time is clarified with time analysis to ensure stable operation.

3.1 Hardware Design

3.1.1 CT and VT selection

The CT and VT are used in μPMU to scale down the signal to the microcontroller level. Due to their critical rules, they are required to be of high accuracy. This is ensured by selecting a high class that meets the accuracy limits and/or robust error estimation and compensation of the CT and VT errors, which is discussed further in section 4. For the current measurements, the CT signals are transferred to a voltage by multiplying the secondary current with the secondary burden impedance. Then the voltage signal from the CT is amplified, so the 120% of the CT rate reflects the full range of the ADC voltage level. Eq.3 shows the current signal amplifier gain value.

$$A_{CT} = \frac{V_{ADC-max}}{1.2Z_B \times I_{CTs}} \tag{3}$$

For the proposed μPMU a CT 200/1 A-11KV, class 0.5 split core is used for the MVS, and for the LVS a CT 800/1 A-400V class 0.5 split core is used. Both CTs use 0.5-ohm pure resistance as burden impedance and the secondary voltage of both CTs have the same range. So, their current channels have the same amplifier design structure in the μPMU main measurement unit. For the voltage measurement, current-type VT is implemented by using isolating CT 1 : 1 ZMPT101B. The input voltage signal is converted to a current signal using primary side limiting resistance that is selected to limit the primary current to 2mA at the maximum input voltage. The secondary current is multiplied by the burden impedance of 0.5-ohm and the secondary voltage signal is amplified to the ADC maximum voltage level. Eq.4 represents the VT amplifier gain.

$$A_{VT} = \frac{R_{P-limit} \times V_{ADC-max}}{V_{in-max} \times Z_{s-B}} \tag{4}$$

This structure provides a bi-polar voltage and current signals to the micro-controller. In regular cases, DC offset is added to shift up the sinusoidal wave above zero. However, to prevent the undefined DC offset and loading effect, which make the calibration of the unit is more difficult and less accurate as in [22], external bipolar ADC is preferred to be used to allow reading both positive and negative parts of the cycle with high accuracy. The amplifier LM324 is used, which is a general-purpose amplifier with high input resistance, linear operation and DC off-set uncertainty of 7mV.

3.1.2 ADC selection and requirements

ADC converts the analog signals of the voltage and current to digital signals to be analyzed by the microcontroller and extracts useful

information from these signals. The ADC should be fast, has a linear operation mode in the operation region, and be of high resolution. Sigma-Delta, Flash, Pipeline, and Successive Approximation Registers are the four types of ADCs that are commonly used. However, SARs are the most used for general purposes due to their high accuracy, low power, zero-cycle latency, and ease to use.

In order to meet all these requirements, AD7606 SAR-type ADC is selected, which has the ability to simultaneous sampling 8-Channels-16 bits of data with a sampling frequency up to $200KHz$. The AD7606 is considered as one of the best selections for Data-Acquisition operation since it can operate with a 5V single supply and can accommodate a true Bipolar analog input $\pm 5V$ and $\pm 10V$, with On-chip 2.5V accurate reference and reference buffer, Analog input clamp protection, Input buffer with 1 $M\Omega$ analog input impedance, Second-order anti-aliasing analog filter that has a 3 dB cutoff frequency of 15 kHz and provides 40 dB anti-alias rejection when sampling at 100 kSPS, a track-and-hold amplifier, Oversampling capability with a pin driven flexible digital filter yields improvements in Signal to Noise Ratio (SNR) to 91.2 dB, reduces the 3 dB bandwidth, and high-speed serial and parallel interfaces to communicate (Serial - Byte - Parallel) [9]. The measured value by the ADC has a resolution of 30 PPM and accuracy of ±12 LSB. Two ADCs are used to measure the 12 μPMU signals using parallel mode operation. Figure.2 shows the timing diagram to read the 12 signals within $2\mu sec$. The following input pins are adjusted to ensure robust sampling with 512 samples/cycle:

1. RANGE: Two range pins are used to control the input operating range of AD7606 (±10,±5) V range. Range Pins can be tied to either supply (±10) or to the ground (±5). In the proposed μPMU the two pins are connected to the ground for (±5) voltage operation.

2. CONVERT: AD7606 has two active high Conversion start pins (A, B), one pin for start every four channels. Both the conversion pins can be tied together and controlled by a single PWM signal from the Micro Controller.

3. Chip select(CS): Active low input pin, which is set to low whenever the host controller wants to perform sampling and read data.

4. READ (RD/SCLK): Active low read signal is used during Parallel operation and for each pulse it will clock all the channel data out.

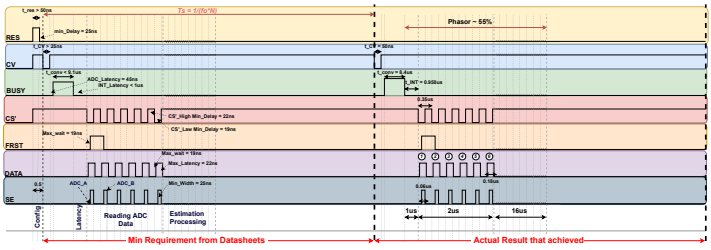


Figure 2: ADC input pins timing diagram.

3.1.3 *Analog filters design*

An Analog filter is used to filter the scaled signal from the CT and VT. Two low pass filters are used. The first filter removes harmonics content with a frequency above the half sampling frequency ($45 - 55\% f_s$). The second is the surge filter, which removes the transient pulses due to switching events. A second-order Butter-Worth low pass filter is designed for this purpose. For the designed μPMU the cuff-off frequency of the filter is required to be in the range of 11-15 kHz, while the surge filter is designed with 100KHz. This function is made with the built-in analog filter of the selected AD7606 with a suitable cut-off frequency for this purpose. The transfer function of the filter is estimated from the data-sheet or estimating the filter frequency gain and phase response in a lab experiment.

3.1.4 *Micro-controller selection*

The microcontroller is the brain of the μPMU, which organizes and controls all the hardware systems to operate correctly, calculate and send the output information. The selected microcontroller should be capable to address the minimum hardware requirements, which allows correct ADC communication, accurate and minimum time of voltage and current phasor estimation using DFT, frequency estimation and ROFC, error compensation, communication with other parts, and send the data to the PDC. Based on these requirements, STM32F407 ARM Cortex M4 ultra-high performance microcontroller is selected. STM32F407 features addressed all the minimum requirements required by the proposed μPMU specification which are:

1. It provides a simultaneous conversion process.

2. Samples are stored in the RAM and shared with a communication buffer through the Direct Memory Access (DMA). In the realized tests, the platform demonstrated enough processing capacity for the calculus and delivery of packages to the PDC.

3. Besides that, it has an Ethernet interface, which is faster than the serial peripheral interface (SPI) and allows the data packages to transfer by directly accessing the memory via DMA without occupying the processor.

4. STM32F407 is based on the high-performance Arm® Cortex®-M4 32-bit RISC core operating at a frequency of up to 168 MHz. The Cortex-M4 core features a Floating point unit (FPU) single precision which supports all Arm single-precision data-processing instructions and data types. It also implements a full set of DSP instructions and a memory protection unit (MPU) which enhances application security.

3.1.5 *GPS synchronization signal and 4G communication*

The μPMU measurements are synchronized with high accuracy clock from the GPS, which consists of 24 satellites in six orbits and provides a Pulse Per Second (PPS) to all GPS modules. The PPS is used in the μPMU to enable the PWM that controls the ADC conversion process. The GPS module is selected to ensure low Synchronization Time Uncertainty (STU) to reduce the synchronization error. GPS SIM808 module is used in the proposed

μPMU. The module is a GSM, GPRS, and GPS three-in-one functions module, which uses the latest GSM/GPS module SIM808 from SIMCOM, supports GSM/GPRS Quad-Band network, combines GPS technology for satellite navigation, and provides communications abilities using 4G. The GPS module has an STU of equal to 10nsec and uses different GPS and 4G antennas to mitigate the excessive communication and overlap problems.

3.1.6 μPMU Additional parts

In order to allow the useful function of the proposed μPMU, powering circuits and accessories are added to the main measurements units. First, AC/DC center-tapped pulse transformer is used to power the circuit with ±5 V. Second, the Nextion LCD touch screen is used to display the output information to the user every second. In addition, built-in storage Ultra-high-speed SD card is added to record the information with communication protocol SDIO for a period of up to 48 hours in the dynamic model. To allow the user access to the μPMU output information from the PC use, A Graphic User Interface (GUI) is designed using LabVIEW to display, save and export the output information. All hardware components are placed on a Printed Circuit Board (PCB).

3.2 Software Design

The main challenges of the designed software are to meet the time requirements of 39.06/32.55 μsec to read the 12 signals from the ADC, update the 12 phasors, and estimate the frequency and ROFC, harmonics content estimation, digital filtering, and error compensation for each updated measurements. Keeping in mind the communication time to not exceed the half-cycle.

3.2.1 Phasor estimation

Phasors play a leading role in the measurement technologies of power systems. They represent the magnitude and phase angle of the fundamental component of the voltage and current in complex representation refer to a specific time reference. DFT is used to estimate the phasor using measured ADC samples. Non-Recursive DFT estimates the phasors with previous without looking to the last estimated phasors, where DFT is applied to all 512 samples. This process required 512×2 product and 512 summation operation. As a result, it consumes a very large time and failed to be implemented in the proposed software structure. Recursive DFT, on the other hand, updates the new phasor based on the previous phasor estimated. This process required two multiplication and one summation process, which reduced the phasors estimation time for the proposed μPMU [23]. The estimated phasor using the Recursive method is defined by Eq. 5.

$$X^h_{N+k} = X^h_{N+k-1} + ((x_{(N+k)} - x_{(k)}) \times [C^h_R(k) - jS^h_I(k)] \quad (5)$$

$$C^h_R = \frac{\sqrt{2}}{N}\begin{bmatrix} cos(\frac{2h\pi 0}{N}) \\ ... \\ cos(\frac{2h\pi(N-1)}{N}) \end{bmatrix}, S^h_I = \frac{\sqrt{2}}{N}\begin{bmatrix} sin(\frac{2h\pi 0}{N}) \\ ... \\ sin(\frac{2h\pi(N-1)}{N}) \end{bmatrix}$$

The DFT coefficient matrices C^h_R and S^h_I are precalculated and stored in the microcontroller to reduce the computational time of the DFT algorithms. It is important to note that the algorithms calculate $| X^h_{N+k} |^2$ to avoid the square process which consumes huge time. The angle of the phasor is measured using the linearization of tan^{-1} by the Tayler series described in [24]. Eq. (6) shows the phasor angle estimation using the linearization function Γ.

$$\theta_{N+K} = \Gamma[Imag(X^h_{N+k})/Re(X^h_{N+k})] \quad (6)$$

The 12 phasors estimation time is 14 μsec with these modified steps.

3.2.2 Digital filter (Simple Average Filter)

In order to compensate for the error due to leakage phenomena under the off-nominal operation, a Simple Average filter is used for both measured magnitude and angle with a window of half cycles. Therefore, the magnitude and phase of the last $N/2$ estimated phasor are stored in the microcontroller and the average value is calculated using Eq.(7).

$$| \overline{X}_{N+K} |^2 = \frac{2}{N} \sum_{K+1+\frac{N}{2}}^{K+N} | X(i) |^2, \overline{\theta}_{N+K} = \frac{2}{N} \sum_{K+1+\frac{N}{2}}^{K+N} \theta(i) \quad (7)$$

Here, $| \overline{X} |^2$ and $\overline{\theta}$ are the averaged values of both magnitude power 2 and the phasor angle.

3.2.3 Frequency estimation

In order to estimate the frequency and ROFC, the angle-based method is used. In this method, the change in the phasor angle is assumed to be a quadratic time function [13, 14]. The phasor angle is stored for the last N estimated phasors and frequency defined by Eq.(8)

$$f = f_o + \triangle f + \frac{df}{dt}t = f_o + \frac{a_1}{2\pi} + \frac{a_2}{\pi}t \quad (8)$$

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} N\Sigma b\Sigma b^{\cdot 2} \\ \Sigma b\Sigma b^{\cdot 2}\Sigma b^{\cdot 3} \\ \Sigma b^{\cdot 2}\Sigma b^{\cdot 3}\Sigma b^{\cdot 4} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ b \\ b^{\cdot 2} \end{bmatrix} \begin{bmatrix} \theta(0) \\ ... \\ \theta(N-1) \end{bmatrix}$$

Here, b is a vector of size N, which represents the corresponding time of each sample $b(n) = n/f_s$.In order to reduce the storage information inside the microcontroller, the b vector is precalculated and only this vector is stored in the memory. These modifications allow estimating the frequency within 2 μsec.

3.2.4 Communication Protocol

Inside the μPMU UART is used as a common communication bus to share the information between GPS, microcontroller, LCD and the directly connected PC GUI. For data export and communication with external PDC, the μPMU follows the IEEE Std C37.118.2 communication protocol to generate the data frames, uses CRC method to check this data and collect the frames in packets to send the information [15].

4 μPMU Error Estimation and Calibration

Errors are associated with the estimated phasors in each process starting from the CT and VT, analog filter, ADC sampling and digitization process, GPS synchronization, and off-nominal operation error. The output phasor calculated in this case is represented by Eq.(9)

$$\vec{X} = \vec{E}_{CT/VT}\vec{E}_{AF}\vec{E}_{ADC}\vec{E}_{GPS}\vec{P}\vec{X}^{act} \quad (9)$$

where $\vec{X}^{act}$ is the actual phasor without error. $\vec{E}_{CT/VT}, \vec{E}_{AF}, \vec{E}_{ADC}, \vec{E}_{GPS}$ and $\vec{P}$ are the complex gain error due to CT/VT scaling, analog filter error, ADC digitization error, GPS synchronization error and off-nominal complex gain P error, respectively. Each error has it is own nature and is required to be compensated, if possible, to ensure high estimation quality [13]. $\vec{E}_{ADC}$ and $\vec{E}_{GPS}$ depend on a random process and these errors are just minimized by the appropriate selection of the ADC and GPS.

4.1 CT and VT Error estimation and Compensation

CT and VT errors are external errors added to the main measurement unit of the μPMU. In traditional PMUs, CT and VT errors are small and negligible compared with the accuracy and resolution limits required by their applications. In contrast, CT and VT errors are higher in distribution since it is noisier. Also, the accuracy and resolution limits are higher in the distribution grid and transducer errors are required to be compensated to meet these limits. In this article, the CT error is the main focus, since both current and voltage measurements rely on CTs. To ensure high-quality measurements, 1) the selected CT should be of a class higher than the accuracy limits, 2) the CT error should be estimated and compensated. Two methods are existed to compensate for the CT error. In the first method, the CT is loaded under different conditions of pure and distorted input current to provide a set of experimental data. These data are used to rather define the empirical calibration equation as in [22] or learned by using an artificial neural network to compensate for the error as in [25]. In the second method, accurate modeling of the CT is used to estimate the primary current based on the CT secondary current waveform, which allows for estimating the complex gain error for the CT at different operation conditions. To achieve this, the Preisach model is used to simulate the CT B-H curve, which can represent the CT major and minor loop under different operating conditions as in [26]. For the proposed μPMU, the empirical equation is used to compensate for the CT error under steady-state. For the dynamic operation, the Discreet Preisach model is used, based on the point-on-Wave measurements provided to evaluate the CT performance under unpredicted waveform [27]. The CT and VT complex gain error estimated during the calibration is represented in Eq.(10).

$$\vec{E}_{CT/VT} = \vec{I_s}/\vec{I_p}\ \ p.u \quad (10)$$

4.2 Analog Filter Error

The analog filter error is represented by a lag phase shift and reduction in the magnitude. In order to estimate the analog filter error, the transfer function of the filter is estimated from the filter frequency and phase response in the experimental lab or from the datasheet information. Eq.(11) represents the analog filter complex gain error based on the filter parameters.

$$\vec{E}_{AF} = [(1-u^2)^2 + (2\zeta u)^2]^{-0.5}\angle -\Gamma[(2\zeta u)/(1-u^2)] \quad (11)$$

Here, $u = 2\pi f/\omega_n$, ζ and ω_n are the analog filter damping coefficient and natural frequency, respectively. The analog filter error magnitude $|\vec{E}_{AF}|$ is linearized using Taylor series in the range of ±5 Hz to increase the algorithm's speed and reduce the computational effort on the microcontroller.

4.3 ADC Error

The ADC error is represented by the digitization process when the sample is approximated to the nearest digital level and the sampling error. The estimated phasor is measured with samples including the digitization error. The ADC complex gain error is represented by Eq.(12).

$$\vec{E}_{ADC} = [DFT(x_{(N+k)} + \gamma_{ADC})]/DFT(x_{(N+k)}) \quad (12)$$

where γ_{ADC} is the ADC approximation value that was added to the samples due to the digitization process. γ_{ADC} is a uniform distribution function and limited between in ± half the interval between the two digital levels $\gamma_{ADC} = rand(\pm 0.5 \times LSB)$. The ADC error is randomly variated and can not be estimated or compensated. However, this error is reduced by using the oversampling process of the AD7606 selected in the proposed μPMU. As the number of bits increases the digitization error reduces. For the sampling error, the PWM control signal with high accuracy needs a high accuracy timer. Systick timer with 24-bit ($\approx 5.95nsec$) is used, which is not perfect for the required sampling frequency because for a 50Hz system the sampling time is 39.06 μsec and the timer reload value is 6562.5≥ (SystemCoreClock/1000) / ($512 \times 50 \times 0.001$). To ensure accurate sampling accumulation and deaccumulation using variable sampling interval control strategy is used [28].

4.4 GPS Error

The μPMU GPS error is represented by the module response time response to the PPS signal. The STU is defined as the maximum response time of the GPS module. The response time τ of the module is assumed to be a Gaussian Distribution function with mean and standard deviation equal to μ= 0.5STU and δ = STU/6, respectively [13]. In the proposed μPMU GPS module SIM808 module is used, which has synchronization time accuracy equals $10nsec$, which gives an error in the angle of 0.00018^o while GPS magnitude is approximately one. The GPS complex gain error is represented by Eq.(13), which represents a lead phase shift to the estimated phasor.

$$\vec{E}_{GPS} \approx 1\angle 360^o \tau f \quad (13)$$

4.5 Off-nominal Error P-gain

Due to the off-nominal operation, when the frequency deviates from the nominal value, the calculated phasor is multiplied by P and Q complex gains, due to the leakage phenomena. The Q gain, which represents a second harmonics ripple, is mitigated using a simple

average digital filter discussed in section 3. The complex gain P error is represented by Eq.(14).

$$\vec{P} = \frac{Sin(\pi \Delta f N/f_s)}{NSin(\pi \Delta f/f_s)} e^{j\pi \Delta f(N-1)/f_s} \tag{14}$$

The P complex gain error magnitude and phase angle are linearized using Taylor series in the range of ±5 Hz to increase the algorithms speed and reduce the computational effort on the microcontroller.

4.6 White Noise

White noise is a natural error added to the measurements due to the environment and electronic distortion. Before exporting the output information, the white noise should be filtered. For this task, the Moving Average Filter (MAF) is used with a window of $N/2$. Eq.(15) shows the MAF representation.

$$y(i)^* = \frac{2}{N} \sum_{n=0}^{\frac{N}{2}-1} y(i-n) \tag{15}$$

where $y(i-n)$ is the output measurements and $y(i)^*$ is the filtered value. Table.2 shows the error contribution of each process on the estimated phasor and the theoretical TVE estimated using Monte Carlo Simulation of 1,000,000 simulations. Transducers error shows the major contribution when they loaded by 5% of the full rate. In order to meet the accuracy limits, the developed μPMU should be calibrated against these errors and compensate for them. Figure.3 shows the μPMU processes structure and operation sequence.

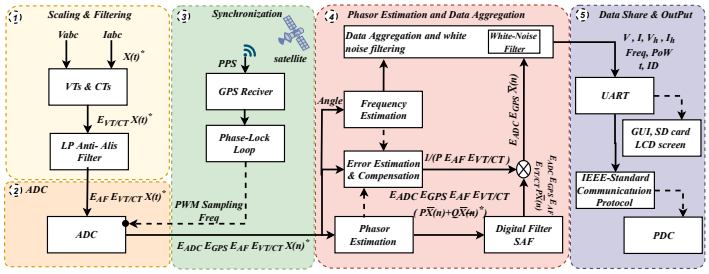


Figure 3: μPMU operation structure.

It is worth noting that in the design of the software and hardware, the measurement accuracy, software speed, and overall cost are the three main performance parameters that we care about during the design. We always looking to select the cheapest hardware that ensures the required level of accuracy with low computational time and reliable operation.

4.7 Taylor series linearization

Error compensation is an important step to improve the proposed μPMU accuracy over the previous work. However, the error compensation algorithm should be designed to be light and fast. Since the total μPMU computational time equals the sampling time, using the non-linear equations of the gain error limits the idea of error compensation. Therefore, all error gains discussed in section 4 are represented by a Taylor series. Eq.(16) shows the linearization of the complex gain error E to L order, which is mainly a function of the measured quantity d.

$$E(d) = a_0 + a_1 d + ... + a_L d^L + e \tag{16}$$

e is the error between the $E(d)$ values and the linearization function. The matrix form of the Eq. (16) for m values is represented by Eq. (17).

$$\begin{bmatrix} E(1) \\ : \\ E(m) \end{bmatrix} = \begin{bmatrix} 1 & d(1) & d(1)^L \\ : & : & : \\ 1 & d(m) & d(m)^L \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ : \\ a_L \end{bmatrix} \tag{17}$$

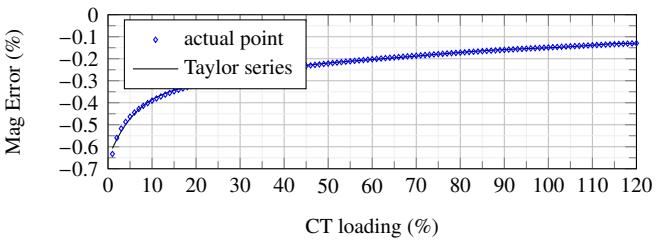
Eq. (17) is solved using Least Square in Eq. (18) to find the A matrix coefficients, using D and E matrix.

$$A = [(D^T . D)^{-1} . D^T].E, \; m \geq (L-1) \tag{18}$$

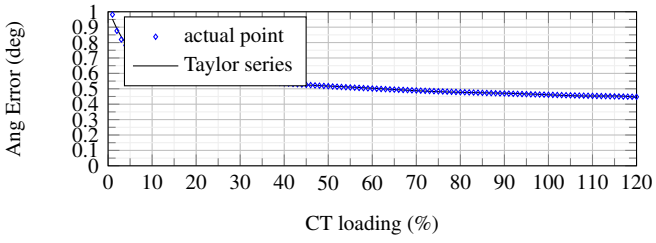
Using high order reduces the error e between the complex gain error and the linearization function. On the other hand, using high order increases the number of sum and multiplication operation, which increase the computational burden of the error compensation stage. Therefore, the linearization function order is selected as the smallest order to represent the complex gain error with an acceptable error range. It is important to note that for each complex gain two linearization functions are estimated for the magnitude and angle.

Figure.4 shows the estimated Taylor series function for the complex gain errors of the magnitude and angle for the MVS CT as an example of the power transducer, analog filter, and the complex gain P, respectively. The analog filter transfer function is estimated using the given data-sheet gain and phase frequency response. It is worth noting that there is no linearization function for the GPS error or the ADC error. Since the error produced by these elements is randomly generated, the errors of both stages are only reduced through the μPMU main measurement unit robust design such as using oversampling ADC properties, and high accuracy GPS module.

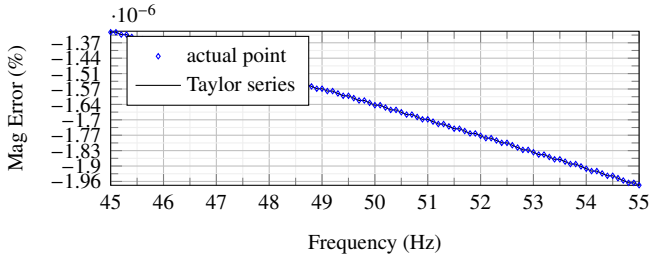
Table. 2 Shows the effect of error compensation, which shows a high improvement in the accuracy of the measurement. It is worth noting that to ensure low operation time, complex gain errors with low effect are not compensated. For example, the magnitude error due to the analog filter is very small compared with the required design limits needed. Adding error compensation for this small error, increase the computational burden on the microcontroller with the limited operation time. On the other hand, the angle error due to the analog filter is high and above the target limits of the proposed design. Therefore, analog filter angle error is considered in the error compensation stage. Gray elements in Table.2 show the error values that are required to be compensated to improve the μPMU accuracy to satisfy the target limits. Table.2 summarizes the performance of the enhancement stage, where the error compensation time reduces from 42 μsec to 0.8 μsec.



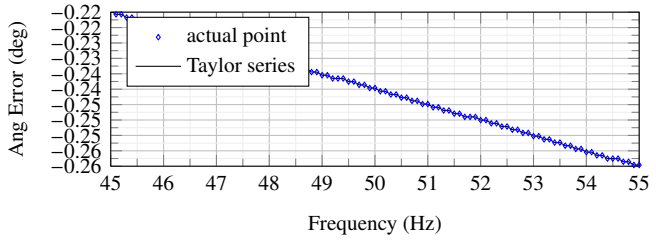
(a) CT magnitude error linearization function



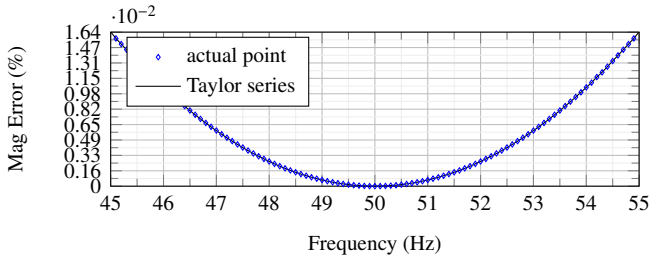
(b) CT angle error linearization function



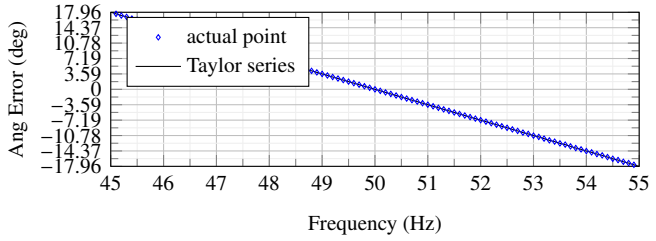
(c) SAF magnitude error linearization function



(d) SAF angle error linearization function



(e) P gain magnitude error linearization function



(f) P gain angle error linearization function

Figure 4: Complex error gains linearization

5 Testing and validation

In this section, the μPMU tests are discussed according to IEEE Std C37. 118.1 and 118.2 to evaluate the main measurement unit operation and performance [14, 15]. The unit steady-state and dynamic operation is examined using six tests and the result is compared with previous work. Opal-RT simulator-OP4510 is used to generate a reference signal to the μPMU in each test and the measurements are sent to the simulator to compare and estimate the error between the two signals. Figure.5 shows the experimental setups.

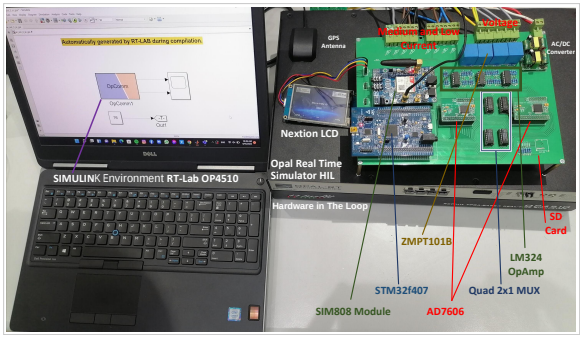


Figure 5: μPMU test and validation environment.

5.1 *Steady-state tests*

In steady-state operation, three tests are conducted. In the first test, the phasor magnitude estimation performance is tested in the range from 0.1-1 p.u with a variation step of 0.1 p.u. The test is made under nominal operation. In the second test, the frequency estimation performance is tested in the range of 49.5-50.5 Hz with a step change of 0.1 Hz. In the third test, the phasor angle estimation performance is tested using input signals with phase shifts 1, 0.1, and 0.01 degrees to estimate the angle accuracy and resolution.

5.2 *Dynamic tests*

In dynamic operation, three tests are conducted to test the performance of the μPMU under dynamic operation. In the first test, an input signal with a ramp frequency change of 0.2 Hz/sec is applied to the μPMU, the main purpose of this test is to evaluate the performance of estimating the ROFC and frequency. In the second test, two input signals are applied to the μPMU, one with a sudden change by 0.1 p.u and 10 degrees to estimate the phasor step response of the μPMU and the other signal with a step frequency change by ± 0.5 Hz to estimate the frequency step response. In the third test, a sinusoidal wave of fundamental and 40% harmonics is measured by the proposed μPMU. The test is made for harmonics content up to 16th harmonic order. The μPMU is adjusted to operate on the dynamic mode to provide the point-on-wave measurement in this test. The main purpose of this test is to check the performance of the μPMU to estimate the harmonics content of the input signal and shows the point-on-wave data.

Table 2: Theoretical error limits using MCS and Taylor linearization performance

Error Source		Maximum Error Without Error compensation		Maximum Error With Error compensation		Taylor estimation performance							
		Mag %	Ang^o	Mag %	Ang^o	L		Error % in Taylor fun		Time μsec before Taylor total 48 μsec		Time μsec after Taylor total 0.8 μsec	
						Mag	Ang	Mag	Ang	Mag	Ang	Mag	Ang
CT /VT		1.5	1.5	0.00549	0.44	8	9	0.366	0.44	-	-	0.25	0.25
Analog Filter		6.5e-9	0.252	1.27e-14	1e-6	2	1	1.9e-6	1e-6	4	12	0.1	0.05
ADC digitization		±9.83e-5	±4.9e-5	± 9.83e-5	±4.9e-5	-	-	-	-	-	-	-	-
GPS		≈0	-1.8e-4	≈0	-1.8e-4	-	-	-	-	-	-	-	-
off-nominal	Q	±0.005	-	±0.005	-	-	-	-	-	-	-	-	-
	P	0.0164	±17.965	1.0496e-4	±4e-13	2	1	4.6e-3	4e-13	14	12	0.1	0.05

6 Result and Discussion

In this section the performance of the proposed μPMU is analyzed under the standard tests of IEEE- C37.118.1 and C37.118.2 is discussed.

6.1 Steady-state operation

6.1.1 Test1 (Magnitude step change)

The result of the magnitude step change is shown in Figure.6. The magnitude estimation under nominal operation shows a very accurate result, the TVE is 0.0318% including the current-type VT without error compensation. The bipolar structure shows more stability and less error compared with adding DC off-set as in [22]. The error in the estimated frequency showed a mean of 3.5 mHz and a standard deviation in the measurement of 1mHZ, which resulted in a maximum variation of 6mHz in a transient change in the voltage. However, compared with previously designed units in the literature the proposed μPMU shows lower variations.

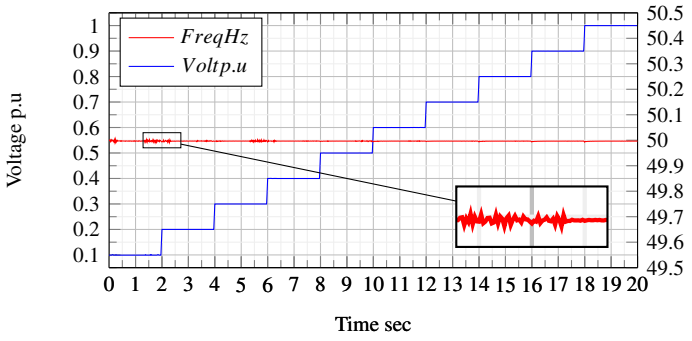


Figure 6: Magnitude step change.

6.1.2 Test2 (Frequency step change)

The result of the frequency step change is shown in Figure.7. The magnitude estimation under off-nominal operation shows degradation in the accuracy. For the magnitude, the TVE increased to 0.7% at the transition period shown by pulses in the graph. Otherwise, the TVE is less than 0.05%. The error in the estimated frequency showed a maximum variation of 0.1 Hz at the transient change, while inside the range stable operation, the frequency maximum variation is 30 mHz. The increased error in both the magnitude and frequency happened due to the off-nominal operation complex gain error P and Q.

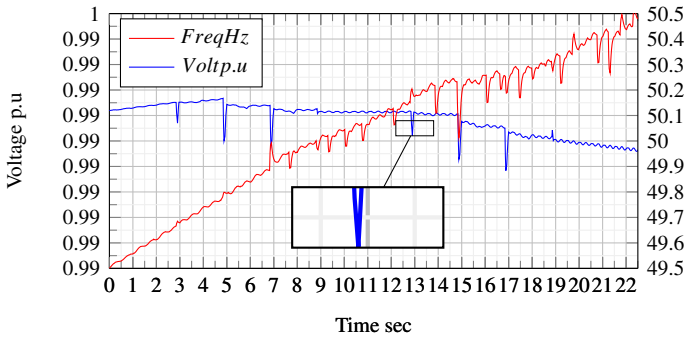


Figure 7: Frequency step change.

6.1.3 Test3 (Angle accuracy and resolution)

Three signals with a phase shift of 1, 0.1, and 0.01 degrees are applied to the μPMU channel. Figure.8 shows the measured angle in all tests. In the three cases, the μPMU showed robust estimation. The mean of the error is constant in the three cases and has a mean of 0.0036^o lag phase shift, due to the analog filter and the complex gain P, and a very low standard deviation equals $2.85^o \times 10^{-5}$. The μPMU angle resolution detected is 0.0036 degrees and reflects the ADC resolution with a sampling frequency of 512 samples/cycle.

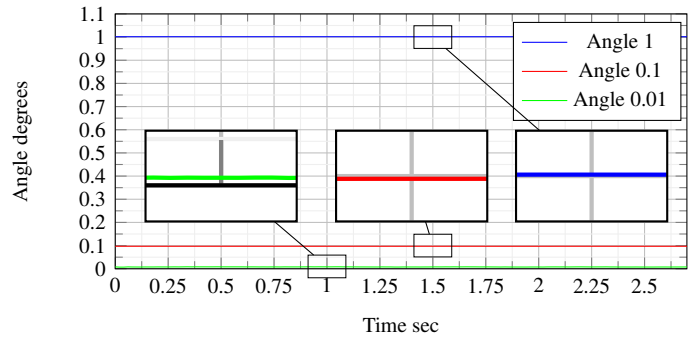


Figure 8: Angle variation.

6.2 Dynamic Operation

6.2.1 Test4 (Ramp frequency change)

When the ramp signal is applied to the μPMU, the estimated ROFC equals the input signal value of 0.2HZ. Figure.9 shows the reported estimated frequency during the test. At the nominal region, the frequency has low ripple and low error, as the frequency deviated with time (increase or decrease), the off-nominal caused a deviation in the measured frequency. As the difference from the nominal frequency increases, the deviation of the measured frequency also increases due to the mismatch of the analog filter and the off-nominal error due to P and Q complex gain frequency.

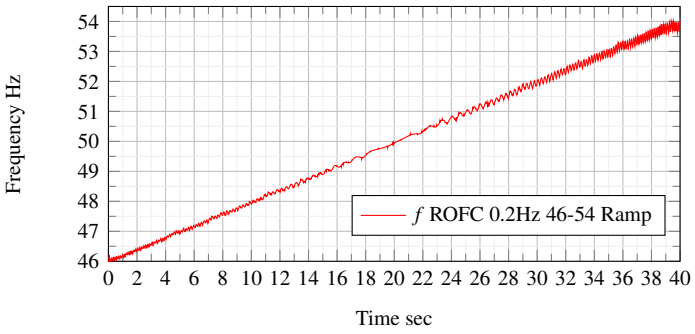


Figure 9: Ramp frequency change.

6.2.2 Test5 (Phasor and frequency step response)

In the case of phasor step change, the phasor magnitude and the angle have the same TVE, accuracy, and resolution in the steady-state. Both magnitude and phase response to the change at the same time and have a response time of 20 msec, which represents the reporting rate and it is also equal to the μPMU latency as shown by Figure.10. This value depends on the selected reporting rate due to the operation mode and can be reduced by adjusting the reporting rate of 2 measurements/cycle. In the case of frequency step change by ± 0.5 Hz, at the nominal operation, the variation of the measured frequency is within the range of 4 mHz, which results in a TVE of 0.008%. The frequency of the two signals is changed with a step of ± 0.5Hz and 49.5, which causes off-nominal operation. Therefore, the second harmonics ripple is added to the measured phasor causing a second harmonics to change in the angle and frequency. The SAF mitigates this ripple to a final variation of ± 16 mHZ. Figure.11 shows the reported frequency measurements.

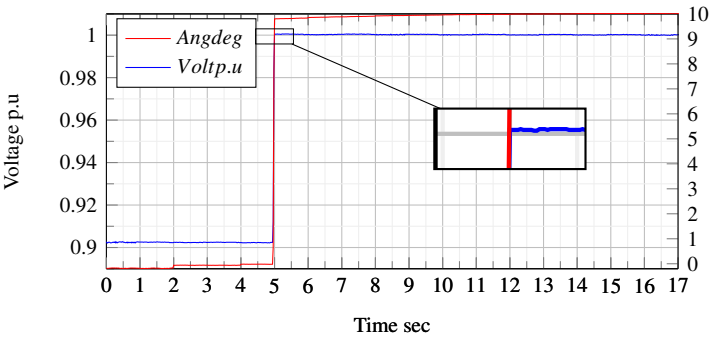


Figure 10: Phasor step response.

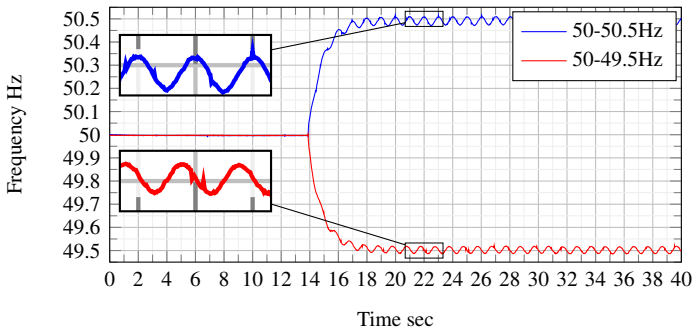


Figure 11: Frequency step response.

6.2.3 Test6 (Harmonics content)

In this test, a sinusoidal wave of 60% fundamental and 40% harmonics from 2nd-16th order is measured by the μPMU. Figure.12 represents the point-on-wave data recorded of the first four signals. Figure.13 the p.u error in the estimated harmonics is shown. The maximum error is found to be 1.36% in the 6th harmonics. In all harmonics, the error is around 1.206%. This magnitude shift is made by the amplifier LM324 DC shift and uncertainty in the gain, which could be improved by using a better amplifier in future design. Table.3 shows a comparison between the proposed unit previous proposed units.

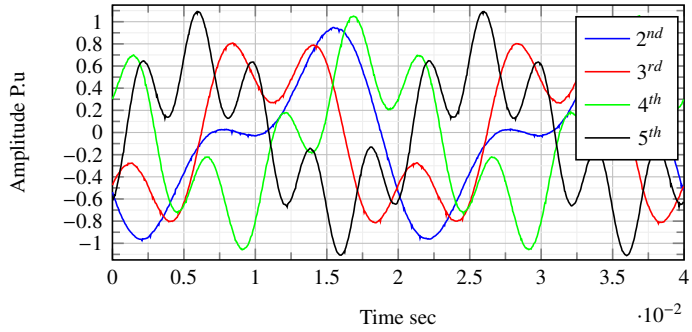


Figure 12: Harmonics Point-on-wave data 2nd-5th.

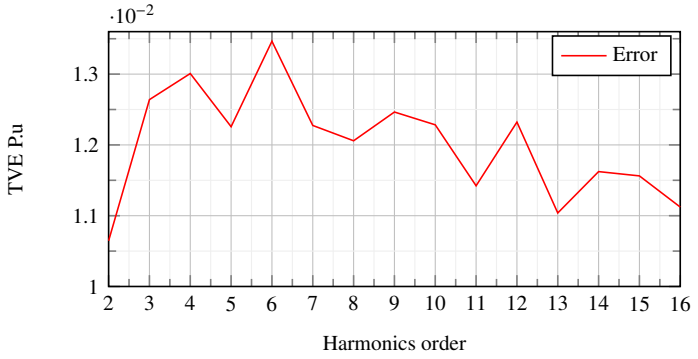


Figure 13: Harmonics estimation error.

Table 3: Comparison of μPMUs

Specifications	HW	Phasor-SW	fs	latency msec	steady state max variation			Dynamic max variation			Cost
					Mag %	Ang deg	Frq %	Mag %	Ang deg	Frq %	K$
Proposed unit	Full structure	Linearized DFT and Error	512	20	0.03	3.6e-3	6e-3	0.05	3.6e-3	0.032	0.25
[10]	Smart meters and GPS	DFT	32	-	0.1	0.3	0.01	0.8	-	0.06	0.3
[11]	Filter, GPS, Mc and communication	DFT	64	22.5	0.062	-	6e-3	0.13	-	0.022	0.4
[12]	ADC, GPS and Mc	Modified DFT	256	-	-	-	0.032	-	-	0.57	-
[13]	ADC, GPS and Mc	i-IpDFT	512	0.07 MC	-	-	-	-	-	-	0.8
PSL	Full structure	-	512	20	0.05	0.003	0.01	-	-	-	12

7 Conclusion

This article presents a roadmap to design distribution level μPMU to meet the requirements of distribution grid applications. The μPMU hardware and software are discussed to ensure the robust design and minimum software operation time. Proposed μPMU shows accurate performance during steady-state and dynamics operation and meets their application requirements with an accuracy and resolution up to 1% and 0.1% for the magnitude and 0.1^o and 0.01^o for the angle. The proposed μPMU address the required limits with a phasor estimation accuracy up to 0.03% for the magnitude and 0.0036 angle resolution, while the frequency is estimated with accuracy up to 0.006% under steady-state. For dynamic operation, both phasor and frequency measurements show a certain level of degradation due to the dynamic and off-nominal error. Power system frequency variation shows a higher effect over magnitude and phase variation. The magnitude and frequency of TVE are degraded to 0.05% and 0.032% under dynamic operation, while the angle variation is the same for both cases. By using linearization, the off-line calculation time is reduced to $18\mu sec$, which allows stable communication. The overall μPMU performance addressed the required specification and showed robust performance under steady-state and dynamic operation.

Acknowledgment

This work was done as part of the research project " Smart Monitoring system for Distribution Power Grid " supported by the Egyptian National Telecom Regulatory Authority (NTRA). Also, the authors would like to thank the Smart-Systems Company for providing the real-time simulator and the μPMU test environment.

References

[1] J. A. Momoh, Smart grid: fundamentals of design and analysis, volume 63, John Wiley & Sons, 2012.

[2] A. Von Meier, E. Stewart, A. McEachern, M. Andersen, L. Mehrmanesh, "Precision micro-synchrophasors for distribution systems: A summary of applications," IEEE Transactions on Smart Grid, **8**(6), 2926–2936, 2017, doi: 10.1109/TSG.2017.2720543.

[3] F. C. Trindade, W. Freitas, "Low voltage zones to support fault location in distribution systems with smart meters," IEEE Transactions on Smart Grid, **8**(6), 2765–2774, 2016, doi:10.1109/TSG.2016.2720544.

[4] P. De Oliveira-De Jesus, C. H. Antunes, "Economic valuation of smart grid investments on electricity markets," Sustainable Energy, Grids and Networks, **16**, 70–90, 2018.

[5] X. Zhao, D. M. Laverty, A. McKernan, D. J. Morrow, K. McLaughlin, S. Sezer, "GPS-disciplined analog-to-digital converter for phasor measurement applications," IEEE Transactions on Instrumentation and Measurement, **66**(9), 2349–2357, 2017.

[6] T. Ahmad, N. Senroy, "Statistical characterization of PMU error for robust WAMS based analytics," IEEE Transactions on Power Systems, **35**(2), 920–928, 2019.

[7] D. Schofield, F. Gonzalez-Longatt, D. Bogdanov, "Design and implementation of a low-cost phasor measurement unit: A comprehensive review," in 2018 Seventh Balkan Conference on Lighting (BalkanLight), 1–6, IEEE, 2018.

[8] P. Romano, M. Paolone, T. Chau, B. Jeppesen, E. Ahmed, "A high-performance, low-cost PMU prototype for distribution networks based on FPGA," in 2017 IEEE Manchester PowerTech, 1–6, IEEE, 2017.

[9] M. C. Garcia, D. Dotta, L. Pereira, M. C. de Almeida, M. R. Paternina, O. L. dos Santos, L. C. da Silva, J. E. d. R. A. Junior, "A development PMU device for living lab applications," Journal of Control, Automation and Electrical Systems, **32**(4), 1111–1122, 2021.

[10] M. C. Garcia, D. Dotta, L. Pereira, M. C. de Almeida, O. L. Santos, L. C. da Silva, "Design and development of D-PMU module for smart meters," in 2020 IEEE Power & Energy Society General Meeting (PESGM), 1–5, IEEE, 2020, doi:10.1109/PESGM41954.2020.9281601.

[11] R. N. Rodrigues, L. H. Cavalcante, J. K. Zatta, L. C. M. Schlichting, "A Phasor Measurement Unit based on discrete fourier transform using digital signal processor," in 2016 12th IEEE International Conference on Industry Applications (INDUSCON), 1–6, IEEE, 2016.

[12] PSL, "Micro-PMU Datasheet," in [Online], Available:https://powerside.com/wpcontent/uploads /2020/01/Powerside-microPMU-DataSheet-.pdf, Accessed on: Aug.1, 2021.

[13] A. A. E. Elsayed, M. A. Mohamed, M. A. Nayel, "Distribution Grid Phasor Estimation and μPMU Modeling," in 2022 IEEE International Conference on Power Electronics, Smart Grid, and Renewable Energy (PESGRE), 1–8, IEEE, 2022.

[14] I. Power, et al., "IEEE Standard for Synchrophasor Measurements for Power Systems–Amendment 1: Modification of Selected Performance Requirements," IEEE Std C37. 118.1 a-2014 (Amendment to IEEE Std C37. 118.1-2011), **2014**, 1–25, 2014.

[15] IEEE, "IEEE standard for synchrophasor data transfer for power systems," IEEE Standard C37. 118.2-2011 (Revision of IEEE Std C37. 118-2005), 1–53, 2011.

[16] NASPI, "Distribution Synchronized Measurements Roadmap Final Report," in [Online], https://www.naspi.org/node/934, Accessed on: Apr.15, 2022.

[17] J. Zhao, G. Zhang, "A robust prony method against synchrophasor measurement noise and outliers," IEEE Transactions on Power Systems, **32**(3), 2484–2486, 2016.

[18] A. G. Phadke, J. S. Thorp, "Phasor measurement units and phasor data concentrators," in Synchronized Phasor Measurements and Their Applications, 83–109, Springer, 2017.

[19] "IEEE Draft Recommended Practices and Requirements for Harmonic Control in Electric Power Systems," IEEE P519/D6ba, September 2013, 1–26, 2013.

[20] A. A. E. Elsayed, M. A. Mohamed, M. Abdelraheem, M. A. Nayel, "Optimal μPMU Placement Based on Hybrid Current Channels Selection for Distribution Grids," IEEE Transactions on Industry Applications, **56**(6), 6871–6881, 2020, doi:10.1109/TIA.2020.3023680.

[21] A. A. Elaziez, M. A. Mohamed, M. AbdelRaheem, M. A. Nayel, "Optimal μPMU placement and current channel selection considering running cost for distribution grid," in 2020 IEEE International Conference on Power Electronics, Smart Grid and Renewable Energy (PESGRE2020), 1–8, IEEE, 2020.

[22] I. Abubakar, S. Khalid, M. Mustafa, H. Shareef, M. Mustapha, "Calibration of ZMPT101B voltage sensor module using polynomial regression for accurate load monitoring," Journal of Engineering and Applied Sciences, **12**(4), 1077–1079, 2017.

[23] A. G. Phadke, J. S. Thorp, Synchronized phasor measurements and their applications, volume 1, Springer, 2008.

[24] G. Daoud, H. Selim, M. M. AbdelRaheem, "Micro phasor measurement unit phasor estimation by off-nominal frequency," in 2018 IEEE International Conference on Smart Energy Grid Engineering (SEGE), 53–57, IEEE, 2018, doi: 10.1109/SEGE.2018.3023680.

[25] M. S. Ballal, M. G. Wath, H. M. Suryawanshi, "A novel approach for the error correction of CT in the presence of harmonic distortion," IEEE Transactions on Instrumentation and Measurement, **68**(10), 4015–4027, 2018.

[26] A. Rezaei-Zare, R. Iravani, M. Sanaye-Pasand, H. Mohseni, S. Farhangi, "An accurate current transformer model based on Preisach theory for the analysis of electromagnetic transients," IEEE Transactions on Power Delivery, **23**(1), 233–242, 2007.

[27] M. Andreev, A. Suvorov, N. Ruban, R. Ufa, A. Gusev, A. Askarov, A. Kievets, "Development and research of mathematical model of current transformer reproducing magnetic hysteresis based on Preisach theory," IET Generation, Transmission & Distribution, **14**(14), 2720–2730, 2020, doi:10.1049/iet-gtd.2018.6796.

[28] W. Yao, L. Zhan, Y. Liu, M. J. Till, J. Zhao, L. Wu, Z. Teng, Y. Liu, "A novel method for phasor measurement unit sampling time error compensation," IEEE Transactions on Smart Grid, **9**(2), 1063–1072, 2016, doi: 10.1049/smg.2016.6796856.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 1, 63-77 (2023)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

An Efficient Way of Hybridizing Edge Detectors Depending on Embedding Demand

Habiba Sultana*, A. H. M. Kamal

Computer Science and Engineering, Jatiya Kabi Kazi Nazrul Islam University, Mymensingh, 2220, Bangladesh

ARTICLE INFO

Article history:
Received: 29 August, 2022
Accepted: 01 January, 2023
Online: 24 January, 2023

Keywords:
Steganography
Edge detection
PSNR

ABSTRACT

Edge detection-based image steganography schemes usually embed data in edge pixels only. However, some schemes embed data in non-edge pixels as well. In that case, the schemes embed more bits in the edges than in the smoothed areas. In all cases, the schemes perform large changes in a tiny area of the image during small data embedding. Detecting such local modifications is comparatively easier for a steganalyzer. As a result, it is preferable to distribute bits evenly across the image. Again, the schemes struggle to hide large messages in a cover image due to the algorithmic approach of hiding a fixed number of bits per pixel. In this research, we have overcome that problem by employing multiple edge detectors in generating a resultant edge image. Depending on the embedding needs, we have checked whether a single edge detector is sufficient to help in conceiving all bits or not. If it is not possible for a single-edge detector, we have hybridized them. Hybridization of edge images is performed either by logical AND, OR or OR with dilation. When the message size is very small, we have generated the resultant edge image by doing a logical AND operation among the edge images. That strategy have reduced the number of edge pixels as well as helped in distributing the to-be-embedded bits over the image in a more evenly manner. Similarly, to meet a larger embedding demand, we have performed a logical OR operation among the same edge images to increase the number of edge pixels. Even, to meet more embedding demand, we have dilated the OR-resultant image. These processes were carried out dynamically in the research according to an embedding demand. The experimental results deduce that this scheme embeds 92.37%, 73.92%, 74.78%, *and* 9.60% *more bits than four competing methods. Similarly, for small embedding demand, the proposed scheme demonstrates* 37.45%, 46.87%, 44.21%, *and* 55.56% *higher PSNR values than the others. Moreover, statistical analyses state that this scheme demonstrates stronger security against attacks.*

1 Introduction

In steganography, an embedding method implants a secret in a cover media such as a text file, digital image, audio, video, IP protocol, bio-signals, DNA sequence, etc [1]. By implanting secrets in a media, these methods modify the contents of that cover media. That modified media is then known as stego media. As a cover media, digital images are widely used in steganography because of their higher degree of redundant information [2]. The performance of image steganography is mainly measured by a set of features like payload, imperceptibility, and security of stego image [3]. Steganography methods work in either spatial domain, transform coefficients or created residues [1]. In the spatial domain, the confidential information is concealed in either pixel values or to their processed values [4]–[6]. There a very commonly used method is least significant bit (LSB) substitution. In the transformed domain, the schemes first transform image contents, e.g., by wavelet transform, Fourier transform, etc., and then implant secrets in these coefficients [7]. In residue-based methods, the schemes implant secrets, generally, in pixel value differences (PVD) and prediction errors [8].

In terms of blindness, these schemes are categorized as reversible [6], [9]–[21] and irreversible [4, 5], [22]–[33] groups. In reversible steganography, the receiver rebuilds the cover image from the stego image in addition to extracting the desired secret message. On the other hand, the irreversible schemes extract the secrets only. Irreversible schemes are easy to implement and provide higher embedding capacity. For this, our research focuses on irreversible techniques. We concentrate our research target on the spatial domain only.

Machine learning is frequently used in cancer and kidney stone

*Corresponding Author: Habiba Sultana, & Email: srity.cse@gmail.com

detection, image retrieval, and brain stroke [34]–[39]. Many of such applications use edge detection algorithm for localizing and visualizing target area in image and data. Before applying machine learning, if one wishes to implant privacy-preserving and security-related data [40]–[42] in the detected edge information that could be a promising technique to be used in the tele-medicine applications. Therefore, it is interesting to associate an edge detection method to divide the image contents into the edge and non-edge areas and to hide the data there [5, 6, 10, 31, 43].

In [22], [24], [33], [6], [2], the author used Canny edge detector to identify edge and non-edge pixels. All of the schemes used the LSB substitution method to hide data bits. In [25], [28] the author worked with different authors and applied a Canny edge detector in both cases to detect edge pixels. In [25] and [28], authors implanted data bits using reduplicated exploiting on the modification direction (REMD) and hybrid Hamming codes, respectively. In [8], the author employed a Canny edge detector as well and applied exclusive OR operation as a part of their embedding process. In [23], the author did the same but partitioned the image first into blocks. In addition to data implantation tasks, these schemes tried to their own ways to maintain a better visual quality in their stego images. In [3] and [31], the author used hybrid edge detection techniques in their data hiding process. They measured pixel value differences (PVD) first to decide the number of implanted bits per pixel and then used the LSB substitution method for data hiding. In [30], the author associated the Canny edge detector and PVD in their data hiding technique as well.

In [1] and [27], the author employed fuzzy edge detector to detect edge pixels and then used LSB substitution approach to implant the secrets into these detected edge pixels. In [1], the author used a chaotic method as a pre-processing task to encrypt the secret message.

Some authors first used multiple edge detectors using diverse edge operators, e.g., Canny, Sobel, and Fuzzy operators, and then hybridized these edge images to increase the number of edge pixels [4] in the resultant edge image. They used the LSB substitution method in their data implantation phase. To minimize the distortion of [4], in [26], the author divided the image into blocks and thereafter, they applied a hybridization process to these detected edge images. In [31], the author used another hybrid edge detection technique. He, additionally, performed a morphological dilation operator in their data-hiding phase to increase embedding capacity. In [32], the author hybridized the edge images by logical AND operator to increase the stego image quality while implanting small sized messages. In [5], the author proposed another edge detection based steganography method. They tested their scheme for Canny, Sobel, and Fuzzy-based edge detectors. They used the renown LSB substitution method to implant data bits. In [29], Ghosal proposed a steganography scheme using the Kirsch edge detection method where they implanted the message bits into each triplet of pixels. Therefore, the embedding capacity is low.

In this study, we have proposed a new hybrid edge detection-based embedding process where it embeds more bits in edge pixels than non-edge ones. The proposed scheme employs multiple edge detection methods and finds the best detector for what the demanded message bits are just conceivable. Depending on embedding demand, it determines whether a single edge detector is capable to help in hiding entire secrets. If the size of the secret message is too small or very high than the embedding capacity by using a single edge detector, it hybridizes the edge images in both cases. When the message length is too small, the proposed scheme hybridizes edge images by logical AND operator to reduce the number of edge pixels. The number of edge images is employed in the AND operation depending on the message length. The resultant edge image helps the embedding algorithm to distribute the secret bits in the cover image more evenly. Again, to implant a large-size message, it performs a logical OR operation among the edge images to increase the number of edge pixels in the resultant edge image. Even, if OR is unable to implant the whole secret message, the scheme employs a morphological dilation operation to further increase the number of edge pixels in the resultant edge image. The scheme does that hybridization and dilation operation dynamically realizing the length of the secret message and computing the embeddable bits through that resultant edge image. Experimental results show that our proposed scheme performs better than the other competing methods for all the performance measuring parameters.

Contributions of this research are listed below:

- This scheme dynamically chooses the best one of the four different embedding techniques depending on the message length.
- We have allowed the scheme to implant a different number of bits in edge and non-edge pixels according to embedding demand.
- Our proposed method increases the visual quality of stego image and embedding capacity as well. At the same time, it shows strong resistance against statistical attacks.

The rest of this paper is organized as follows: section 2 concisely presents the related works. The proposed method is described neatly in section 3. Section 4 demonstrates the simulated results of our scheme. The results of testing the robustness of the proposed scheme against attacks are devoted in section 5. Finally, section 6 concludes the article.

2 Related Works

2.1 A Brief on Edge Detectors

The sharp changes in the image brightness are called the edges or boundaries of the image. Edges may exist in horizontal, vertical, or diagonal directions. The method which is used to detect the edges of an image is called edge detection. A filter, known as kernel or operator is used to identify the edges in an image. Very commonly used edge detectors are canny, sobel, log, Prewitt, kirsch, laplacian, and fuzzy. Generally, edge detectors are used in pattern recognition, feature extractions, and image morphology. In the field of detection-based, edge detectors are used to improve the security of data hidden. These schemes first detect the edge pixels and non-edge pixels in a cover image and implant only edge pixels or both categories. We have studied a good number of articles on that state art. Among those, we found the works of [5], [31], [33] and read them very

carefully and attentively and built the foundation of our proposed work on them.

2.2 Single edge-based image steganography

In [5], the author proposed an edge-detection-based steganographic method. This scheme copied the cover image and cleared $n-$bits LSBs from it. Then it applies various popular edge detectors such as canny, Sobel, fuzzy, etc., and generates an edge image. This scheme classifies the contents of cover image pixels as edge and non-edge pixels based on that edge image. This scheme then implants x bits of secrets into edge pixels and y bits into non-edge pixels and generates a stego image. This is the embedding process. In the extraction phase, the receiver extracts the secret messages from the stego image using the reverse process of embedding.

2.3 Hybrid edge-based image steganography

In 2018, Rasol [33] proposed an image steganography using a hybrid edge detection technique. In this scheme, the authors apply canny and Sobel edge detection methods and generate edge area. They combined those edge areas using logical OR operation. On the other hand, they also add a special character at the end of the message and convert it into binary according to ASCII. Then they embed x bits into the edge area and y bits into the non-edge area using the LSB method and thus generate a stego image and send it to the receiver. The receiver performs the reverse of the embedding process and extracts the secret message.

2.4 Dilated hybrid edge-based image steganography

In [31], the author proposed a dilated hybrid edge detection-based image steganography scheme. This method has three phases such as preprocessing, embedding, and extraction. Like as [5], this scheme also copies the cover image and cleared $n-$bits LSBs from it. This scheme then applies $m-$number of edge detectors such as $e_1, e_2, ..., e_m$ and combine two edge detectors using logical OR operation. This scheme also used morphological operators such as dilation to increase the number of edge pixels. Based on the dilated hybrid edge image, all the contents of the cover image pixels are classified as edge pixels and non-edge pixels. The XOR operator is used in the embedding process to improve security. This scheme implants x bits information in edge pixels and y bits into non-edge pixels and generates a stego image. In the extraction phase, the receiver extracts the secret message using the reversible method of embedding.

2.5 Edge-based image steganography

In [32], the author proposed an image steganography method based on hybrid edge detection. This scheme all most similar to setiadi's method [31]. The difference between those schemes is, [32] is applicable for small message sizes with maintaining good visual quality and [31] has good embedding capacity with maintaining visual quality. The scheme [32] used logical AND operation instead of logical OR operation.

Table 1 gives a summary of this work.

Table 1: Summary of related works. Uses of multiple edge detectors and dilating the hybrid edge image are the key distinguishing features of this scheme.

Criteria	**[32]**	**[5]**	**[33]**	**[31]**
Cleared LSBs?	Yes	Yes	No	Yes
Hybridize edge images?	Yes	No	Yes	Yes
Dilate edge image?	Yes	No	No	Yes
Use an edge detector?	Yes	Yes	Yes	Yes
Encrypt message?	No	No	Yes	Yes
Embed as (x,y) bits?	Yes	Yes	Yes	Yes
Message type?	Text	Binary	Text	Text

3 Proposed method

The proposed work consists of three phases: pre-processing, data embedding, and data extracting. The description of these phases is given below:

3.1 Pre-processing phase

As our target is to implant the message in both edge and non-edge pixels, we perform single edge detectors or hybridize edge detectors based on message length m_L and maximum achievable payload. We take an instance of the cover image and cleared $n-$ bits LSBs from it. Then we apply $m-$ number of edge detectors. We select the best suitable edge detector in the following way:

$$rE = \begin{cases} \text{E} & \text{when } m_L > P_s \text{ and } m_L < P_h \\ AND(\text{E_1,E_2....E_m}) & \text{when } m_L \leq P_s \\ OR(\text{E_1,E_2....E_m}) & \text{when } m_L \geq P_h \\ dilation(E) & \text{otherwise} \end{cases} \tag{1}$$

We select single edge detector E when message length m_L is greater than the probable highest payload P_h and less than the probable smallest payload P_s. When message length m_L is less than P_s then we hybridize two or more edge images using AND operation. we also hybridize two or more edge images using OR operation when message length m_L is greater than the probable highest payload P_h. In another case, we used morphological operator dilation when the message length is large. Let the cover image is C and an instance of it by I. We first clear $n-$bits of LSBs from every pixel of I by equation(2).

$$I(i, j) = I(i, j) - f(I(i, j), 2^n); \tag{2}$$

where function f returns the remainder value when one divides I(i,j) by 2^n.

We have applied $m-$number of edge detection operators, e.g., canny, sobel, fuzzy, Robert, Prewitt, log, etc on the cleared image I to detect edge pixels, separately. We have generated the edge image by equation (3).

$$eI(i) = \psi(I, \Omega); \tag{3}$$

where ψ is one of the m edge detection operators, i.e., $\Omega\epsilon\{canny, sobel, log, fuzzy, Robert, Prewitt, etc.\}$ and $1 <= I <=$

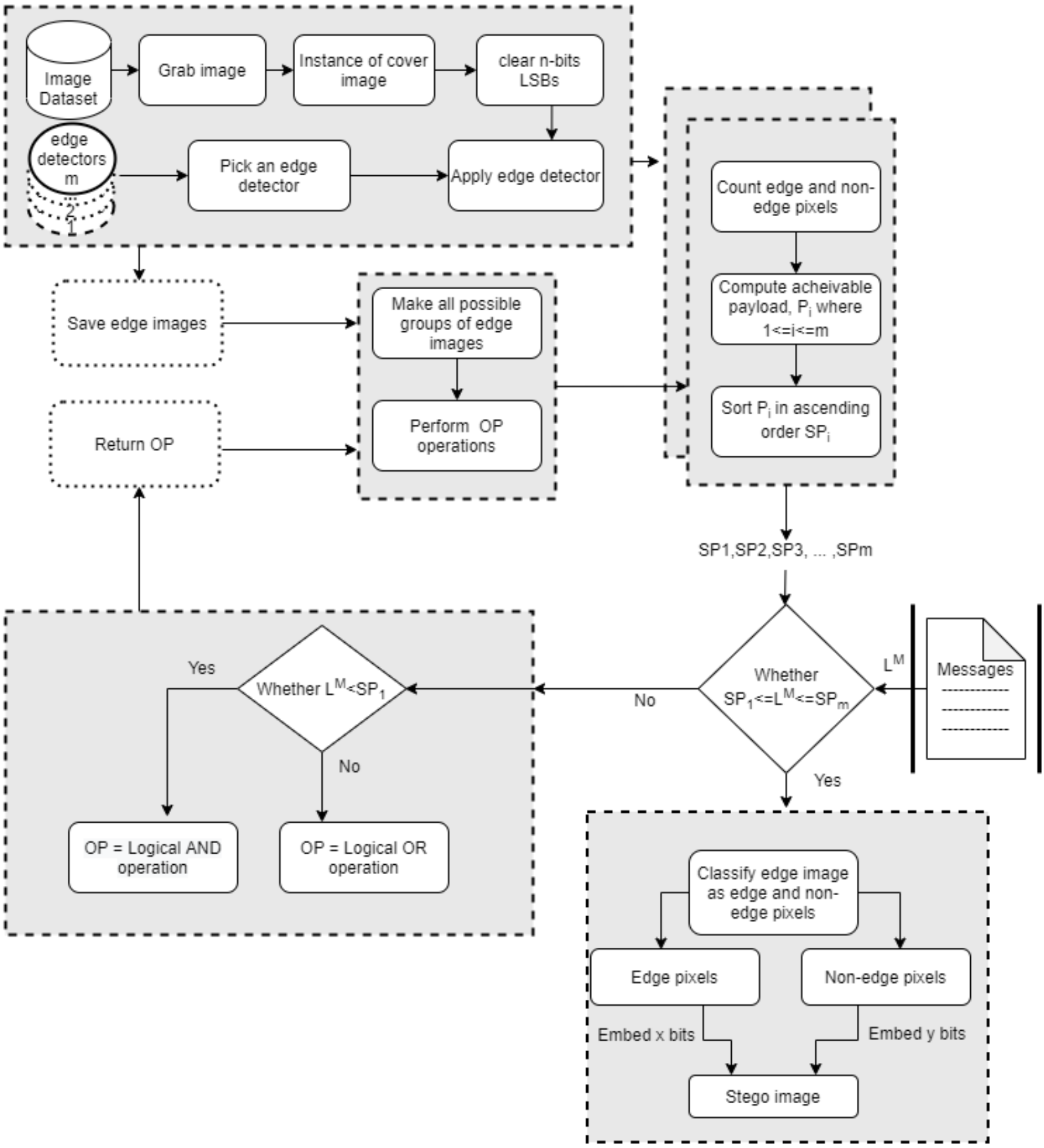


Figure 1: Preprocessing and embedding Phase.

m and ψ returns the edge image eI from I for a specific edge detector Ω. We also make all possible combinations of edge images if needed. We performed logical AND, OR, or dilation operations in each group. Each edge image is a binary image. For each pixel, the edge image holds a 0 or 1. A 1-in-edge image means the corresponding pixel of I is in the detected edge.
We compute the maximum achievable payload P_i by using equation (4). Payload is the total embedded bits.

$$P_i = ePN * x + neN * y; \tag{4}$$

Where ePN is the edge pixel and neN is the non-edge pixels and x and y are the number of embedded bits, $1 <= x <= 5$ and $1 <= y <= 4$.
We then sort P_i in ascending order i.e. $P_1 < P_2 < P_3 < ..... < P_i$. We calculate the total embedded message length m_L by equation (5).

$$m_L = length(message); \tag{5}$$

Next, we check which one first meets the requirement of embedding payload m_L, say P_k using equation (1). P_k is the resultant edge image.
Those are the pre-processing stage. This is illustrated in Figure (1) up to (4) blocks.

3.2 Data embedding phase

Now it is the time to implant secrets in the image I. The pre-processing and embedding process is illustrated in Figure (1). The data implantation steps are as follows:

- An algorithm is developed to classify edge and non-edge pixels and their location in I based on resultant edge image P_k by equation (6).

$$[ePN, ePNP, neN, neNP] = F(I, P_k); \tag{6}$$

Where F returns edge pixels ePN, their positions $ePNP$, non-edge pixels neN and their positions $neNP$ in I. The function F performs in following:

Function $F(I, P_k)$
Compute the size of image I. Let it is (h, w)
$[ePN, ePNP] = G_e(I, P_k, h, w)$
$CP_k = (P_k - 1) * (-1)$
$[neN, neNP] = G_e(I, CP_k, h, w)$
return $[ePN, ePNP, neN, neNP]$

Where G_e is defined below.

Function $G_e(Q, R, h, w)$
$k = 0$
for $i = 1toh$
for $j = 1tow$
if $R(i, j) == 1$
$R_1(k) = Q(i, j)$
$R_2(k, 1) = i, R_2(k, 2) = j$
return R_1, R_2

- Now, we implant x bits and y bits of secrets in each edge and non-edge pixel, respectively of the cover image by the LSB substitution method. Let x bits of information be b_x and y bits of information are b_y. In that, the substitution task is performed by equation (7).

$$\begin{cases} S(s,t) = I(s,t) + \phi(b_x) \\ S(u,v) = I(u,v) + \phi(b_y) \end{cases} \tag{7}$$

where $\phi(b_x)$ stands for decimal conversion of binary b_x. $s = ePNP(i, 1)$, $t = ePNP(i, 2)$, $u = neNP(j, 1)$, $v = neNP(j, 2)$ and $1 \leq i \leq No_O_Edge_Pixels$, $1 \leq j \leq No_Of_nonEdge_Pixels$. Here b_x will be different for each of the s and t. The same is true for b_y. This means that each time a different b_x and b_y of the secret will be implanted. That stego image S is then sent to a receiver end. The receiver end next extracts the implanted secrets from S.

3.3 Data extraction phase

In the extraction phase, the receiver receives the stego image and stego key. The receiver gets the necessary information from the stego key such as the number of cleared bits n, the name of selected edge detectors, the number of bits embedded in edge and non-edge pixels, and message length. Like the sender, the receiver copies the stego image S to I. It then clears n bits of LSBs from I by equation (1). Let, that $n-$LSBs cleared image is also I. The scheme that applies $m-$ number of edge detectors on I from stego key. We have then separated the edge and non-edge pixels and their corresponding positions in I by equation (6). Next from each of the edge and non-edge located pixels, i.e., from (s, t) and (u, v), we have measured d_x and d_y using equations (8) and (9).

$$d_x = S(s,t) - I(s,t); \tag{8}$$

$$d_y = S(u,v) - I(u,v); \tag{9}$$

Here $s, t, u, v, i, and\ j$ are defined in the previous subsection. Next, we extracted the binaries of the secret by equation (10).

$$\begin{cases} b_x = \phi^{-}1(d_x) \\ b_y = \phi^{-}1(d_y) \end{cases} \tag{10}$$

Where $\phi^{-}1$ means binary conversion of decimal value d_x.

4 Result analysis and discussion

In this section, we show the experimental results conducted to evaluate the performance of the proposed scheme with the works of Sultana [32], Bai [5], Rasol [33] and Setiadi[31]. We first selected ten frequently used images, an image dataset, and a message dataset. We set up our experiment and then analyzed the results.

4.1 Experimental Setup

We worked on MATLAB's edition R(2017a) on windows 7. The experiments were performed on a desktop that is specified by an Intel (R) Core (TM) i5-8500T CPU @ 2.10 GHz 2.11 GHz processor and RAM of 8.00 GB. In the proposed system we used two different types of input data one is the secret message, i.e. to be implanted data, and the other is the cover image. We first collected some sample messages from different sources, as shown in Table 2.

The sample message could be a text, binary, or any other format. We used our prepared function ConBin to convert the non-binary input data to binary. For example, text data is converted to binary from the ASCII values of the text contents. We work for different sizes of message lengths. As a cover media, We collected ten frequently used standard images as shown in Figure 2 to conduct all our primary experiments. We also used 499 images of the BOSS dataset. We converted the color of the images to grayscale and resized them to 512 x 512. As the contents of the dataset are images, we worked with intensity values of pixels. Thus the final inputed values to our embedding algorithm are binary for secret message and pixel values for cover media. We measured the performance of the schemes by several feature values, such as edge pixel generation capability, embedding capacity, peak signal-to-noise ratio (PSNR), structural similarity index matrix (SSIM), standard deviation, correlation coefficient, entropy, cosine similarity, and Pixel difference histogram, etc.
Table 2 is given as a message dataset of this work.

Table 2: List of message dataset with message length and type

DatasetName	Message Length	Message Type
SupervisorMessage	398833	text
M1	274661	binary
M2	442483	text
M3	693637	binary

A supervisor Message is a text-type message from my supervisor. The message is
Though the life of a Ph.D. researcher is a matter of struggle, it is enjoyable as well. Bethe cause finding some novelty is always challenging and overcoming such challenge gives a researcher heavenly happiness.....,.

4.2 Mathematical Representation of Feature values

Let the number of edge and non-edge pixels in a cover image are ePN and neN, respectively. Then the maximum achievable payload PL is defined by equation (11)

$$PL = ePN * x + neN * y; \tag{11}$$

Where x and y are the numbers of bits embedded in the edge and the non-edge pixels. We also measured the capacity. Capacity can be defined as the number of implanted bits per pixel. Embedding capacity EC is measured by equation (12)

$$EC = \frac{P}{h * w}; \tag{12}$$

Where, P is the total number of implanted bits in the cover image, h and w are the image height and width. Maintaining image quality is a challenging task and for this purpose, we used PSNR and SSIM which are commonly used image distortion measurement parameters. PSNR is measured by equation (13)

$$PSNR = 10 \log_{10} \frac{255^2}{MSE}; \tag{13}$$

where,

$$MSE = \frac{1}{h * w} \sum_{i=1}^{w} \sum_{j=1}^{h} (S_{i,j} - C_{i,j})^2; \tag{14}$$

here, S is the stego image and C is the original cover image. Next, SSIM is calculated by equation (15)

$$SSIM = \frac{(2\mu_c\mu_s + C_1)(2\sigma_{cs} + C_2)}{(\mu_c^2 + \mu_s^2 + C_1)(\sigma_c^2 + \sigma_s^2 + C_2)} \tag{15}$$

here, μ_c and σ_c are the mean and variance of pixel values in the cover image. Likewise cover, μ_s, and σ_s are the same for the stego image. C_1 and C_2 are two constants and we set $C_1 = 0.0001$ and $C_2 = 0.0009$ for experiment. There are many methods of analyzing the robustness against various attacks. Famous techniques are entropy measurement, standard deviation measurement, analyzing correlation among the pixels, checking the cosine similarity between the cover and stego image, and histogram of the Pixel difference between the stego and cover image. The entropy is measured by equation (16)

$$H = -\sum_{k} P_k \log_2(P_k); \tag{16}$$

Where, P_k is the probability associated with gray value k and $1 \leq k \leq 255$.
Standard deviation is defined by equation (17)

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2}. \tag{17}$$

Where N is the number of data points, x_i each of the values of the data, and $\overline{x}$ is the mean of x_i.
Population correlation is defined by equation (18)

$$P_{cs} = \frac{\sigma_{cs}}{\sigma_c \sigma_s}; \tag{18}$$

Where σ_c and σ_s are population standard deviations in cover C and stego S. Again, σ_{cs} is the co-variance between the cover and stego image. Equation (19) gives us the cosine similarity values

$$f_{\cos sim}(C, S) = \cos\theta = \frac{\sum_{i=1}^{h} \sum_{j=1}^{w} C(i,j) S(i,j)}{\sqrt{\sum_{i=1}^{h} \sum_{j=1}^{w} C(i,j)} \sqrt{\sum_{i=1}^{h} \sum_{j=1}^{w} S(i,j)}}; \tag{19}$$

Where C and S are cover and stego images.

4.3 Experimental results and discussions

In the experiment, we first applied canny, sobel, log, Prewitt, and Roberts edge detectors in five LSBs cleared images to identify edge and non-edge pixels. Canny, sobel, log, Prewitt, and Roberts-based edge detector functions of MATLAB return an edge image for a given input image. The edge image is a binary image. The obtained edge images generated from ten input images, are shown in Table 3. In the previous section, according to the embedding rules we implant x bits of information in edge pixels and y bits of information in non-edge pixels and that represent as a tuple (x, y) where $x > y$. Table 4 summarises the number of edge pixels that were found in ten sample images by different methods. Table 4 provided statistics collected from 5-LSBs cleared images.
We calculated the maximum achievable payload of each edge image

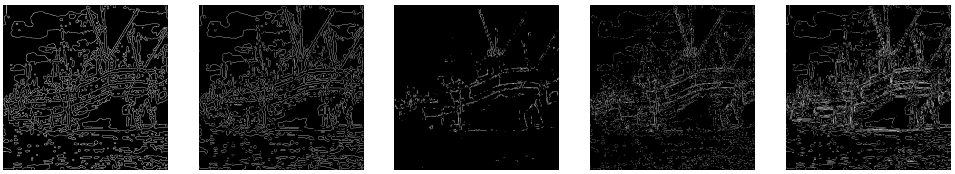


Figure 2: Sample cover images.



Figure 3: Stego images for the cover images of Figure 2.

Table 3: Edge images generated from ten cover images. The images were formed for Canny, Log, Prewitt, Sobel, and Robert edge detectors from 5-LSBs cleared images (n=5).

ImageName	Methods				
	Canny	Log	Prewitt	Sobel	Roberts
F16.jpg					
babon.jpg					
basket.jpg					
boat.jpg					
brbra.jpg					
lena.jpg					
livingroom.jpg					
pepper.jpg					
walkbridge.jpg					
wheel.jpg					

which is shown in Figure 4. That figure states that canny is the highest and Prewitt is the lowest value of the maximum achievable payload. At the same time, for the experimental purpose, we take four sample messages and the length are 398833 bits, 274661 bits, 442483 bits, and 693637 bits. We demonstrate various experimental results for those messages.

We analyzed the performance in embedding capacity as well. Embedding capacity is graphically shown in Figure 5 for 499 images of the BOSS dataset.

We also analyzed the visual quality and structural originality of stego images. Visual quality is measured by PSNR values and is sketched in Figure 6 for a message length of 398833 bits. PSNR value in our scheme is higher than the others which are 37.45%, 46.87%, 44.21%, and 55.56%. It is clear from the diagram that the proposed scheme has a higher PSNR value than the competing schemes. The structural similarity index value, SSIM, for message length = 398833 bits is listed in Table 5. The table confirms that the SSIM values of all the schemes are both high and very close to each other.

Table 7 and Table 8 contain the values of PSNR and SSIM for message lengths 274661 bits, 442483 bits, and 693637 bits. We analyzed the time complexity of diverse schemes by measuring their required data embedment times for a message length of 398833 bits. The time complexity of the schemes is measured experimentally and tabulated in Table 6.

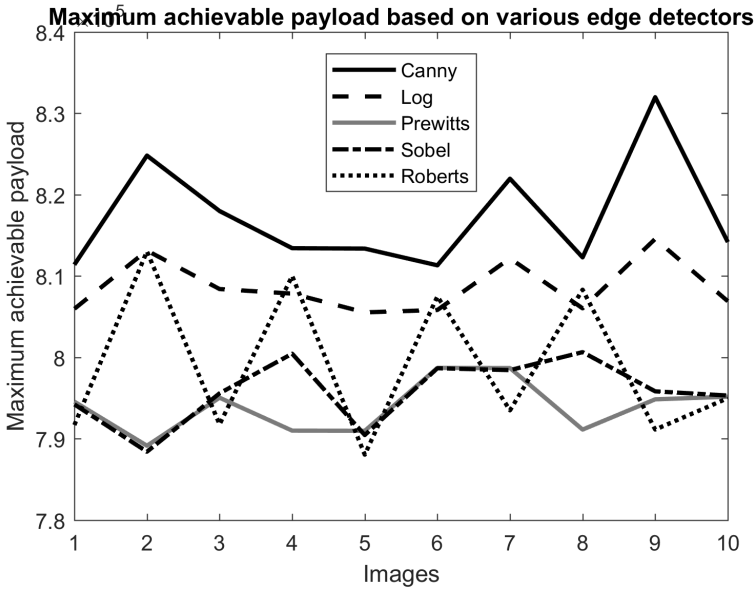


Figure 4: Maximum achievable payload based on diverse edge detectors where x = 4 and y = 3. The figure states that canny has the highest achievable payload than log, Prewitt's, sobel, and Roberts.

Table 4: A comparison of the number of edge pixels on various edge detectors for cleared images (n=5).

ImageName	Methods				
	Canny	Log	Prewitt	Sobel	Roberts
F16.jpg	24966	19532	8119	7786	5239
babon.jpg	38383	26700	2700	1966	26603
basket.jpg	31560	21967	8603	9145	5323
boat.jpg	26991	21406	4555	14018	23576
brbra.jpg	26941	19099	4536	4010	1595
lena.jpg	24884	19391	12270	12229	21061
livingroom.jpg	35543	25742	12253	11998	7049
pepper.jpg	25860	19595	4686	14211	21890
walkbridge.jpg	45563	28134	8405	9388	4685
wheel.jpg	27745	20467	8726	8870	8544

Table 5: SSIM values when message length = 398833 bits.

ImageName	SSIM values				
	Proposed	[32]	[5]	[33]	[31]
FF16.jpg	0.993	0.992	0.975	0.984	0.961
babon.jpg	0.972	0.971	0.924	0.946	0.911
basket.jpg	0.997	0.995	0.987	0.989	0.979
boat.jpg	0.993	0.962	0.964	0.980	0.948
brbra.jpg	0.994	0.995	0.966	0.986	0.945
lena.jpg	0.987	0.954	0.953	0.973	0.928
livingroom.jpg	0.997	0.994	0.986	0.988	0.978
pepper.jpg	0.990	0.950	0.954	0.979	0.933
walkbridge.jpg	0.998	0.996	0.986	0.989	0.984
wheel.jpg	0.996	0.994	0.989	0.990	0.982

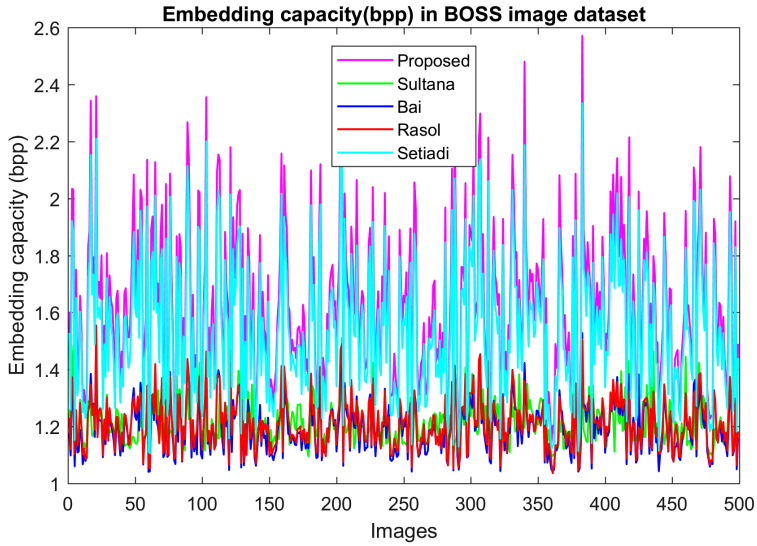


Figure 5: Performance comparison of the proposed scheme with Sultana [32], Bai [5], Rasol [33], and Setiadi [31] in terms of capacity in the BOSS image dataset.

Table 6: Elapsed time for message length = 398833 bits.

ImageName	Elapsed time				
	Proposed	[32]	[5]	[33]	[31]
F16.jpg	13.32	13.32	13.32	13.47	12.20
babon.jpg	13.45	13.23	13.66	13.75	9.89
basket.jpg	13.69	13.38	13.61	13.59	11.01
boat.jpg	13.28	13.41	13.50	13.68	11.18
brbra.jpg	13.18	13.00	13.49	13.32	11.37
lena.jpg	13.19	13.43	13.52	13.45	11.51
livingroom.jpg	13.48	13.39	13.58	13.72	9.72
pepper.jpg	14.38	13.39	13.47	13.47	11.26
walkbridge.jpg	13.60	13.27	13.81	13.85	9.44
wheel.jpg	13.49	13.42	13.43	13.46	11.64

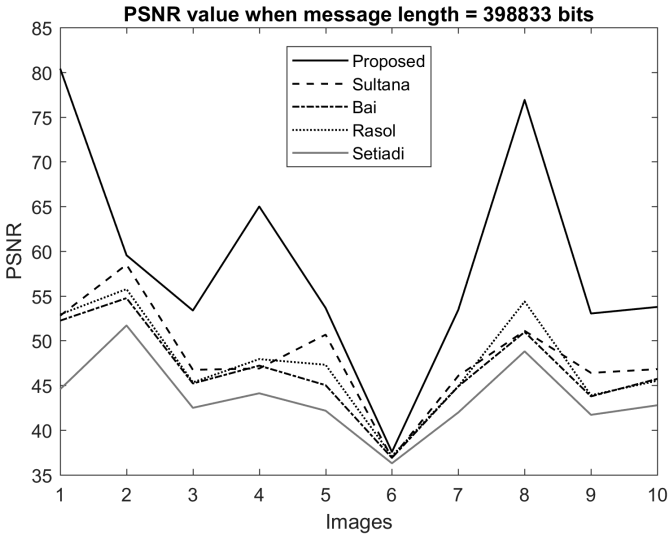


Figure 6: Performance comparison of the proposed scheme with Sultana [32], Bai [5], Rasol [33], and Setiadi [31] in terms of PSNR when message length = 398833 bits. The figure states that the proposed scheme dominates the other competing schemes.

Table 7: SSIM values of different schemes

Methods	ImageName	Message length (bits)		
		274661	442483	693637
Proposed method	F16.jpg	0.994	0.967	0.948
	boat.jpg	0.991	0.953	0.924
	pepper.jpg	0.988	0.941	0.905
[32]	F16.jpg	0.993	0.992	0.992
	boat.jpg	0.972	0.962	0.962
	pepper.jpg	0.962	0.950	0.950
[5]	F16.jpg	0.978	0.975	0.975
	boat.jpg	0.972	0.964	0.964
	pepper.jpg	0.963	0.954	0.954
[33]	F16.jpg	0.987	0.984	0.984
	boat.jpg	0.985	0.980	0.980
	pepper.jpg	0.982	0.979	0.979
[31]	F16.jpg	0.969	0.957	0.955
	boat.jpg	0.961	0.940	0.929
	pepper.jpg	0.953	0.924	0.913

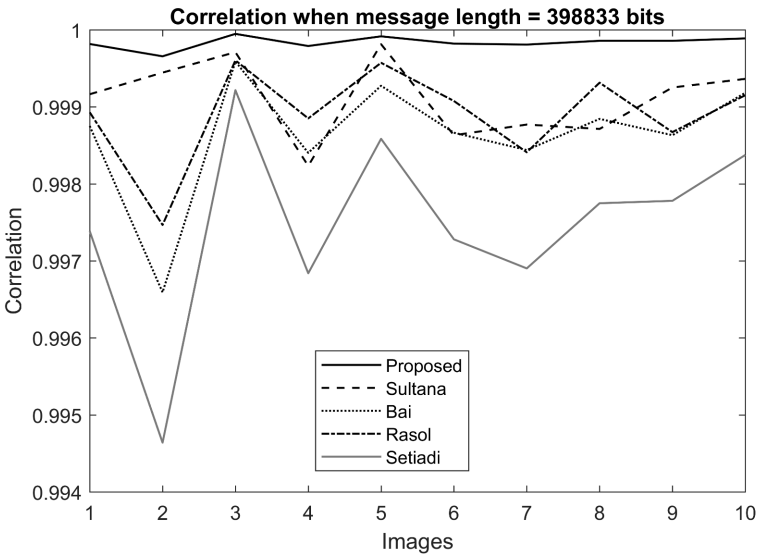


Figure 7: Performance comparison of the proposed scheme with Sultana [32], Bai [5], Rasol [33], and Setiadi [31] in terms of Correlation coefficients when message length = 398833 bits.

Table 8: PSNR values of different schemes

Methods	ImageName	Message length (bits)		
		274661	442483	693637
Proposed	F16.jpg	62.47	51.19	42.69
	boat.jpg	54.96	46.28	41.79
	pepper.jpg	57.92	50.40	46.89
[32]	F16.jpg	53.12	52.78	52.78
	boat.jpg	48.00	46.88	46.88
	pepper.jpg	52.58	51.08	51.08
[5]	F16.jpg	52.95	52.23	52.23
	boat.jpg	48.16	47.20	47.20
	pepper.jpg	52.17	50.96	50.96
[33]	F16.jpg	53.63	52.93	52.93
	boat.jpg	48.73	47.93	47.93
	pepper.jpg	55.10	54.37	54.37
[31]	F16.jpg	46.37	44.16	43.91
	boat.jpg	46.11	43.73	43.15
	pepper.jpg	50.69	48.13	47.51

Table 9 contains the time complexity for message lengths 274661 bits, 442483 bits, and 693637 bits.

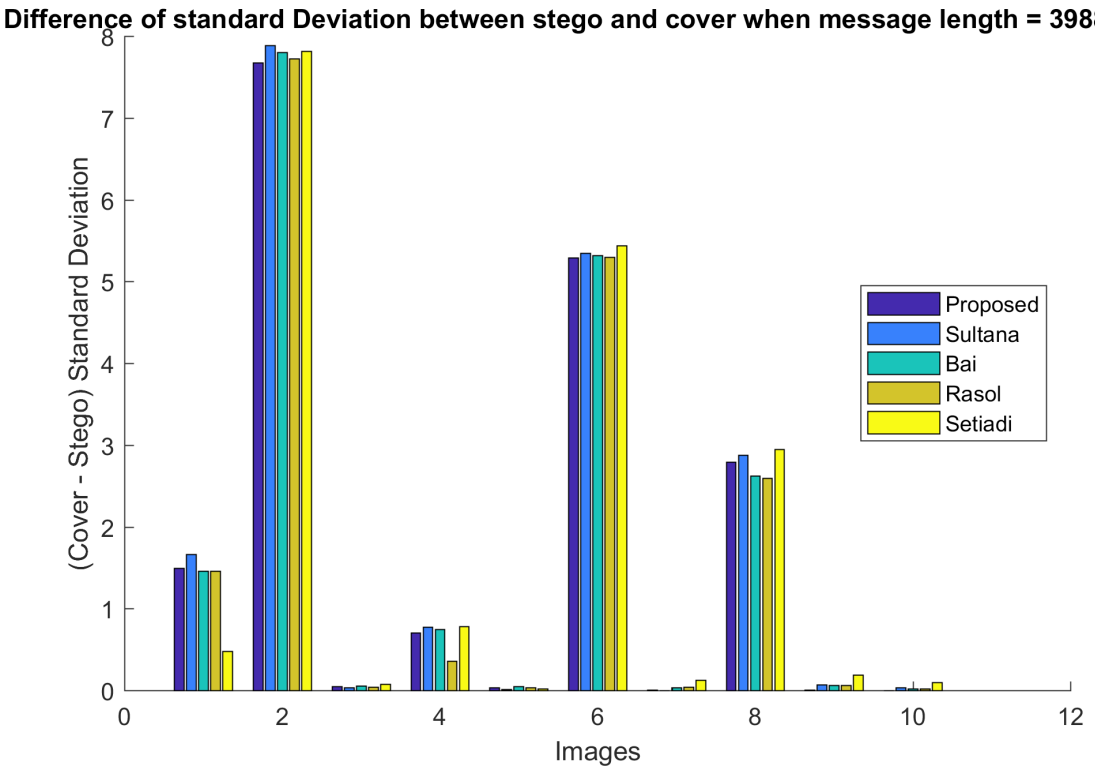


Figure 8: Difference of standard Deviations of cover and stego images when message length = 398833 bits. The figure states that the differences are very small and close to each other.

Table 9: Elapsed time of different schemes

Methods	ImageName	Message length (bits)		
		274661	442483	693637
Proposed method	F16.jpg	13.18	13.48	14.12
	boat.jpg	12.95	13.58	15.07
	pepper.jpg	12.90	14.16	14.69
[32]	F16.jpg	11.72	13.69	13.27
	boat.jpg	10.29	13.33	13.38
	pepper.jpg	10.51	14.13	13.68
[5]	F16.jpg	10.86	13.41	13.54
	boat.jpg	10.68	13.42	13.33
	pepper.jpg	11.17	13.65	13.59
[33]	F16.jpg	11.47	13.52	13.35
	boat.jpg	11.21	13.52	13.38
	pepper.jpg	11.53	13.42	13.32
[31]	F16.jpg	8.59	13.36	13.91
	boat.jpg	7.66	12.04	13.88
	pepper.jpg	7.23	12.78	14.25

Table 10: Correlation values of different schemes

Methods	ImageName	Message length (bits)		
		274661	442483	693637
Proposed method	F16.jpg	0.999	0.998	0.996
	boat.jpg	0.999	0.997	0.995
	pepper.jpg	0.999	0.998	0.996
[32]	F16.jpg	0.999	0.999	0.999
	boat.jpg	0.998	0.998	0.998
	pepper.jpg	0.999	0.998	0.998
[5]	F16.jpg	0.998	0.998	0.998
	boat.jpg	0.998	0.998	0.998
	pepper.jpg	0.999	0.998	0.998
[33]	F16.jpg	0.999	0.998	0.998
	boat.jpg	0.999	0.998	0.998
	pepper.jpg	0.999	0.999	0.999
[31]	F16.jpg	0.998	0.997	0.996
	boat.jpg	0.997	0.996	0.996
	pepper.jpg	0.998	0.997	0.997

Table 11: Standard deviations values of different schemes

Methods	ImageName	Message length (bits)		
		274661	442483	693637
Proposed method	F16.jpg	1.49	1.45	0.38
	boat.jpg	0.71	0.76	0.50
	pepper.jpg	2.58	2.65	2.77
[32]	F16.jpg	1.65	1.66	1.66
	boat.jpg	0.74	0.77	0.77
	pepper.jpg	2.86	2.87	2.87
[5]	F16.jpg	1.46	1.46	1.46
	boat.jpg	0.72	0.74	0.74
	pepper.jpg	2.61	2.62	2.62
[33]	F16.jpg	1.46	1.46	1.46
	boat.jpg	0.34	0.35	0.35
	pepper.jpg	2.59	2.59	2.59
[31]	F16.jpg	0.53	0.47	0.47
	boat.jpg	0.85	0.80	0.83
	pepper.jpg	2.93	2.93	2.97

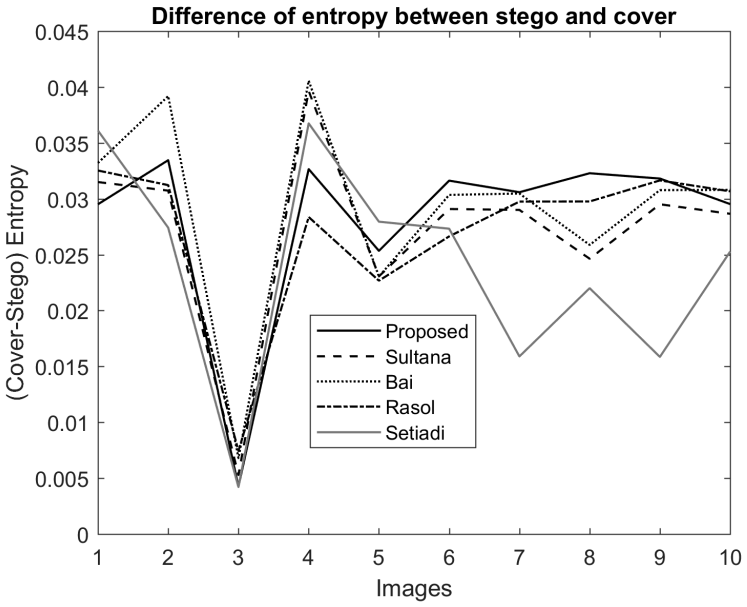


Figure 10: Performance comparison of the proposed scheme with Sultana [32], Bai [5], Rasol [33], and Setiadi [31] in terms of entropy values when message length = 398833 bits.

5 Robustness of the proposed scheme against attacks

We statistically analyzed our scheme using various parameters such as correlation coefficient, standard deviation, entropy, cosine similarities, and pixel difference histogram to check its robustness against various attacks, We first measured correlation coefficients $\rho_s C$ between the cover and stego image for message length 398833 bits. $\rho_s C = 0$ stands for no relationship between two images. $\rho_s C > 0$ means a positive correlation between the cover and stego image and lies perfect relationship when it reaches 1. Similarly, a negative value of $\rho_s C$ indicates a negative relationship. Results of $\rho_s C$ are depicted in Figure 7. Through proposed method shows a higher correlation value, its difference from others is insignificance.

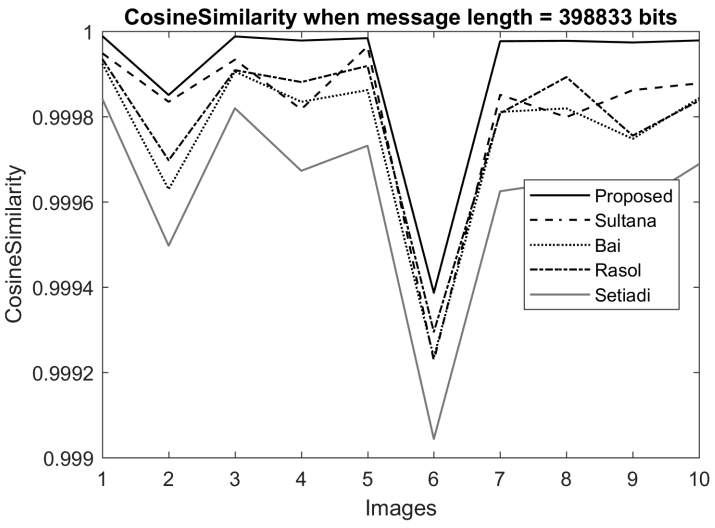


Figure 9: Performance comparison of the proposed scheme with Sultana [32], Bai [5], Rasol [33], and Setiadi [31]in terms of Cosine similarity when message length = 398833 bits.

Table 12: Cosine similarity values of different schemes

Methods	ImageName	Message length (bits)		
		274661	442483	693637
Proposed method	F16.jpg	0.9999	0.9998	0.9997
	boat.jpg	0.9999	0.9997	0.9995
	pepper.jpg	0.9999	0.9997	0.9995
[32]	F16.jpg	0.9999	0.9999	0.9999
	boat.jpg	0.9998	0.9998	0.9998
	pepper.jpg	0.9998	0.9997	0.9997
[5]	F16.jpg	0.9999	0.9999	0.9999
	boat.jpg	0.9998	0.9998	0.9998
	pepper.jpg	0.9998	0.9998	0.9998
[33]	F16.jpg	0.9999	0.9999	0.9999
	boat.jpg	0.9999	0.9998	0.9998
	pepper.jpg	0.9999	0.9998	0.9998
[31]	F16.jpg	0.9998	0.9998	0.9998
	boat.jpg	0.9997	0.9996	0.9995
	pepper.jpg	0.9997	0.9996	0.9995

Rather, as with others, it represents a higher correlation between cover and stego image. We also measured the standard deviation of pixel values from, their mean, separately in cover and stego image is σ_c and σ_s for message length 398833 bits. We then calculated their difference by $sigma_{=}\sigma_c - \sigma_s$. Ideally, σ_d should be zero for a non-tempered image. The results are shown in Figure 8.

Table 10 and Table 11 contains the values of correlation coefficients and standard deviations for message length 274661 bits, 442483 bits, and 693637 bits.

To verify further with similar statistics, we measured cosine similarities between the cover and stego images for a message length of 398833 bits. That value is 1 for two identical images and 0 for two fully mismatched images. The results are demonstrated in Figure 9.

The figure illustrates that our proposed method shows higher values than the other competing schemes.
We computed, the entropy values H as well in cover and stego images for message length 398833 bits. Next, we calculated their differences. That difference value is zero for two identical images. Results are plotted in Figure 10. The figure shows that none of the results are greater than 0.04, i.e., these are very small and close to zero.

We also used, another statistical tool pixel difference histogram (PDH) to identify the stego images. The PDH of the original images and corresponding stego images are shown in Figure 11 and Figure 12 for message length 398833 bits.
Thus, it can be deduced from the results of these experiments that our method is strong enough to protect against attacks on implanted data.

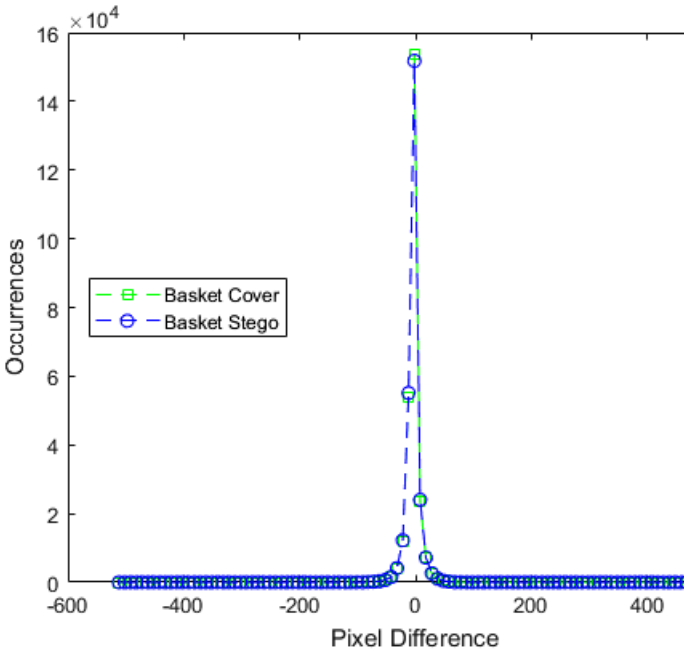


Figure 11: Performance comparison of the proposed scheme with Sultana, Bai, Rasol, and Setiadi in terms of pixel difference histogram when message length = 398833 bits for basket cover image.

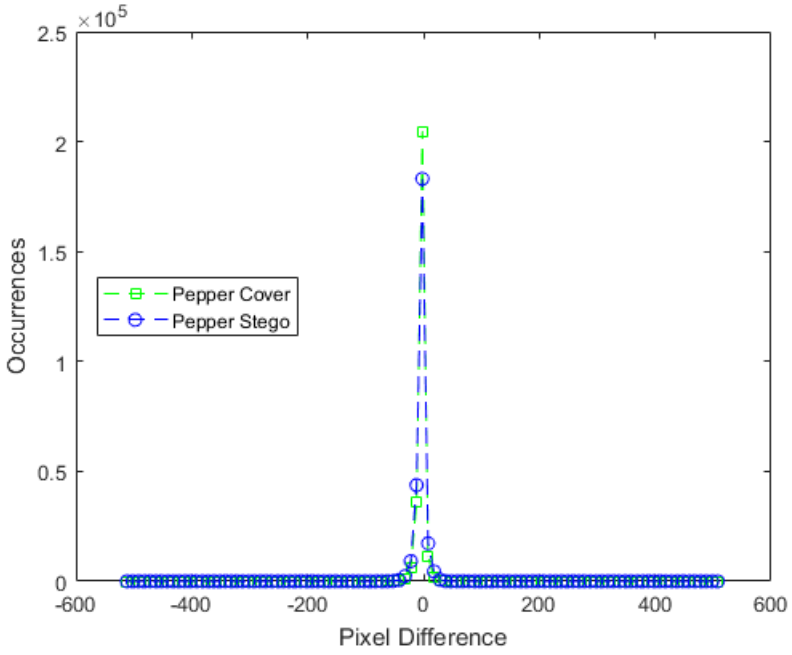


Figure 12: Performance comparison of the proposed scheme with Sultana, Bai, Rasol, and Setiadi in terms of pixel difference histogram when message length = 398833 bits for peppers cover image.

Table 12 and Table 13 contains the values of cosine similarity and entropy for message length 274661 bits, 442483 bits, and 693637 bits.

Table 13: Entropy values of different schemes

Methods	ImageName	Message length (bits)		
		274661	442483	693637
Proposed method	F16.jpg	0.027	0.040	0.052
	boat.jpg	0.030	0.046	0.068
	pepper.jpg	0.028	0.029	0.046
[32]	F16.jpg	0.024	0.031	0.031
	boat.jpg	0.020	0.039	0.039
	pepper.jpg	0.011	0.024	0.024
[5]	F16.jpg	0.022	0.033	0.033
	boat.jpg	0.023	0.040	0.040
	pepper.jpg	0.014	0.025	0.025
[33]	F16.jpg	0.022	0.032	0.032
	boat.jpg	0.019	0.028	0.028
	pepper.jpg	0.021	0.029	0.029
[32]	F16.jpg	0.021	0.043	0.048
	boat.jpg	0.019	0.048	0.069
	pepper.jpg	0.0008	0.032	0.044

6 Conclusion

The edge-detection-based embedding schemes, generally, implant either in edge pixels only or a different number of bits in edge and non-edge pixels. Few of them use multiple edge detectors and hybridize them to increase the number of edge pixels in the resultant edge image. Even, either maintaining the visual quality of stego image or meeting the embedding demand is still functioning as a challenging matter. This research overcomes that problem by realizing the situation and then dynamically choosing the number of implanted bits per pixel, the required edge detectors, and the best hybridization technique.

Edge detectors are selected and their hybridization is performed based on message length. This scheme, thus, increases or decreases the number of edge pixels and fixes the number of implanted bits per edge and non-edge pixels according to the embedding demand. The experimental result deduces that this scheme embeds 92.37%, 73.92%, 74.78%, and 9.60% more bits than [32], [5], [33], [31], respectively. Similarly, for a small embedding demand, the proposed scheme demonstrates 37.45%, 46.87%, 44.21%, and 55.56% higher PSNR values than [32], [5], [33], [31], respectively. Moreover, the statistical analyses state that this scheme exhibits stronger security against attacks.

In our future work, we wish to apply this high-performing embedding scheme in electronic kit development for use in healthcare and forensic-related applications. We also plan to apply a machine learning algorithm for segmenting the image areas into different complex levels so that we can implant a different number of bits in those staged areas, i.e. levels.

Conflict of Interest There is nothing related to conflict of interest.

Acknowledgment The first author is a research fellow of the ICT division of the Ministry of Post, Telecommunication, and Information Technology of the Government of Bangladesh. Therefore, we want to devote a thank to the concerned ministry, and at the same time, we would like to acknowledge that support.

References

[1] C. Vanmathi, S. Prabu, "Image steganography using fuzzy logic and chaotic for large payload and high imperceptibility," International Journal of Fuzzy Systems, **20**(2), 460–473, 2018, doi:10.1007/s40815-017-0420-0.

[2] E. J. Kusuma, O. R. Indriani, C. A. Sari, E. H. Rachmawanto, et al., "An imperceptible LSB image hiding on edge region using DES encryption," in 2017 International Conference on Innovative and Creative Information Technology (ICITech), 1–6, IEEE, 2017, doi:10.1109/INNOCIT.2017.8319132.

[3] M. Hussain, A. W. A. Wahab, N. Javed, K.-H. Jung, "Recursive information hiding scheme through LSB, PVD shift, and MPE," IETE Technical Review, **35**(1), 53–63, 2018, doi:10.1080/02564602.2016.1244496.

[4] W.-J. Chen, C.-C. Chang, T. H. N. Le, "High payload steganography mechanism using hybrid edge detector," Expert Systems with applications, **37**(4), 3292–3301, 2010, doi:10.1016/j.eswa.2009.09.050.

[5] J. Bai, C.-C. Chang, T.-S. Nguyen, C. Zhu, Y. Liu, "A high payload steganographic algorithm based on edge detection," Displays, **46**, 42–51, 2017, doi:10.1016/j.displa.2016.12.004.

[6] H. Sultana, A. Kamal, "An Edge Detection Based Reversible Data Hiding Scheme," in 2022 IEEE Delhi Section Conference (DELCON), 1–6, IEEE, 2022, doi:10.1109/DELCON54057.2022.9753404.

[7] P. W. Adi, F. Z. Rahmanti, N. A. Abu, "High quality image steganography on integer Haar Wavelet Transform using modulus function," in 2015 International Conference on Science in Information Technology (ICSITech), 79–84, IEEE, 2015, doi:10.1109/ICSITech.2015.7407781.

[8] H. Al-Dmour, A. Al-Ani, "A steganography embedding method based on edge identification and XOR coding," Expert systems with Applications, **46**, 293–306, 2016, doi:10.1016/j.eswa.2015.10.024.

[9] A. Kamal, M. M. Islam, "A prediction error based histogram association and mapping technique for data embedment," Journal of Information Security and Applications, **48**, 102368, 2019, doi:10.1016/j.jisa.2019.102368.

[10] H. Sultana, A. Kamal, M. M. Islam, "Enhancing the Robustness of Visual Degradation Based HAM Reversible Data Hiding." J. Comput. Sci., **12**(2), 88–97, 2016, doi:10.3844/jcssp.2016.88.97.

[11] A. Kamal, M. M. Islam, "Enhancing embedding capacity and stego image quality by employing multi predictors," Journal of Information Security and Applications, **32**, 59–74, 2017, doi:10.1016/j.jisa.2016.08.005.

[12] A. Kamal, M. M. Islam, "Boosting up the data hiding rate through multi cycle embedment process," Journal of Visual Communication and Image Representation, **40**, 574–588, 2016, doi:10.1016/j.jvcir.2016.07.023.

[13] A. Kamal, M. M. Islam, "Capacity improvement of reversible data hiding scheme through better prediction and double cycle embedding process," in 2015 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), 1–6, IEEE, 2015, doi:10.1109/ANTS.2015.7413636.

[14] A. Kamal, M. M. Islam, "An image distortion-based enhanced embedding scheme," Iran Journal of Computer Science, **1**(3), 175–186, 2018, doi:10.1007/s42044-018-0016-3.

[15] A. Kamal, M. M. Islam, Z. Islam, "An embedding technique for smartcard-supported e-healthcare services," Iran Journal of Computer Science, **3**(4), 195–205, 2020, doi:10.1007/s42044-020-00055-1.

[16] A. Kamal, M. M. Islam, "Enhancing the embedding payload by handling the affair of association and mapping of block pixels through prediction errors histogram," in 2016 International Conference on Networking Systems and Security (NSysS), 1–8, IEEE, 2016, doi:10.1109/NSysS.2016.7400691.

[17] A. Kamal, M. M. Islam, "Enhancing the performance of the data embedment process through encoding errors," Electronics, **5**(4), 79, 2016, doi:10.3390/electronics5040079.

[18] A. Kamal, M. M. Islam, "Facilitating and securing offline e-medicine service through image steganography," Healthcare Technology Letters, **1**(2), 74–79, 2014, doi:10.1049/htl.2013.0026.

[19] W. Hong, "Adaptive reversible data hiding method based on error energy control and histogram shifting," Optics Communications, **285**(2), 101–108, 2012, doi:10.1016/j.optcom.2011.09.005.

[20] H. Yao, C. Qin, Z. Tang, Y. Tian, "Improved dual-image reversible data hiding method using the selection strategy of shiftable pixels' coordinates with minimum distortion," Signal Processing, **135**, 26–35, 2017, doi:10.1016/j.sigpro.2016.12.029.

[21] S. Yi, Y. Zhou, "Binary-block embedding for reversible data hiding in encrypted images," Signal Processing, **133**, 40–51, 2017, doi:10.1016/j.sigpro.2016.10.017.

[22] S. Khan, N. Ahmad, M. Ismail, N. Minallah, T. Khan, "A secure true edge based 4 least significant bits steganography," in 2015 International Conference on Emerging Technologies (ICET), 1–4, IEEE, 2015, doi:10.1109/ICET.2015.7389227.

[23] G. Swain, "Adaptive pixel value differencing steganography using both vertical and horizontal edges," Multimedia Tools and Applications, **75**(21), 13541–13556, 2016, doi:10.1007/s11042-015-2937-2.

[24] S. Sun, "A novel edge based image steganography with 2k correction and Huffman encoding," Information Processing Letters, **116**(2), 93–99, 2016, doi:10.1016/j.ipl.2015.09.016.

[25] C.-F. Lee, C.-Y. Weng, K.-C. Chen, "An efficient reversible data hiding with reduplicated exploiting modification direction using image interpolation and edge detection," Multimedia Tools and Applications, **76**(7), 9993–10016, 2017, doi:10.1007/s11042-016-3591-z.

[26] H.-W. Tseng, H.-S. Leng, "High-payload block-based data hiding scheme using hybrid edge detector with minimal distortion," IET Image Processing, **8**(11), 647–654, 2014, doi:10.1049/iet-ipr.2013.0584.

[27] S. Kumar, A. Singh, M. Kumar, "Information hiding with adaptive steganography based on novel fuzzy edge identification," Defence Technology, **15**(2), 162–169, 2019, doi:10.1016/j.dt.2018.08.003.

[28] C.-F. Lee, C.-C. Chang, X. Xie, K. Mao, R.-H. Shi, "An adaptive high-fidelity steganographic scheme using edge detection and hybrid hamming codes," Displays, **53**, 30–39, 2018, doi:10.1016/j.displa.2018.06.001.

[29] S. K. Ghosal, A. Chatterjee, R. Sarkar, "Image steganography based on Kirsch edge detection," Multimedia Systems, **27**(1), 73–87, 2021, doi:10.1007/s00530-020-00703-3.

[30] B. Vishnu, L. V. Namboothiri, S. R. Sajeesh, "Enhanced image steganography with PVD and edge detection," in 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 827–832, IEEE, 2020, doi:10.1109/ICCMC48092.2020.ICCMC-000153.

[31] D. R. I. M. Setiadi, "Improved payload capacity in LSB image steganography uses dilated hybrid edge detection," 2022, doi:10.1016/j.jksuci.2019.12.007.

[32] H. Sultana, A. Kamal, "Image Steganography System based on Hybrid Edge Detector," in 2021 24th International Conference on Computer and Information Technology (ICCIT), 1–6, IEEE, 2021, doi:10.1109/ICCIT54785.2021.9689777.

[33] D. R. I. M. Setiadi, J. Jumanto, "An enhanced LSB-image steganography using the hybrid Canny-Sobel edge detection," Cybernetics and Information Technologies, **18**(2), 74–88, 2018, doi:10.2478/cait-2018-0029.

[34] K. M. Sagayam, A. A. Anton Jone, K. Cengiz, L. Rajesh, A. A. Elngar, "Breast Cancer Detection based on 3-D Mammography Images using Deep Learning Strategies," Journal of Information Technology Management, **14**(4), 2–18, 2022, doi:10.22059/JITM.2022.88132.

[35] P. Elayaraja, S. Kumarganesh, K. Martin Sagayam, H. Dang, M. Pomplun, "An efficient approach for detection and classification of cancer regions in cervical images using optimization based CNN classification approach," Journal of Intelligent & Fuzzy Systems, (Preprint), 1–11, 2022, doi:10.3233/JIFS-212871.

[36] R. Jayasingh, J. K. RJS, D. B. Telagathoti, K. M. Sagayam, S. Pramanik, O. P. Jena, S. K. Bandyopadhyay, "Speckle noise removal by SORAMA segmentation in Digital Image Processing to facilitate precise robotic surgery," International Journal of Reliable and Quality E-Healthcare (IJRQEH), **11**(1), 1–19, 2022, doi:10.4018/IJRQEH.295083.

[37] T. B. Mary, P. M. Bruntha, M. Manimekalai, K. M. Sagayam, H. Dang, et al., "Investigation of an Efficient Integrated Semantic Interactive Algorithm for Image Retrieval," Pattern Recognition and Image Analysis, **31**(4), 709–721, 2021, doi:10.1134/S1054661821040234.

[38] A. D. Andrushia, K. M. Sagayam, H. Dang, M. Pomplun, L. Quach, "Visual-Saliency-Based Abnormality Detection for MRI Brain Images—Alzheimer's Disease Analysis," Applied Sciences, **11**(19), 9199, 2021, doi:10.3390/app11199199.

[39] J. Andrew, T. Mhatesh, R. D. Sebastin, K. M. Sagayam, J. Eunice, M. Pomplun, H. Dang, "Super-resolution reconstruction of brain magnetic resonance images via lightweight autoencoder," Informatics in Medicine Unlocked, **26**, 100713, 2021, doi:10.1016/j.imu.2021.100713.

[40] J. A. Onesimu, J. Karthikeyan, D. S. J. Viswas, R. D. Sebastian, "Security and Privacy Challenges of Deep Learning: A Comprehensive Survey," Research Anthology on Privatizing and Securing Data, 1258–1280, 2021, doi:10.4018/978-1-7998-8954-0.ch059.

[41] J. A. Onesimu, J. Karthikeyan, Y. Sei, "An efficient clustering-based anonymization scheme for privacy-preserving data collection in IoT based healthcare services," Peer-to-Peer Networking and Applications, **14**(3), 1629–1649, 2021, doi:10.1007/s12083-021-01077-7.

[42] Y. Sei, J. Andrew, H. Okumura, A. Ohsuga, "Privacy-Preserving Collaborative Data Collection and Analysis with Many Missing Values," IEEE Transactions on Dependable and Secure Computing, 2022, doi:10.1109/TDSC.2022.3174887.

[43] A. Jan, S. A. Parah, M. Hassan, B. A. Malik, "Realization of Efficient Steganographic Scheme Using Hybrid Edge Detection and Chaos," Arabian Journal for Science and Engineering, 1–14, 2022, doi:10.1007/s13369-022-06960-w.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 4, 72-78 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Machine Learning Model Selection considering Tradeoffs between Accuracy and Interpretability

Zhumakhan Nazir*, Temirlan Zarymkanov, Jurn-Guy Park

School of Engineering and Digital Sciences, Computer Science Department, Nazarbayev University, Nur-Sultan, 010000, Kazakhstan

ARTICLE INFO

Article history:
Received: 31 May, 2022
Accepted: 14 July, 2022
Online: 16 August, 2022

Keywords:
Interpretable Machine Learning (IML)
Explainable ML
Model Selection techniques

ABSTRACT

Applying black-box ML models in high-stakes fields like criminology, healthcare and real-time operating systems might create issues because of poor interpretability and complexity. Also, model building methods that include interpretability is now one of the growing research topics due to the absence of interpretability metrics that are both model-agnostic and quantitative. This paper introduces model selection methods with trade off between interpretability and accuracy of a model. Our results show 97% improvement in interpretability with 2.5% drop in accuracy in AutoMPG dataset using MLP model (65% improvement in interpretability with 1.5% drop in accuracy in MNIST dataset).

1. Introduction

This paper is an extension of the work originally presented in ICITEE 2021 with 1) addition of classification problems and 2) more clarified outcomes (i.e. graphs, tables) [1].

ML models are widely used in various fields including public health and the judicial system. However, the majority of the state-of-the-art estimators could be categorized as 'black-box' models with poor accountability and transparency [2]. For instance, the CNN model learned to detect metal token on the corner of the radiology image instead of the image itself (no accountability) [3]; because the model is black-box it is hard to notice such behavior (no transparency).

In [4], research interest in interpretability in model building is rising. Unfortunately, because of the absence of quantitative assessment metrics, evaluating interpretability is not a trivial goal. According to [5], [6], interpretability is inversely proportional to accuracy. Therefore, one realistic approach is to trade accuracy for interpretability, specifically, is it possible to create simpler (easily interpretable) models with high enough accuracy (drop in accuracy to a certain threshold)?

To address the above problem, we acquire a simple and effective numerical interpretability metric-simulatability operation count (SOC) [7], following the major contributions of this work:

- Evaluate interpretability and accuracy of commonly used models for regression and classification tasks: tree-based models, multi-layer perceptron (MLP) and support vector machine (SVM).
- Propose and apply methodology for a trade-off between interpretability and accuracy to enhance interpretability of the models, by letting accuracy to drop up to certain limits.

2. Motivation and Related work

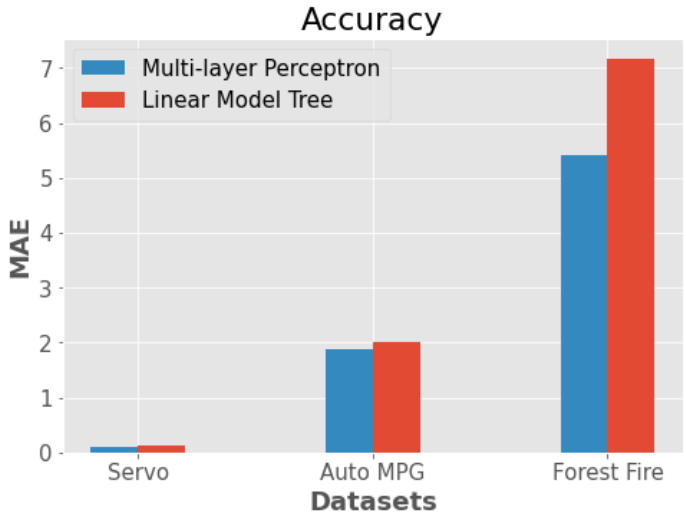
2.1. Motivation

Even though supremacy of black-box models led to their extensive usage, they have lower interpretability compared to tree-based models (e.g., linear model tree), which can compete with other models on both regression and classification tasks. Complexity of black box models result in higher accuracy in general. However, they have lower interpretability with respect to tree-based models (e.g. decision tree) which can achieve competitive performance on both classification and regression tasks. As depicted in Figure 1a, the linear model tree (LMT), compared to MLP regression, has almost the same accuracy results (MAE) on AutoMPG and Servo datasets, and worse results (higher MAE) on Forest Fire dataset. While the interpretability level of LMT is remarkably higher (lower SOC) than MLP in the AutoMPG and Servo datasets (Figure 1b).

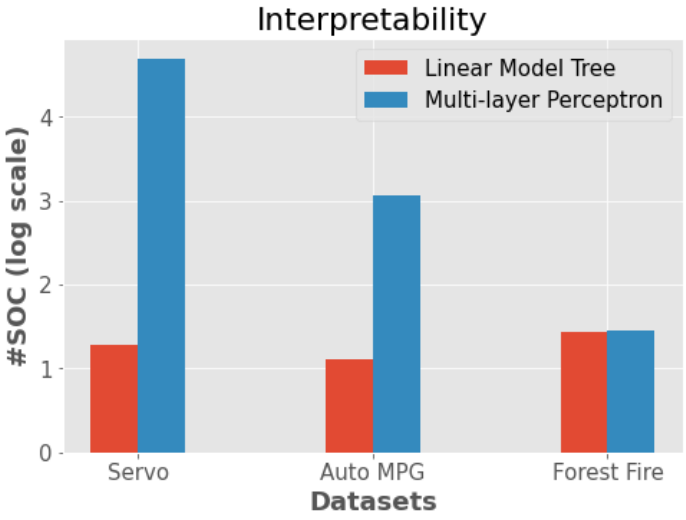
For comparatively simple datasets (AutoMPG and Servo), LMT model can be used to increase interpretability with a small accuracy degradation. On the other hand, when the degradation of

*Corresponding Author: Zhumakhan Nazir, zhumakhan.nazir@nu.edu.kz

accuracy cannot be neglected for complex datasets (i.e. Forest Fire), we can prune a complex model by hyper-parameter tuning in order to raise interpretability level (e.g. by decreasing the number of neurons and/or hidden layers in MLP) within practical accuracy range. As a by-product, size of the model may be reduced, training and inference speed may be increased.



(a) Accuracy (MAE)



(b) Interpretability (#SOC)

Figure 1: Motivating Example: Accuracy and interpretability of LMT and MLP Algorithms

2.2. Related Work

Interpretability/explainability of ML models and explainable AI are now emerging research areas due to the wide usage of AI technologies [8]. According to [2], it is favored using simple and interpretable models as they are capable of replacing sophisticated 'black box' models. In [9], if-then-based rules are extracted from SVM using a two-step method: first run SVM on data and obtain the set of support vectors, then another interpretable model is trained. In [10], the author proposed a human-based proxy metric that is derived from evaluation of model interpretability by humans or a black-box model's post-hoc interpretation. Authors of [11] studied a simulatability and a 'what if' local explainability of logistic regression, neural network and decision tree. They proposed the metric of interpretability as the run time Operation Count (OC). According to [5] the Simulatability Operation Count (SOC) evaluates interpretability for several regression models through the proposed formula. The experiments in this paper use SOC formulas for comparing interpretability of selected models in our experiments.

3. Methodology

3.1. SOC metric

Interpretability of algorithms can be evaluated in terms of simulatabilty. Simulatability Operation Count (SOC) - the number of arithmetic operations needed to execute an algorithm. According to [8], SOC can be a proxy metric for simulatability. For instance, a linear regression model with 10 variables does 10 multiplications and 9 additions, thus its SOC is 19. More detailed derivation of SOC of estimators can be found in [5].

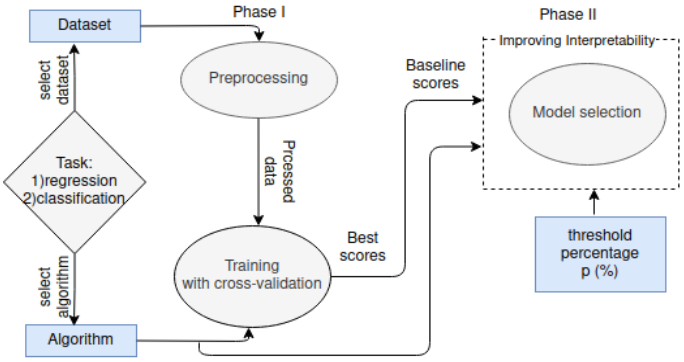


Figure 2: An Overview of the Trade-offs Methodology

3.2. Workflow of the experiment

The workflow of the experiment is shown in Figure 2. Phase I is divided into Data Preprocessing, followed by Model Training. In the first stage 1) categorical entries of datasets are converted to numerical with OrdinalEncoder [12]; 2) StandardScaler [12] is applied to reduce effects of entries on regression coefficients; 3) outliers which has z-score bigger than 3 are removed from the dataset [13]; 4) correlated entries are dropped to prevent multicollinearity (i.e. Variance Inflation Factor (VIF) is larger than 10) [14]. In the next stage (model training), hyper-parameters are selected by applying GridSearchCV implementation of sklearn [12].

Phase II consists of a model selection method that we are proposing. In the first step, the SOC scores of the chosen estimators at the previous training phase will be evaluated. Next, we repeatedly run a model selection process to decrease SOC scores by letting the accuracy to drop by up to a limit set by threshold percentage ($p\%$) from the highest accuracy values achieved in the training stage (trade-offs between accuracy and interpretability). The threshold percentage is chosen arbitrarily between 0 and 15%, but in reality its optimality depends on the specifics of the task (i.e. error tolerance and requirements like transparency and accountability). For example, on Figure 1a LMT and MLP have almost similar accuracy, on Figure 1b LMT has much lower SOC. In such cases, LMT is a suitable candidate for the tasks that require interpretability of algorithms.

Such trade-off can be done by tuning parameters of models influencing the SOC according to Table 1.

Table 1: SOC formula of algorithms [7]

Estimator	K_t or A_t	SOC formula
LMT	N/A	2D + 2P + 1
DT	N/A	2D + 1
MLP	A_t Relu Sigmoid Tanh	$2 \times N_{H+1} + \sum_{h=1}^{H}(2 \times N_h + A_t) \times N_{h+1}$ $A_t = 1$ $A_t = 4$ $A_t = 9$
SVM	K_t Linear Polynomial Sigmoid RBF	SV × (K_t+ 2) $K_t = (2P - 1)$ $K_t = (2P + 1 + d)$ $K_t = (2P + 10)$ $K_t = (3P + 1)$

Following the feature selection stage, the number of variables in the dataset (P) is fixed, the depth (D) could be decreased to reduce SOC in Decision Tree (DT) and in LMT.

The type of activation functions (A_t), the number of hidden layers (H) and neurons (N) can be tuned in MLP to obtain lower SOC values. Lastly, selecting a simpler kernel function (K_t , e.g. Linear) and decreasing the number of support vectors (SV) (using NuSVR and NuSVC [12]) and by tuning hyperparameters reduces SOC in SVM.

Table 2: Servo Features

Feature	Description
motor screw pgain vgain class	A,B,C,D,E A,B,C,D,E 3,4,5,6 1,2,3,4,5 0.13 to 7.10

4. Experimental Results

4.1. Experimental Setup

Datasets for Regression: For our experiments three test datasets (from complex to simple) are used: Forest Fire (complex) [15], Auto MPG (medium) [16] and Servo (simple) [17].

1) *Servo*: There are 5 variables with target class, and 167 data instances. Value of each variable is discrete, except for the target, it is continuous within the range [0.13, 7.1] and is a servo-mechanism's raise time. Detailed descriptions are explained in the TABLE 2.

2) *Auto MPG*: There are 9 variables and 398 data instances, the target variable is 'mpg' (miles per gallon). Detailed descriptions are in the TABLE 3.

Table 3: Auto MPG features

Feature	Description
mpg	**miles per gallon, continuous output variable**
model year	version of a car
cylinders	power unit of engine
displacement	measure of the cylinder volume
horsepower	power of engine produces
weight	weight of car
acceleration	amount of time taken for car to reach a velocity of 60 miles per hour
origin	multi-valued discrete
name	name of the car

3) *Forest Fire*: There are 517 data instances and 13 vari- ables including a target class 'area'. Full descriptions are in the TABLE 4.

Table 4: Forest Fire features

Feature	Description
area	**in ha, 0 means less than** 1ha/100 (=100m2)
X,Y	coordinates of place of fire
month, date	categorical value from jan. to dec. and mon. to sun. correspondingly
temp, wind, rain	meteorological data
RH	relative humidity
FFMC,DMC,DC,ISI	components of Fire Weather Index (FWI) of the Canadian system

Datasets for Classification: The classification datasets include Iris (simple) [18], MNIST (medium) [19] and Pima Indian Diabetes (complex) [20].

4) Iris: The dataset contains 5 features with 1 target class and 150 instances. Features of Iris dataset are real values and described in TABLE 5.

Table 5: Iris Features

Feature	Description
sepal length	1.0 - 6.9 cm
sepal width	0.1 - 2.5 cm
petal length	4.3 - 7.9 cm
petal width	2.0 - 4.4 cm
class	**Setosa, Versicolour, Virginica**

5) MNIST: The dataset has 784 features and 70000 (60k training and 10k test images) instances, predicting one of 10 digits. Features of MNIST consists of a 28x28 array of real values of all pixels in the picture.

6) Diabetes: There are 768 instances of 9 features of real values, which are described in TABLE 6. The outcome is positive or negative for the diabetes test.

Table 6: Diabetes Features

Feature	Description
Pregnancies	0 - 17
Glucose	0 - 199
Blood Pressure	0 - 112
Skin Thickness	0 - 99
Insulin	0 -846
BMI	0.0 - 67.1
Diabetes Pedigree Function	0.078 - 2.42
Age	21 - 81
class	**0, 1**

Algorithms for Regression: LMT is a model [21] and its implementation was according to M5 design [17]; Scikit- learn library's [12] MLP Regressor and SVR were used in our experiment. Accuracy metrics is a Mean Absolute Error (MAE)

Algorithms for Classification: DT, MLP Classifier and SVM implementations of scikit-learn library [12] were used to deal with classification task. The percentage of correct predictions (Accuracy) is used as an accuracy metric.

4.2. Results and Analysis

Preprocessing for Regression: Preprocessing stage allowed to reduce Servo dataset to 152 data instances, Auto MPG to 367 ('horsepower' and 'displacement' are dropped because of collinearity issue, 'name' variables is not used), and Forest Fire to 468.

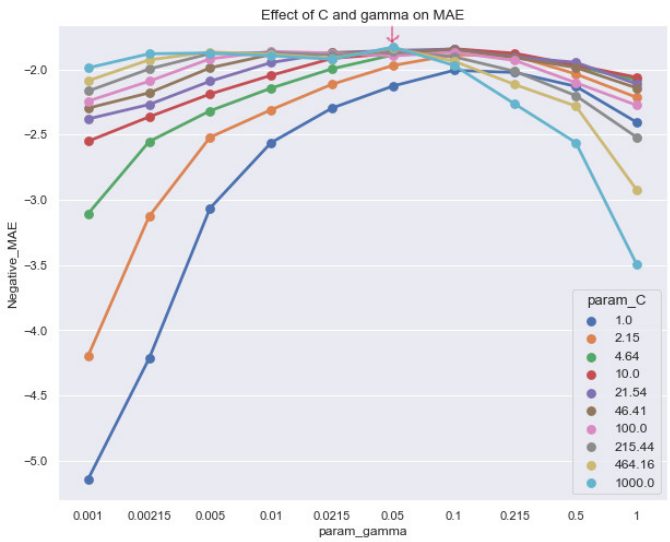


Figure 3: SVR training on Auto MPG dataset.

Model Training for Regression: As mentioned earlier, algorithms are trained with GridSearchCV allowing us to test a broad range of hyper-parameters. Figure 3 shows the process of training SVM on Auto MPG. The lowest error is at C = 1000 and gamma = 0.05 and sub-optimal configuration is obtained by concave down graph.

Table 7: Accuracy Performance (MAE) of Trained Models with references. (Lowest error values in bold).

	Servo	Auto MPG	Forest Fire
LMT	0.133	1.889	6.847
MLP	**0.096**	1.890	5.376
SVR	0.183	**1.830**	**5.212**
Lin. Reg.	0.863	2.304	6.723
Other Ref.	0.220 [22]	2.020 [23]	6.334 [24]

Overall outcomes of the model training phase are in Table 7. SVR performs better than other models on Auto MPG and Forest Fires datasets and MLP on Servo dataset. Performances of the Scikit Learn's Linear Regression and other references are provided for comparison.

Model Training for Classification: Like the regression training phase, GridSearchCV is used to find best combinations of hyperparameters for the models. One of the examples of parameter-tuning is shown in Figure 4, where optimal values are gamma = 0.00003 and C = 4.64.

Table 8: Accuracy Performance of Trained Models with references. (Highest accuracy values in bold).

	Iris	MNIST	Diabetes
DT	96.7%	79.0%	75.4%
Random Forest	97.7%	**97.2%**	76.5%
MLP	83.3%	94.9%	75.7%
SVM	**98.7%**	96.0%	
Other Ref.	**77.1%**		
	98.7% [25]	99.7% [26]	76.0% [27]

Training stage's results are provided in Table 8. On MNIST dataset Random Forest has the best results, while SVM outperforms other models on IRIS and Diabetes dataset. As a comparison, the results of the Random Forest model from the Scikit Learn library were provided.

Model Selection for Regression: Following two approaches to improve interpretability are discussed in this paper: 1) model can be substituted by simpler model and 2) the same model is simplified by tuning its hyperparameters (e.g. reducing the number of neurons or layers in MLP).

Figure 5 shows the results of the first approach and the idea behind it is to demonstrate the behavior of models optimized for interpretability applying the trade-offs method. The point (0, 0) corresponds to the baseline accuracy and SOC score of the estimator on the given dataset. These are the first (top most)

entries of each algorithm and dataset pair on Table 9. For example, for LMT and Servo, baseline score is (0.133, 19). The percentage of the increase or decrease is calculated with respect to the baseline score. Next entry in LMT and Servo on Table 9 is (0.134, 17), which corresponds to (0.75 = (0.134 − 0.133)/0.133 × 100, 10.52 = (19 − 17)/19 × 100) point on the blue graph (encircled with red) on Figure 5.

All estimators (MLP, SVR and LMT) behave similarly on regression task. For example, on the Servo dataset SOC is reduced significantly (approximately by 85%, 17% and 11%) for small raise in error (2.1%, 1.6% and 0.75% respectively). MLP is the most accurate estimator for Servo (with 0.096 MAE value). If 2% reduction in accuracy is feasible for Servo dataset, MLP's interpretability could be increased by 85%. Alternatively, if MAE value of 0.133 is acceptable for Servo dataset, LMT algorithm with SOC value of only 19 could be used instead of MLP.

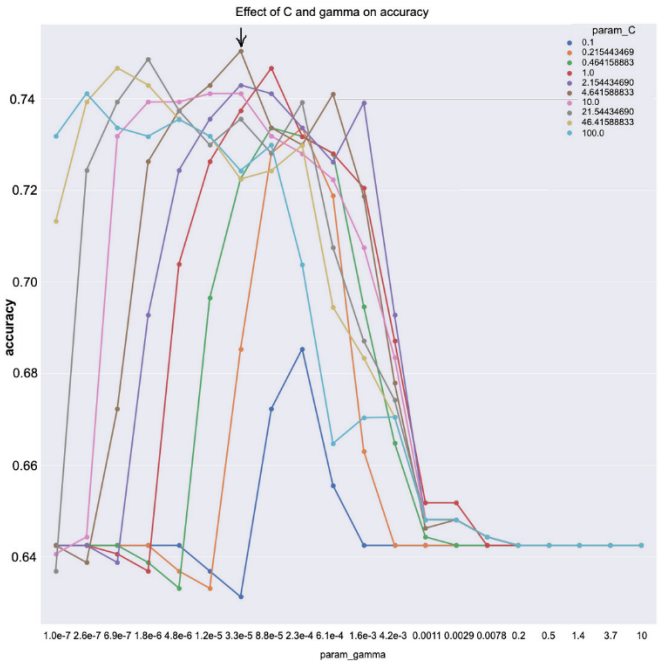


Figure 4: Training SVM on Diabetes dataset.

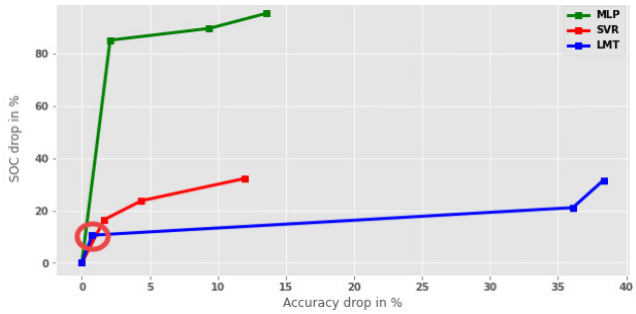


Figure 5: Comparison of Models in Accuracy and Interpretability on the Servo dataset.

Table 9 is a supplement of Figure 5 with additional results of Forest Fire and Auto MPG datasets. Same as in Figure 5, notable advancement is achieved in interpretability with a small degradation in accuracy. For the Auto MPG and Forest Fire datasets using LMT, the most accurate results are obtained with tree depth of 1 (see Figure 7), hence the model cannot be simplified further.

Figure 6 shows the results of the second approach with three datasets using MLP. The similar behavior is observed on all datasets - SOC of the MLP model is decreased notably with small degradation in accuracy. The elbow (turning) points are feasible candidate points for effective trade-off between interpretability and accuracy, since after these points (from left to right) the slopes of the graphs sharply drop. For instance, in Auto MPG dataset (line in red) interpretability is improved (reduction of SOC) by 97% with 2.5% reduction (raise in MAE) in accuracy. The similar pattern is observed in the rest of the datasets.

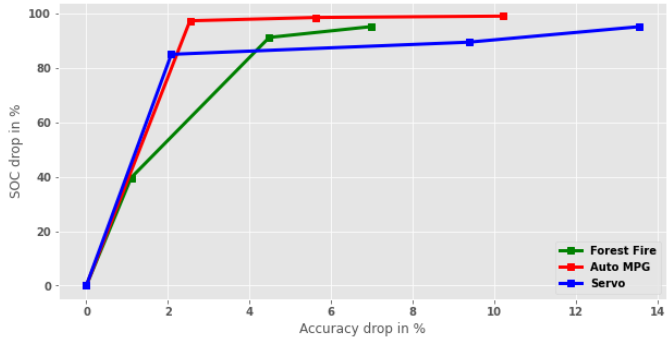


Figure 6: Trade-off between Accuracy and Interpretability in MLP estimator.

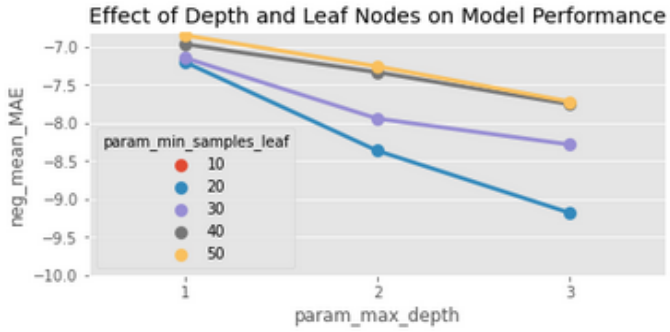


Figure 7: Performance of LMT on Forest Fire dataset.

Model Selection for Classification: For the classification task, similar approaches as for regression were applied; and Figure 8 summarizes the results of the first approach where trade-offs between accuracy and interpretability for all the models (DT, MLP and SVM) on MNIST dataset are depicted. It could be seen from Figure 8 that all graphs start at point (0, 0), which are the baseline scores for accuracy and inter-pretability. These baseline score correspond to the last entries (with highest accuracy and SOC values) of each algorithm and dataset pair on Table 10. For instance, for MLP and MNIST baseline score is (94.9%, 10719). Percentage change in SOC or accuracy are calculated with respect to the baseline score, for example, next entry in MLP and MNIST on Table X is (93.4%, 3879), and it gives a red point (1.5% = (94.9% - 93.4%), 64% = (1-(3879/10719))*100) on Figure 8. Overall, models perform in the same way on the classification task. For instance, interpretability could be increased dramatically (by 7.3%, 64%, and 40%) in exchange for a small decrease in accuracy (6.62%, 1.5%, and 3.2% correspondingly). SVM is the most accurate model (accuracy 96%) for MNIST dataset,

since DT is interpretable intrinsically, its interpretability could not be improved effectively with trade-offs method.

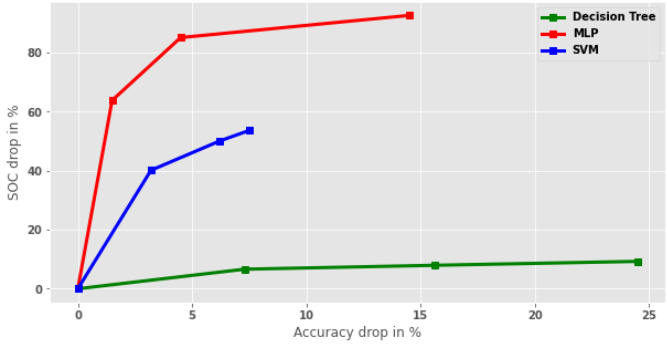


Figure 8: Comparison of Models in Accuracy and Interpretability on the MNIST dataset.

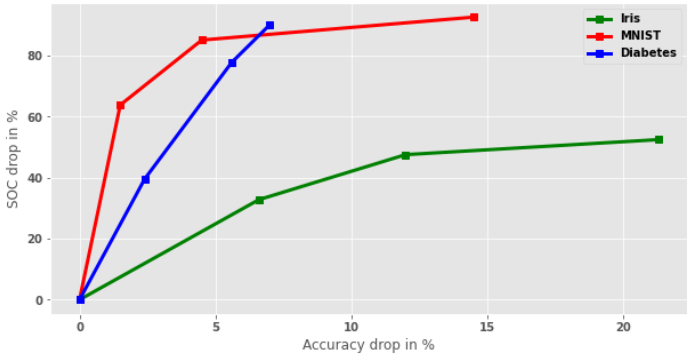


Figure 9: Trade-off between Accuracy and Interpretability in MLP estimator for classification.

Table 10: Comparison of models in terms of accuracy and interpretability (acc. is short for accuracy) for classification.

	DT acc., SOC	MLP acc., SOC	SVM acc., SOC
Iris	66.7%, 11 93.3%, 13 96.0%, 15 96.7%, 17	62.0%, 29 71.3%, 32 76.7%, 41 83.3%, 61	96.7%, 117 97.3%, 162 98.0%, 198 98.7%, 252
Mnist	54.5%, 137 63.4%, 139 71.7%, 141 79.0%, 151	80.4%, 794 90.4%, 1599 93.4%, 3879 94.9%, 10719	88.5%, 45220 89.8%, 48720 92.8%, 58380 96.0%, 97500
Diabetes	72.3%, 19 73.7%, 21 73.8%, 23 75.4%, 27	68.7%, 103 70.1%, 231 73.3%, 627 75.7%, 1038	75.7%, 5185 76.6%, 5338 77.1%,5712 - -

Table 10 is a more detailed version of Figure 8, and it shows significant increases of interpretability by allowing some drops in accuracy.

The second approach was to test one of the models on three datasets, for example MLP (Figure 9). The model performs similarly on all datasets and SOC could be improved at cost of lowering accuracy. For example, 64% increase of interpretability would require 1.5% reduction in accuracy on MNIST.

5. Conclusion

We introduced a methodology for trade-offs between interpretability and accuracy by inheriting the quantitative and model-agnostic metric - SOC. The LMT model, through its powerful but simple architecture (combination of linear regression and decision tree models), is the most interpretable estimator amongst the considered regression estimators; it has comparable accuracy to MLP in simple-medium datasets like Auto MPG and Servo.

The Decision Tree algorithm has the highest interpretability compared to other evaluated models due to its simplicity on classification task. It outperforms MLP on Iris and shows competitive results on Diabetes dataset. However, it has the lowest accuracy on MNIST with a large gap from other algorithms.

This paper demonstrates the tradeoff method between accuracy and interpretability using SOC metric. SOC is a model agnostic quantitative metric, hence it allows fair comparison between different types of estimators. In our experiments decreasing SOC leads to a simpler model with less memory requirement and faster inference speed. However, in general lower SOC may not always result in models with small memory requirement (i.e. replacing complex operation with simpler one) and faster inference speed (parallelizable model with high SOC can be faster than purely sequential model with low SOC on parallel hardwares).

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was partly supported by the Nazarbayev Univer- sity (NU), Kazakhstan, under FDCRGP grant 021220FD0851.

References

[1] Z. Nazir, D. Kaldykhanov, K.K. Tolep, J.G. Park, “A Machine Learning Model Selection considering Tradeoffs between Accuracy and Interpretability,” 2021 13th International Conference on Information Technology and Electrical Engineering, ICITEE 2021, 63–68, 2021, doi:10.1109/ICITEE53064.2021.9611872.

[2] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” Nature Machine Intelligence, **1**(5), 206–215, 2019, doi:10.1038/s42256-019-0048-x.

[3] J.R. Zech, M.A. Badgeley, M. Liu, A.B. Costa, J.J. Titano, E.K. Oermann, “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study,” PLoS Medicine, **15**(11), 1–17, 2018, doi:10.1371/journal.pmed.1002683.

[4] C. Molnar, G. Casalicchio, B. Bischl, “Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges,” Communications in Computer and Information Science, **1323**(01), 417–431, 2020, doi:10.1007/978-3-030-65965-3_28.

[5] U. Johansson, C. Sönströd, U. Norinder, H. Boström, “Trade-off between accuracy and interpretability for predictive in silico modeling,” Future Medicinal Chemistry, **3**(6), 647–663, 2011, doi:10.4155/fmc.11.23.

[6] T. Mori, N. Uchihira, “Balancing the trade-off between accuracy and interpretability in software defect prediction,” Empirical Software Engineering, **24**, 779–825, 2019, doi:10.1007/s10664-018-9638-1.

[7] J.-G. Park, N. Dutt, S.-S. Lim, “An Interpretable Machine Learning Model Enhanced Integrated CPU-GPU DVFS Governor,” ACM Trans. Embed. Comput. Syst., **20**(6), 2021, doi:10.1145/3470974.

[8] A. Adadi, M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” IEEE Access, **6**, 52138–52160,

2018, doi:10.1109/ACCESS.2018.2870052.
[9] M.A.H. Farquad, V. Ravi, S.B. Raju, "Support vector regression based hybrid rule extraction methods for forecasting," Expert Systems with Applications, **37**(8), 5577–5589, 2010, doi:https://doi.org/10.1016/j.eswa.2010.02.055.
[10] F. Doshi-Velez, B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," ArXiv E-Prints, arXiv:1702.08608, 2017.
[11] D. Slack, S.A. Friedler, C. Scheidegger, C.D. Roy, "Assessing the Local Interpretability of Machine Learning Models", NeurIPS Workshop on Human-Centric Machine Learning, 2019, doi:10.48550/ARXIV.1902.03501.
[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, **12**, 2012.
[13] E-handbook of statistical methods, NIST/SEMATECH, 2012, doi: https://doi.org/10.18434/M32189.
[14] R. O'Brien, "A Caution Regarding Rules of Thumb for Variance Inflation Factors," Quality & Quantity, **41**, 673–690, 2007, doi:10.1007/s11135-006-9018-6.
[15] P. Cortez, A. Morais, "A Data Mining Approach to Predict Forest Fires using Meteorological Data," 2007.
[16] J.R. Quinlan, "Combining Instance-Based and Model-Based Learning," in Proceedings of the Tenth International Conference on International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA: 236–243, 1993.
[17] J.R. Quinlan, "Learning With Continuous Classes," World Scientific: 343–348, 1992.
[18] R.A. FISHER, "THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS," Annals of Eugenics, **7**(2), 179–188, 1936, doi:https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.
[19] L. Deng, "The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]," Signal Processing Magazine, IEEE, **29**, 141–142, 2012, doi:10.1109/MSP.2012.2211477.
[20] J. Smith, J. Everhart, W. Dickson, W. Knowler, R. Johannes, "Using the ADAP Learning Algorithm to Forcast the Onset of Diabetes Mellitus," Proceedings - Annual Symposium on Computer Applications in Medical Care, **10**, 1988.
[21] L. Dillard, lmt.py, 2017, Link: https://gist.github.com/logandillard/lmt.py.
[22] P. Cortez, M.J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," Information Sciences, **225**, 1–17, 2013, doi:https://doi.org/10.1016/j.ins.2012.10.039.
[24] P.E. Utgoff, ed., "Machine Learning, Proceedings of the Tenth International Conference, University of Massachusetts, Amherst, MA, USA, June 27-29, 1993," Morgan Kaufmann, 1993, doi:10.1016/c2009-0-27798-1.
[24] A. Stanford-Moore, "Wildfire Burn Area Prediction," 2019.
[25] Z. Hussain, H. Ibraheem, M. Aljanabi, A. Ali, M.A. Ismail, S. Kasim, T. Sutikno, "A new model for iris data set classification based on linear support vector machine parameter's optimization," International Journal of Electrical and Computer Engineering (IJECE), **10**, 1079, 2020, doi:10.11591/ijece.v10i1.pp1079-1084.
[26] D.C. Cireşan, U. Meier, L.M. Gambardella, J. Schmidhuber, "Deep, Big, Simple Neural Nets for Handwritten Digit Recognition," Neural Computation, **22**(12), 3207–3220, 2010, doi:10.1162/NECO_a_00052.
[27] B. Chandra, V.P. Paul, "A Robust Algorithm for Classification Using Decision Trees," in 2006 IEEE Conference on Cybernetics and Intelligent Systems, 1–5, 2006, doi:10.1109/ICCIS.2006.252336.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 5, 73-78 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Frequency Oscillation Suppression of Interlinked Solar PV-Thermal Power System using HVDC Link

Gulshan Sharma*

Department of Electrical Engineering Technology, University of Johannesburg, Johannesburg, 2006, South Africa

ARTICLE INFO

Article history:
Received: 26 August, 2022
Accepted: 31 August, 2022
Online: 10 October, 2022

Keywords:
AC Tie-line
Dynamic Responses
HVDC link
PV-Thermal Interlinked LFC
PID Design Tuning Methods

ABSTRACT

This paper looks for the merging of solar PV power plant with thermal power generation and interlinked through tie-lines approaching into an interlinked PV-Thermal model for the design and analysis of load frequency controllers. Moreover, the different models of controllers are also proposed and tested to regulate the frequency of PV-Thermal interlinked system for a certain disturbance in one of the regions of an interlinked system. The choice of auxiliary controllers is assessed utilizing assorted tuning strategies and the comparative examination of all models are shown via graphical response and via calculating numerical values to bring back the frequency and tie-line power deviations back to the standard level and to analyze the viable control activity. In addition, an HVDC link is also used between linked PV-Thermal areas to stabilize the frequency and tie-power quickly as well as to stabilize the system operation for alteration in changing loading conditions of the interlinked system.

1. Introduction

Load frequency control (LFC) is a critical process in the power system, and it determines the quality energy supply to the cutting-edge society. We fulfil this requirement by matching the power system generation with the power demand and hence frequency manages to its original value i.e., 50 Hz. However, the power demand continuously changes over the complete day and hence it shifts the frequency value from 50 Hz to either up or low from the original value and therefore the power customers are affected. Subsequently, load frequency controllers are required on the generators in position to keep the power system generation up or moo as per the power demand. This gives the electrical energy to the clients with least change from the standard and acceptable level (i.e., 50Hz).

The frequency stability reflects approximately the beginning time and moment of control activity. As there's an unbalance between the power generation and power demand, the first control action (Primary control) is initiated and tries to increase or decrease the power generation as per the requirement. The primary control loop is with the synchronous generators and known as first level of control. In any case, the action of primary control loop is limited, slow and hence unable to take the frequency of the system back to its original value and consequently the auxiliary control activities (i.e., secondary control) plays a vital part in achieving quick reclamation to power system frequency to perceived level.

Apart from this, the regions are interconnected with the means of tie-lines which generally interchange power between the regions, and these are AC tie-lines. But they suffer with large number of technical problems and one of the major issues are transferring of disturbances between the regions and that is why these drawbacks can overcome via moving the HVDC tie-line with AC tie-lines between the regions. Advance, it is continuously fitting that each region ought to meet its claim power request whereas overseeing the power give-and-take over the tie-lines to the plan value [1-2].

In any case due to continuous change in power demand as per the client's requirement ultimately results in frequency as well as tie-line power alteration from the beginning level and hence a secondary control plan is persistently required in the system. The deviation in frequency with power deviations over tie-lines combined in a linear form is recognize as area control error (ACE). South Africa is moving ahead in orchestrate to meet its current and future power demand through cheap as well as clean sources such as sun fuelled control well known as PV based control. In any case, it is enormously troublesome to ACE deviations in the power system as the power generated through PV is exceedingly convulsive in nature.

Still, the PV based power generation is conclusion of the country considering its boundless openness, brought compelling

*Corresponding Author: Gulshan Sharma, gulshanmail2005@gmail.com

and defilement free keeping the environment of the country clean and sickness free as well as in arrange to cater the current as well as rising power demands of South Africa. An introductory exertion in frequency management reflects about is to investigate the controller plan based on classical approach [3-4].

In any case, this approach appears to be totally unsuccessful in view of set working conditions of the power system. The controller plan approach for frequency regulation have progressed over the past few years. Assorted calculations such as optimal control by seeing each state [5]; output vector feedback design seeing crucial states [6], changing structure [7] and robust controller designs [8-9] are well attempted over the last number of years to reach the LFC benchmarks. Be that as it may, most of the frequency steadiness remedies are given for well establish plants i.e. thermal power or for combination of hydro-thermal and exceptionally scarce approaches [10-13] were available for solar PV generation and integration to the existing systems. The solar PV can generate power in one of the regions and can be interlinked to thermal power via tie-lines in order to have stable operation. Further an HVDC link may prove to be more efficient in meeting the standards of solar PV-Thermal as well as to deliver quality energy to people of South Africa.

It is well-known that the managers and labour working in assorted control businesses are very comfortable with the different structures of auxiliary controller which are "PI (proportional-integral), PD (proportional-derivative), proportional-integral-derivative (PID)" to oversee industry operations in day-to-day life in different spaces as the operation of these controllers and setting the pickup values are much easier for them and these can be achieved by considering the graphical system response [14] and hardly required any numerical calculations. Although, the pickup values and the controller design play a challenging role in achieving the required control target and sometimes it is also advisable to design and check the diverse controller designs for a certain application [15]. Within the scope, this paper is set to;

- Present mathematical model of PV and thermal power interlinked through tie-line i.e., PV-Thermal. To discuss the transfer function model of HVDC link.
- Create the conventional validation of solar PV in region one with thermal power in region two and these two power generated regions are connected via help of HVDC link parallel to AC tie-lines.
- Further, to propose and to calculate the pick-up value of various auxiliary controllers for LFC using various tuning methods and to see the merits of one over other design for step disturbance.
- The outcomes are shown via graphs for various states to find the best design for the proposed LFC.
- At last, the performance of auxiliary controllers are also matched by calculating the error (integral time multiplied by absolute error) for all LFCs with and without HVDC link to show the positive benefits of using HVDC running parallel to tie-lines.

2. Details of PV-Thermal LFC Model

The PV cell is made up of current source and it is directly related to solar radiation with one diode parallel to it. It also has a small resistance in series and its schematic representation is shown in Figure 1 [10-13]. Further the output of this cell depends on solar radiation and surface temperature. Generally, the MPPT technique is used to achieve the maximum output from PV and hence it enhances the output as well as efficiency of PV.

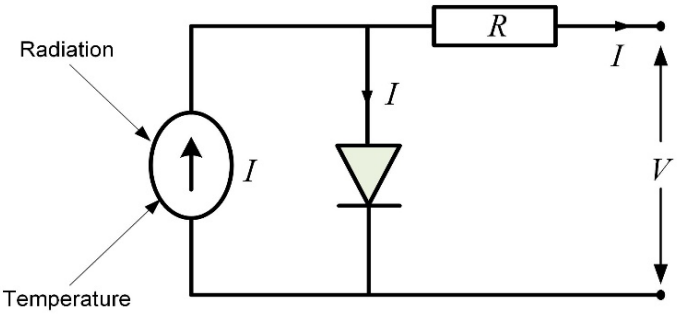


Figure 1: Representation of PV cell

The model of PV is a combination of PV panel, an inverter to convert DC power to AC power, MPPT system and filter. The complete representation in form of transfer function (TF) is as follows [10-13]:

$$G_{PV} = \frac{-18s+900}{s^2+100s+50} \quad (1)$$

The present work is the interconnection of PV with thermal power finally coming into an PV-Thermal model for frequency regulation of the power system. The TF of PV is given in equation (1) and can be used for load frequency control studies. The thermal power system includes governor to regulate the flow of steam and its output is connected to turbine with finally the shaft of turbine coupled to the power system generator. The PV-Thermal areas are interlinked via AC tie-line. The governor for thermal area in TF is as follows [13]:

$$\frac{K_g}{T_g s + 1}$$

With re-heater in TF as.

$$\frac{K_r T_r s + 1}{T_r s + 1}$$

The steam turbine TF representation is as;

$$\frac{K_t}{T_t s + 1}$$

Finally, the TF of power system generator is depicted as.

$$\frac{K_P}{T_P s + 1}$$

The interlinked PV-Thermal model for LFC is shown in Figure 2 and these areas are interlinked via AC tie-line.

However, these tie-line struggle to damp out the system oscillations quickly as well as disturbances are transferred from one area to another and hence affect the customers. These problems can be solved by running HVDC link parallel to AC.

The HVDC interface comprises of a rectifier and an inverter. The terminating points are balanced to control the rectifier and the

inverter. The first-order exchange work of dc-link is communicated as:

$$K_{DC}/(1+sT_{DC}) \quad (2)$$

Where, K_{DC} is the gain of the link as well as T_{DC} represent the responding time of HVDC link.

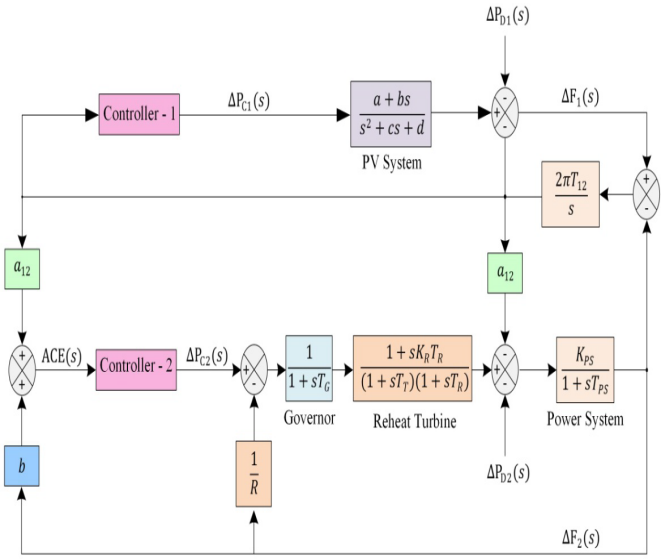


Figure 2: PV-Thermal LFC model

3. Design of Auxiliary Controllers for PV-Thermal System

One of the most common auxiliary controllers used in different types of power as well as in control industries is PID as it has three diverse control abilities which are "proportional, integral & derivative" and these abilities help the system response to come back to initial value quickly after certain disturbance. The PID work on the concept of feedback system; as PID is connected in feedback and hence it calculates the difference between actual and desired value i.e., error and tries to minimize the error by regulating the input of the system. In present research work, the aim of connecting the PID is to increase or decrease the electrical power in solar PV as well as in thermal power plants as per the requirement of electrical power and hence to attain the balance between the two so that frequency of each area and connecting tie-lines deviations should be minimized. As PID has three diverse control abilities and that is why pickup values of each control need to be calculated correctly. At first, the disturbance of value 0.01 per unit (p.u.). is imposed in area one of an interlinked PV-Thermal and at the same time, the pickup value of proportional control needs to be increased in small steps until the system start showing complete oscillations as per the Ziegler–Nichols (ZN) practice [15].

The pickup after which the system loses its stability is known as extreme pick up (Ku) for thermal region and the time between crests (Tu) is well-known in order to calculate the picks up of PID for different modes.

The same formulation is additionally taken after for PV system also. At last, with the assistance of ZN practice the pick-ups of different modes of PID are calculated and given in Table 1 for solar PV region and Table 2 for thermal system and calculated value of error i.e., ITAE is shown in Table 3 with and without inclusion of HVDC link.

Table 1: PID Gains for solar PV system

Auxiliary Controller	Kp (Proportional Gain)	Ki (Integral Gain)	Kd (Derivative Gain)
Proportional Integral (PI)	1.79	1.542	
Proportional Derivative (PD)	3.19		0.559
Classic PID	2.39	3.428	0.419
Pessen Integral Rule	2.80	5.000	0.588
Some Overshoot	1.30	1.902	0.622
No Overshoot	0.80	1.15	0.373

Table 2: PID Gains for thermal power

Auxiliary Controller	Kp (Proportional Gain)	Ki (Integral Gain)	Kd (Derivative Gain)
PI	1.79	2.4	
PD	3.19		0.36
Classic PID	2.39	5.33	0.27
Pessen Integral Rule	2.80	7.77	0.37
Some Overshoot	1.33	2.96	0.40
No Overshoot	0.80	1.77	0.24

Table 3: ITAE values with and without HVDC Link

Auxiliary Controller	ITAE	ITAE with HVDC
PI	4.224	0.05129
PD	0.116	0.116
Classic PID	0.0065	0.00104
Pessen Integral Rule	0.0058	0.00099
Some Overshoot	0.0202	0.01547
No Overshoot	0.0565	0.02095

4. Simulation Details and Analysis for LFC

The research work is aimed to check the feasibility of generating electrical energy via means of renewable solar PV in one area and interlinking to power generation via means of thermal power in another area and connected through tie-lines resulting into an interlinked PV-Thermal system used for designing and

analysis of load frequency controllers to regulated frequency and interlinked tie-lines deviations for a certain disturbance.

In addition, the PV-thermal areas are connected via additional HVDC tie-line running parallel to AC tie-line between two interlinked areas. The HVDC tie-lines parallel to AC have the capability to improve the load frequency control responses due to lesser time constant and will show reduce oscillations, better settling time and steady state error to zero value.

The various sub-designs of PID such as "PI, PD & PID" are design and tested for interlinked model to check the superiority of one design over other design. The auxiliary controllers are design to minimize frequency and tie-power deviations to minimum or zero value.

The Zeigler-Nichols tuning method is used to move forward the framework reactions. The picks up of different modes of controllers are assessed with the assistance of ZN practice. The obtained pickup values for various PIDs are given in Table 1 for solar PV and for thermal plants in Table 2. The diverse calculated error values for each mode of auxiliary controllers are given in Table 3. The Table 3 also shows the ITAE with and without inclusion of HVDC link. Now the integral time absolute error (ITAE) definition is chosen to obtain the output of various modes of auxiliary controllers and ITAE is built on the basis of ACE.

The step load is applied in thermal power and PV-Thermal responses for various modes are obtained by considering HVDC parallel to AC tie-line between PV-Thermal regions. The execution of interconnected PV-Thermal is assessed on gotten picks up as given in Table 1 & Table 2 for different strategies and the comparative comes about for frequency, AC tie-power alteration & ACE of each area is appeared in Figures 3-7.

A look at these outcomes about clearly appears that PD mode is unable to take the system results back to reference value within 50 seconds also. The calculated ITAE is 0.116 and it is highest in the group of these control actions.

The results secure through PI and No Overshoot have the supported motions for frequency, AC tie-line power and ACEs for area 1 & 2 which is additionally not alluring seeing the required control and ITAE is 4.224 for PI without HVDC link and it reduces to 0.05129 with HVDC.

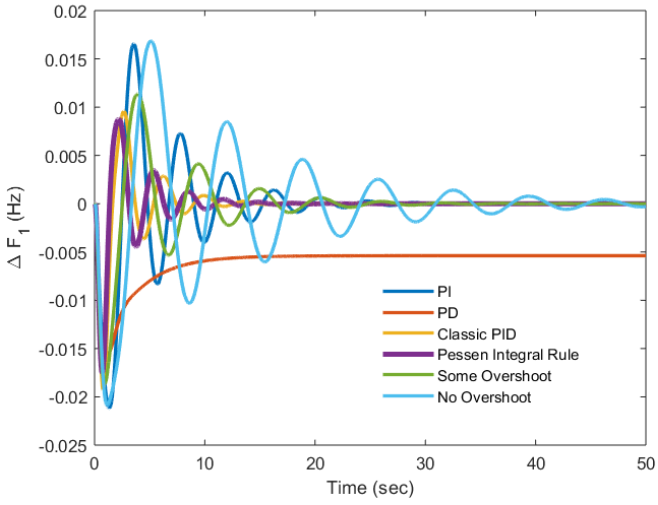


Figure 3: Graphical outcomes of PV-Thermal system

At the same time the ITAE for No Overshoot is 0.0565 without HVDC link & it reduces additionally with HVDC. The results about advertised by Some Overshoot strategy are superior to PI, PD & No Overshoot nevertheless having number of oscillations and settling time is as large as of 20 seconds for all outcomes. The results obtained via classic PID is much better for all graphical results in view of reduced overshoot, lesser oscillations and much lesser time to reach the steady state value when matched with regards to that obtained through other studied auxiliary load frequency controllers.

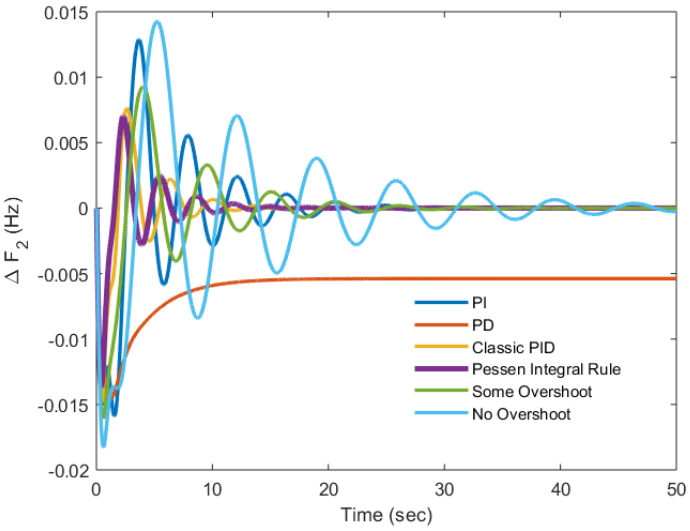


Figure 4: Graphical outcomes of PV-Thermal system

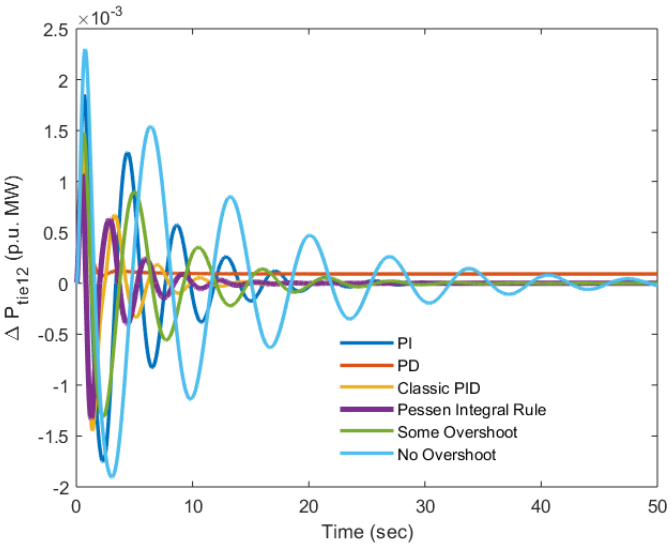


Figure 5: Graphical outcomes of PV-Thermal system

The frequency responses by using the Pessen Integral Rule have (-0.02) Hz of variation from the original value for solar PV area with response settling in 14 seconds. At the same time, the frequency variation is (-0.015) Hz for thermal power plants with settling in 10 seconds & able to go back to the reference or original value much faster than solar PV plants. The same trend is seen for response of tie-line power, ACEs for area 1 and 2 and hence it is concluded that Pessen Integral Rule based auxiliary controller shows much better output in reaching the targets of PV-Thermal system in least conceivable time.

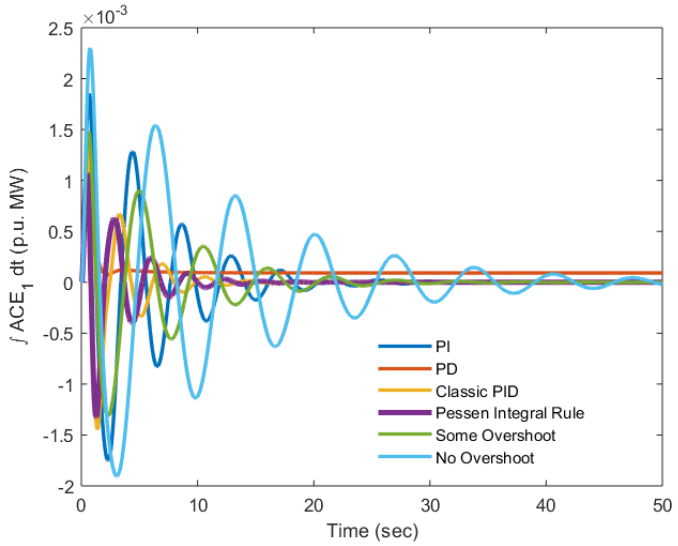


Figure 6: Graphical outcomes of PV-Thermal system

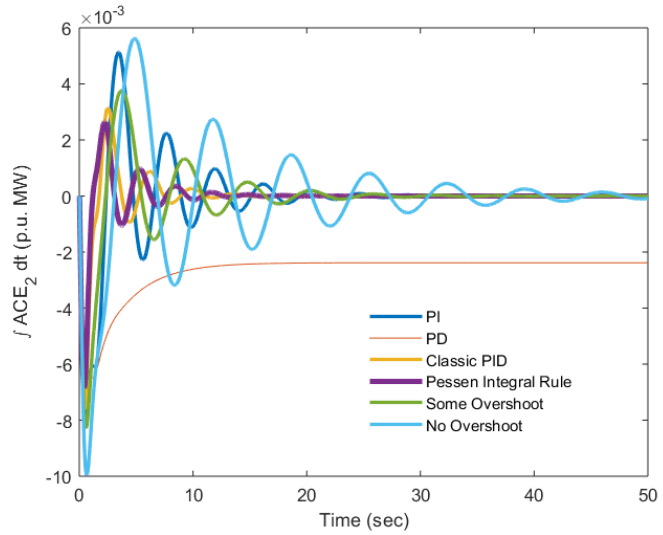


Figure 7: Graphical outcomes of PV-Thermal system

The obtained ITAE is 0.0058 and it comes out to be 0.00099 and this value of ITAE is least ITAE in complete group of investigated auxiliary controllers. It is additionally seen that the settling time and top of overshoot is also less for responses of area-2 when matched with area-1. The reason is that thermal power are well establishing in terms of design and development over the last number of years where as solar PV plants depend highly on solar radiation for generating the output and the solar radiation is highly intermittent in the nature. Further, it is also noted that HVDC link has minimized the ITAE of all auxiliary controllers. When ITAE of various modes are compared and all ITAE values are obtained with respect to HVDC running similar to AC between PV-Thermal regions, it is further verified that ITAE of Pessen Integral Rule outperform all other modes and reach to best possible minimum limit and hence justify the positive impact of this mode for PV-Thermal System.

5. Conclusions

This paper attempts to achieve the of linking of Solar PV power plants with the help of tie-lines to thermal power coming into an interlinked system. In addition, an impact of HVDC link with AC tie-line is also investigated in this work. The distinctive modes of PID are outlined by developing and checking various PID structures. The design and investigations of various load frequency controllers revealed that the design accomplished by Pessen Integral Rule is quick enough to bring back the responses of PV-Thermal system such as (frequency of area-1 and 2, tie-line power and ACEs of area-1 and 2) deviations to original value after the load disturbance. The output of all responses is smooth with very few oscillations and coming back to original value much faster in comparison to other load frequency controllers. It is also perceiving that error value i.e., ITAE obtain via Pessen method is not as much of when compared with ITAE of other PID controllers. The significant reduction in error values is observed for all auxiliary controllers with inclusion of HVDC link. With HVDC link also, ITAE achieved via Pessen Integral Rule is minimum when compared with other auxiliary controllers and hence this method of PID outperforms over other auxiliary controllers for present studies. However, the research work can be extended to calculate the values of various PIDs by using optimization techniques. The effect of non-linearity and time delay with sensitivity analysis also need to be study for PV-Thermal system. The work may be extended to deregulated environment in view of various market transactions and to study the ancillary services such as LFC in deregulated environment.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] E. Çelik, N. Ozturk, and E. H. Houssein, "Influence of energy storage device on load frequency control of an interconnected dual-area thermal and solar photovoltaic power system," Neural Computing and Applications, 1-17, 2022, doi: 10.1007/s00521-022-07558-x.

[2] F. T. Tomy, R. Prakash, "Load frequency control of a two-area hybrid system consisting of a grid connected PV system and thermal generator," Int J Res Eng Technol., **3**(7), 573–580, 2014.

[3] J. Nanda, S. Mishra, and L. C. Saikia, "Maiden application of bacterial foraging based optimization technique in multi-area automatic generation control," IEEE Trans. Power Syst., **24**, 602-609, 2009, doi: 10.1109/TPWRS.2009.2016588.

[4] Ibraheem, P. Kumar, and D. P. Kothari, "Recent philosophies of automatic generation control strategies in power systems," *IEEE Trans. Power Syst*., **20**(1), 346–357, 2005, doi: 10.1109/TPWRS.2004.840438.

[5] Ibraheem, K. R. Niazi, and G. Sharma, "Study on dynamic participation of wind turbines in AGC of power system," Electric Power Comp. and Syst., **43**(1), 44-55, 2014, doi: 10.1080/15325008.2014.963266.

[6] G. Sharma, Ibraheem, and K. R. Niazi, "Optimal AGC of asynchronous power systems using output feedback control strategy with dynamic participation of wind turbines," Electric Power Component and Systems, **43**(4), 384-398, 2015, doi: 10.1080/15325008.2014.949916.

[7] N. N. Bengiamin, and W. C. Chan, "Variable structure control of electric power generation," IEEE Trans Power Apparatus Syst., PAS-101, 376–380, 1982, doi: 10.1109/TPAS.1982.317117.

[8] G. Sharma, I. Nasiruddin, K. R. Niazi, and R. C. Bansal, "Robust automatic generation control regulators for a two-area power system interconnected via AC/DC tie-lines considering new structures of matrix Q", IET-Generation Transmission and Distribution, **10**(14), 3570-3579, 2016, doi: 10.1049/iet-gtd.2016.0321.

[9] Y. Wang, R. Zhou, and C. Wen, "Robust load-frequency controller design for power system", IEE Proc. C, **140**(1), 11-16, 1993, doi: 10.1049/ip-c.1993.0003.

[10] E. S. Ali, "Speed control of DC series motor supplied by photovoltaic system via firefly algorithm," Neural Comput Appl., **26**(6), 1321–1332, 2015, doi: 10.1007/s00521-014-1796-5.

[11] A. S. Oshaba, E. S. Ali, and S. M. Abd-Elazim, "PI controller design for MPPT of photovoltaic system supplied SRM via BAT search algorithm," Neural

Computing Application, **28**(4), 651-667, 2015, doi: 10.1007/s00521-015-2091-9.

[12] A. S. Oshaba, E. S. Ali, and S. M. Abd-Elazim, "Speed control of SRM supplied by photovoltaic system via ant colony optimization algorithm," Neural Computing Application, **28**(2), 365-374, 2015, doi: 10.1007/s00521-015-2068-8.

[13] S. M. Abd-Elazim, and E. S. Ali, "Load frequency controller design of a two-area system composing of PV grid and thermal generator via firefly algorithm," Neural Computing Application, 1-10, 2016, doi: 10.1007/s00521-016-2668-y.

[14] A. Panwar, G. Sharma, I. Nasiruddin, and R. C. Bansal, "Frequency stabilization of hydro–hydropower system using hybrid bacteria foraging PSO with UPFC and HAE," Electric Power Systems Res., **161**, 74-85, 2018, doi: 10.1016/j.epsr.2018.03.027.

[15] J. G. Ziegler, and N. B. Nichols, "Optimum setting for automatic controllers," Transaction ASME, **64**, 759–768, 1942.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 6, 77-86 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

The Security of Information Systems and Image Processing Supported by the Quantum Computer: A review

Tarek Nouioua*,[1,2], Ahmed Hafid Belbachir[2]

[1]*Department of Mathematics and Computer Sciences, Faculty of Exact sciences and sciences of Nature and life, University of Tebessa, Tebessa, 12000 , Algeria*

[2]*Laboratory of Analysis and Application of Radiation (LAAR), Engineering Physics Department, Faculty of Physics, University of Sciences and Technology of Oran Mohamed Boudiaf, Algeria, Oran, 31000, Algeria*

ARTICLE INFO

Article history:
Received: 30 August, 2022
Accepted: 24 October, 2022
Online: 15 November, 2022

Keywords:
Quantum Computer
Qubit
RSA Algorithm
Information Security
Ternary Quantum Systems
Quantum Image Security

ABSTRACT

The knowledge and understanding of the technology of quantum computers and their superiority over classical computers are still insufficient or uncertain for many communities of researchers, manufacturers, investors and the general public. For this reason, we try in this article to present and explain some of the basic concepts of quantum computers. We explain how the quantum phenomena may be employed to conceive a quantum computer by defining the qubit that will represent the data entity corresponding to the bit in the classical computer and how this computer can effectively be powerful. We address the issue of strengthening the information system security through a simulation of a spy hunter and the importance of image processing security using the quantum computer, which will minimize the data processing time regardless of the amount of data to be processed. The security of the images will lead us to introduce the new prospects of using multilevel systems instead of binary systems, which will exponentially increase the gain in the size of the memory used.

1 Introduction

This paper is an extension of a work initially presented at the International Conference on Recent Advances in Mathematics and Informatics (ICRAMI) [1]. The main objective is to contribute to attracting attention to the importance of the quantum computer in the future for those who now use classical computers in all their different fields or who work in the development of technology and the computer industry and anyone related to this advanced technology.

To be convincing, an example of a spy hunter simulation will be presented and discussed. The relevance of using quantum computers in image processing security will also be presented and discussed. The impact of using multilevel or ternary quantum systems compared to binary quantum systems will be presented to argue their advantage in minimizing the memory size in image processing.

The technological evolution of the processors has considerably improved their efficiency, namely the calculation time and the energy consumption. However, the processor's physical core is changing as well as the nature of the data which will be represented by the qubit instead of the bit, by studying quantum physics.

Studying quantum physics and aiming to make it representative of the information will help us conceive the quantum computer and use it efficiently. Following that idea, several laboratories invest budgets in research to have the first quantum computer and benefit from quantum physics.

Quantum physics history goes back to the 1940s and plays a central role in manufacturing computers' electronic components. In 1982 [2], the author was the first to propose the concept of the quantum computer, and in 1986 by another paper [3], he confirmed his thought. However, quantum computing got its big break in the 1990s. The most important being that idea [4, 5], where the author used a quantum algorithm for the factorization of prime numbers of size n in a time $O((n)^3)$ and space $O((\log n))$.

Another idea of a quantum algorithm in 1992 was developed [6], which did not have significant importance. But in 1996, another one [7], developed a quantum algorithm which consists in searching one or more elements out of N elements within a time proportional to $\sqrt{N}$, with a storage space which is proportional to $logN$.

*Corresponding Author: Tarek Nouioua, Department of Mathematics and Computer Sciences, Faculty of Exact sciences and sciences of Nature and life, University of Tebessa, Algeria, tarek.nouioua@univ-tebessa.dz

Security requirements are essential for a secure transaction on a transmission canal using an asymmetric encryption algorithm. Information systems exchange data over the transmission canal using the RSA (Rivest-Shamir-Adleman) algorithm; since the encryption key is different from the decryption key, it is difficult to break the RSA keys.

So far, no one has been able to break RSA keys. Once we are in the era of quantum computers, it will be possible to factorize large numbers in a fast time and, consequently, RSA keys may be broken and the security of the information systems would be at risk. Therefore it will be necessary to think of a post-quantum cryptography [8].

Quantum Image Processing (QIP) using Quantum image-based data security techniques are more efficient and provide higher security than conventional image processing. The processing speed and the amount of data processed in image processing are the most advantages guaranteed by QIP. The strength of this approach is subject to quantum physics rules which we do not have in classical image processing.

The rest of this paper is organized as follows. In 2, Literature review is presented. In 3, Basics of quantum computer are presented. In 4, Computation using Quantum computer will be described. In 5, The quantum computer and data security are presented. And finally, in 6 The conclusion.

2 Literature review

It is noteworthy that after a rising technological revolution, it has been quite a while since a scientific work on information processing using quantum physics was published [2, 3, 9].

In [10], the authors presented in their survey on quantum computing the basic components that are required to build a real quantum computer. They also have pointed out the basics of each capability to be translated from a classical environment to a quantum environment and vice versa.

In [11], the authors proposed initial ideas and several schemes for quantum computing architectures that satisfy the physical constraints; the architectures restrict the way to map the logical qubits used to describe the algorithm to the physical qubits to realize the corresponding functionality.

In a paper [12], the authors presented the difference between classical and quantum cryptography and have pointed out the implications of quantum computing in current cryptography and introduced the basic post-quantum algorithm and said that it deals with different quantum keys distribution methods and mathematically based solutions.

In [13], the authors discussed the connection between the strong testing, which they referred to as the purity testing and the quantum ciphertext authentication (QCA), and have reported that it may offer higher security.

In [14], the authors reported on the relevance of using quantum computing in the biological sciences and how the problems posed by quantum algorithms in this area could provide increased computational efficiency, which was much lacking.

In [15], the authors mentioned that the major challenges faced in designing a quantum computer are related to the errors generated by the quantum computer during computation, which decreases its efficiency. To minimize these errors, they suggested the deployment of quantum error-correcting code, and to achieve fault-tolerance, it will be important to make advancement in engineering.

In their work [16], the authors compared different methods of image storage, image representations and image retrievals in a quantum system. They have also presented and discussed the advancement in quantum image processing.

Another review article[17], the authors reported an outline of the QIMP as well as a diagnostic analysis of the enhancement of existing models of quantum image representations, the design of quantum algorithms to solve sophisticated operations, and the further development of physical hardware and software architecture to both capture and manipulate quantum images.

In a recent paper [18], the authors highlighted that a prominent example is secure communication and presented the cryptographic requirement of transmitting confidential messages from one location to another.

In [19], the authors presented in their work the possibilities that quantum mechanics can add to strengthen the cypher code to be unbreakable.

In [20], the authors discussed details in their work towards applying quantum computation to image processing which can be improved through the application of Quantum Computing.

In a recent work [21], the authors presented in their paper the role of quantum mechanics on image and data processing, they presented three quantum algorithms for comparing the similarity between two quantum images and declared that their proposed algorithms achieve exponential acceleration than the existing quantum and classical methods in all three cases.

Table 1 represents a summary of the different works of the selected articles in literature in relation to our present paper.

3 Basics of a quantum computer

The item *bit* is the basic element for data representation in conventional computer science, which can only have one value among two values 0 or 1. In quantum physics, a third situation is possible, the *qubit* (quantum bit), which represents the new item for the basic element of data representation, can be in a superposition of two states $|0\rangle$ and $|1\rangle$ that can be represented by (1):

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{1}$$

where α , $\beta \in \mathbb{C}$ with $|\alpha|^2 + |\beta|^2 = 1$ and $|\psi\rangle$ is a unit vector in a complex vector space of dimension 2, whose basis vectors are denoted as in (2):

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad and \quad |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{2}$$

A geometric representation in a three-dimensional sphere called Bloch Sphere [22] (see Figure 1), is used to describe many operations of a single *qubit*, for multiple qubits, there is no generalization known of the Bloch sphere. In such a situation, in the whole sphere, we can have an infinity of information represented by an infinity of states.

Table 1: A Review Summary

Focused study / Releted work	Quantum computer architecture	Quantum cryptography	Quantum Image processing	Other Quantum interests
Paper [10]	✓			
Paper [11]	✓			
Paper[12]		✓		
Paper [13]		✓		
Paper [14]				✓
Paper [15]	✓			
Paper [16]			✓	
Paper [17]		✓		
Paper [18]		✓		
Paper[19]		✓		
Paper [20]		✓		
Paper [21]		✓		
Present paper	✓	✓	✓	

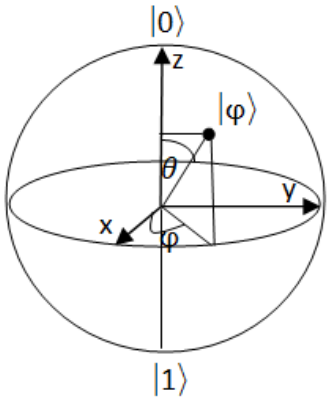


Figure 1: *Qubit* representation by Bloch Sphere

In computer science, all useful information is represented using a succession of bits of 0's and 1's in low-level codes that can be decoded using low-level to high-level transcription programs for different purposes, such as displaying, reading and writing or even other useful purposes, each bit has to be 0 or 1 but not both.

By using a superposition of two states $|0\rangle$ and $|1\rangle$, the *qubit* have the potential to represent both values 0 and 1 at the same time in a state $|\psi\rangle$ that can be noted in (3).

$$|\psi\rangle = \frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle \tag{3}$$

In Equation (3), at the moment of measurement, the state $|\psi\rangle$ will be $|0\rangle$ with a probability equal to $(\frac{1}{\sqrt{2}})^2$ or $|1\rangle$ with a probability equal to $(\frac{1}{\sqrt{2}})^2$, and this is known in quantum physics as decoherence [23], which has to be exactly the same *bit* case of coventional computer.

To make calculations using *qubits*, we must have to leave the measurement at the end to avoid losing the state superposition due to *qubit* decoherence, which means we must maintain the *qubit* coherence as long as possible, for this reason, we should eliminate the constraints that help the *qubit* decoherence and especially minimize the *qubit* interaction with its environment. In [24], was mentioned five requirements, one of them being to minimize the interaction of the *qubit* with its environment, otherwise, the computation will collapse and then the complete system will collapse.

If we need a numerical analysis of the possible coherence time of a *qubit*, in an experiment [25], where the authors made atoms interact with photons one by one, measured the time it takes to observe progressive decoherence, and they obtained a time of about 100 μ seconds for only ten atoms, and this was considered too small to be measured which means that when we increase the number of atoms, the time will decrease exponentially and this will affect the coherence of the *qubit*.

For a given computation time and considering the microprocessor's technological evolution; with 100 μ seconds we can perform a lot of computations, but considering the gate's latency, we will have to increase this time so that it is appropriate to a given computation.

We can have a comparison with the frequency of a conventional processor to see how many quantum processes it can manage, for that purpose, we can convert the qubit coherence time into a frequency that we can obtain a ratio $S_{C/Q}$ that we can calculate using (4); a number which represents the number of cycles that a conventional processor can handle.

$$S_{C/Q} = \frac{F_{CP}}{F_{DQub}} \tag{4}$$

where:

- F_{CP} is the coventional Processor frequency,
- F_{DQub} is the *qubit* decoherence rate,
- $S_{C/Q}$ represents the speed ratio between the conventional processor frequency and the *qubit* decoherence rate.

If we consider a conventional processor frequency of 3.2 GHz, and if we transform the *qubit* coherence time into a frequency, we will have 10 KHz. By using (4), we can calculate the ratio between the conventional processor frequency and the coherence frequency that gives us 320000 times faster, which means that the conventional processor can manage 320000 *qubit's* cadencies (see Figure 2).

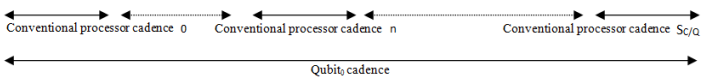


Figure 2: Illustration of cadency of $S_{C/Q}$ *qubits* by a conventional processor rate time

To increase the number $S_{C/Q}$ it is only sufficient to increase the frequency of the conventional processor, or to reduce the decoherence rate (i.e. to extend the decoherence time), and for that purpose Chapter 10 of the book [26] discusses various approaches to extend the decoherence time effectively.

The conventional computer's technological evolution makes it possible to have the classical processor's frequency increased. But it is a difficult mission to reduce *qubit* decoherence rate because it is not easy to isolate the *qubit* from its environment.

3.1 Concepts and Physical realisation of the qubit and the quantum computer

To have a quantitative evaluation and to see the measurable size of a *qubit*, in [27],The authors succeeded in fabricating a 60 μm *qubit* in a 6 GHz resonant system. The realized *qubit* is considered giant compared with the size of an atom, which means that with this *qubit* size, we have the horizon opened to the fabrication of quantum electronic chips that are also considerably larger than the size of an atom.

A *qubit* is a fundamental unit of data representation and a basic element of the quantum computer's hardware architecture. This architecture, quite particular compared to the classical one, needs to meet certain engineering requirements to be able to engineer a quantum computer. These requirements are linked to quantum physics specificities, which cannot be found in the classical one.

One of the relevant concepts in quantum physics is energy, which forms discontinuous scales known as a quantum leap. In their experiment [28], by using Barium, the authors measured the quantum leap, which confirms that, by the quantum physics nature, an electron can occupy two energy levels at the same time; that is known as a states superposition.

For the concept of entanglement, where two electrons or two photons will be linked to each other; namely, if we act on one automatically we act on the other instantaneously, and if one chooses its state whether it is $|0\rangle$ or $|1\rangle$ the other one automatically is in the opposite state. This phenomenon was described as strange and mentioned in a scientific paper [29] known as the EPR paradox, since no velocity, until today, is higher than the velocity of the light.

In a recent experiment [30], Chinese researchers, have proven and demonstrated that even at large distances (1203 km), photons were entangled.

For example, taking two superposed electrons where we have four possibilities $|11\rangle$ $|10\rangle$ $|01\rangle$ $|00\rangle$. The most relevant case is $|10\rangle$ $|01\rangle$ called the entangled states, where the two electrons are irremediably linked to each other, when measuring one of them we necessarily interact with the other one, i.e. the entangled systems are linked each to other, and this is proved by experiments [30, 31].

4 Computation using Quantum computer

Computing with a quantum computer is quite different from computing with a classical computer. The entanglement and the superposition of the states of a *qubit* are very fragile that we can only make measurements at the end of the calculation for fear that the entanglement and the superposition of the states will collapse. To overcome this problem, we can use an ingenious solution which consists in duplicating thousands of copies of the *qubit* and making the calculation, two scenarios are to be considered; we check the calculation and it is good so we go ahead, otherwise, we stop it and start again.

When doing computations, we cannot escape from the computational errors which are so far in the range of $\frac{1}{1000}$, and by taking into account the number of *qubits* to be handled, we can expect to have one million of *qubits* to carry out an appropriate computation.

If we assume having a decoherence time four times the time observed in the experiment [25], i.e. a decoherence time of 400 μ seconds, which represents a frequency of 2.5 kHz and let's take (4) to calculate the quotient $S_{C/Q}$ for a conventional processor frequency of 3.2 GHz, we obtain 1280000 cadences. If we duplicate copies of the *qubits* and we verify at a given moment if the calculation is good or not, and let's say that everything is good and that the calculation we are doing is correct without collapsing as shown in Figure 3, in this case, the total time T is equal to the time at the end of the computation, i.e. equal to the *qubit* decoherence, but if we were in the second case (see Figure 4), where the computation is not correct and therefore we stop it and start again, in this case, the total time is $T1$ and is different from T, it will be equal to $m \times$ Conventional Processor instruction time, where m represents the number of repetitions.

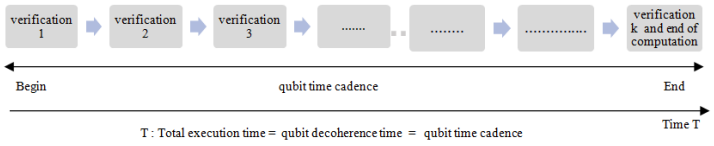


Figure 3: Time computation T: Computation verification without errors

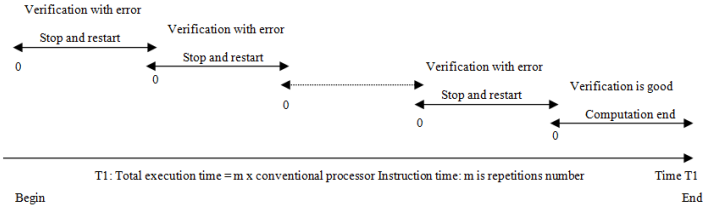


Figure 4: Time computation T1 : Computation verification error with stop and restart

We point out that even though the calculation is correct, this does not mean that there are no errors; this is due to the probabilistic measurements and the superposed states that occur each with a probability. Also, the quantum computer runs on a different clock separated from that of a conventional computer.

Quantum programming cannot replace conventional programming completely, it is estimated only for solving some complex

problems that are difficult to solve using a classical computer. In [32], a list of 50 sample problems was compiled that may be interesting to use as examples for writing quantum programs.

Quantum computer can be considered as a co-processor for the conventional computer, in which our programs are divided into two different parts; a conventional programming part for the conventional computer side, and another quantum programming part for the quantum computer as shown in Figure 5.

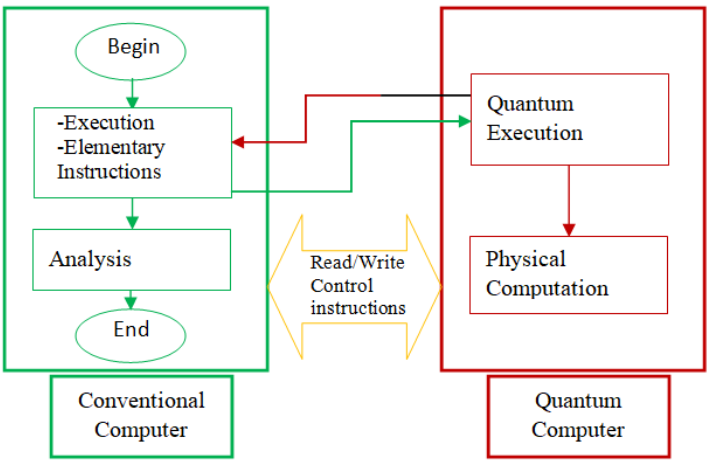


Figure 5: Illustration of two programming parts

5 The Quantum computer and data security

5.1 *Internet Security and the RSA encryption method*

Internet at high speed is difficult to define, as what constitutes "high speed" changes over time as technology advances and applications develop. There are advantages and weaknesses in each technology. The final selection should attend to the specific requirements of the Internet application, mainly determined by the factors that affect Internet objectives: low power, low cost, high number of devices, and medium-long range [33]. The high-speed of the Internet has affected many areas, especially those that use it as a medium for communication and data exchange.

In the era of information technology, data exchange has rapidly progressed, especially with the availability of high-speed communication such as the internet. However, data exchange is subject to hackers' attacks to hack the information, which has led us to use technical solutions to protect our information systems as well as the data exchanged on the Internet.

The RSA encryption method [34, 35], is an asymmetric algorithm used to exchange data on the Internet. This algorithm depends on two keys: a public key and a private key. No one has been able to break these keys, but once we have the quantum computer, these keys may be compromised, and the information system will then be at risk.

To address this risk, we develop quantum cryptography methods to secure the information system and the data exchange algorithm.

The technology used to design a qubit plays a huge role in the accuracy of computations performed with a quantum computer, but the question arises seriously: is quantum computing accurate? In order to provide an answer, we need to look at one of the latest works done. In [36], the authors mentioned that quantum operations are accurately characterized using gate set tomography (GST) , resulting in average single-qubit gate fidelities of up to 99.95%, average two-qubit gate fidelities of 99.37%, and two-qubit preparation/measurement fidelities of 98.95%. They precised that these three metrics are approaching the performance demanded in fault-tolerant quantum processors. which means that by the time the first generation of quantum computer appears, we will be able to achieve 100% accuracy.

To take advantage of quantum physics, we can benefit from the concepts of entanglement and superposition of qubit states, and since the measurement can only be performed at the end of the computation to avoid the system collapse. We have simulated an *8-qubits* (i.e. a *quByte*) code, Based on the original sample code example [37], to illustrate how can we do to perform a secure data exchange.

The spy hunter continuously checks and tunes the transmission channel to get accurate information by hacking the data. They use sophisticated tools and devices to get powerful that give them techniques to decrypt any data. With the quantum computer, spies can easily intercept and decipher data, and it is of great interest to understand quantum physics to know how to manage a secure transmission medium.

By capturing the data, they behave as if they had not caught it and send it back to the receiver (Bob). When Bob receives the data, and before reading, Alice informs him that she had applied a quantum gate (and maybe a combination of quantum gates) to the data, and then Bob applies the inverse of the quantum gate to the received data and compares it with Alice's data. If their data are identical, Bob and Alice conclude that the data has not been hacked, otherwise, they understand that it has been hacked.

In the case of applying quantum gates, and even if the hackers intercept the data, they will not be able to read it because they will never be able to guess the combination of quantum gates applied to the data before sending it.

Figure 6 represents the result of the simulation using QCEngine, and Figure 7 represents the result of the simulation using Qiskit.

For illustration presented on Algorithm 1, Alice wants to send some data to Bob (as example here data are an arbitrary information coded on 8 Qubits).

It is necessary to secure our information systems by thinking about post-quantum-cryptography solutions; i.e. working a lot in the technological field of quantum cryptography to prevent the RSA key from being broken because 1024-bit or even 2048-bit keys will be broken using quantum computers.

One day we will be in an era where the use of the quantum computer will be available to everyone, and then the spies of the computer will have the possibility to factorize long numbers, thereafter, they will be able to break the RSA algorithm keys. However, we have seen from the presented example that it will be guaranteed to ensure the security of our information systems and the strength of the RSA keys, while based on the characteristics of quantum physics.

The field of quantum programming is expanding, and we look forward to the first real and commercialized quantum computer

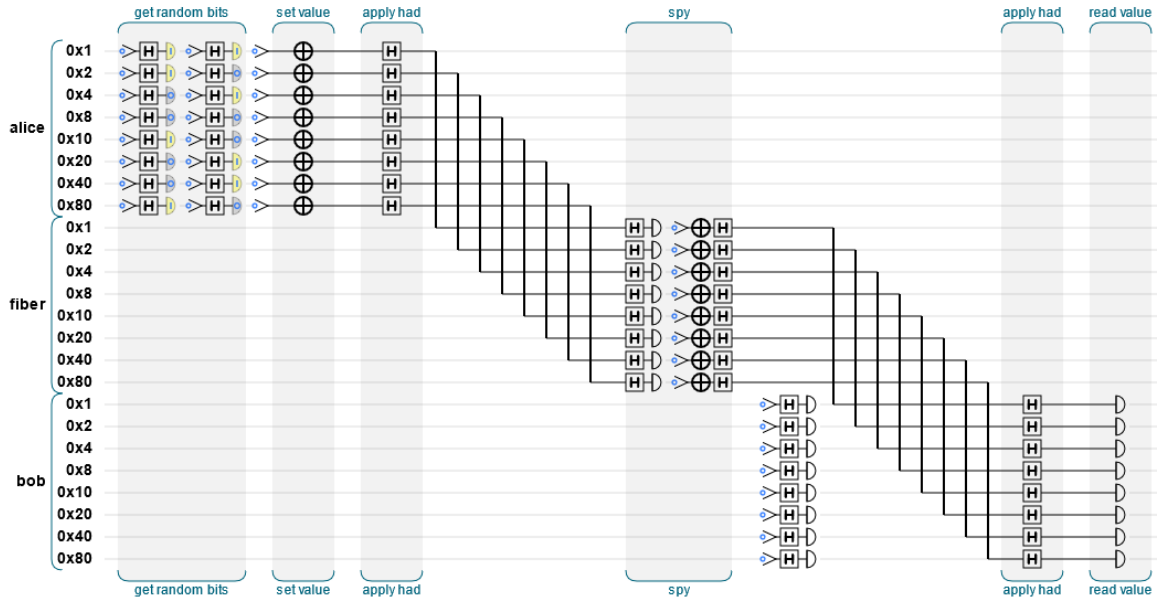


Figure 6: Illustration of spy hunter quantum program simulation using QCEngine

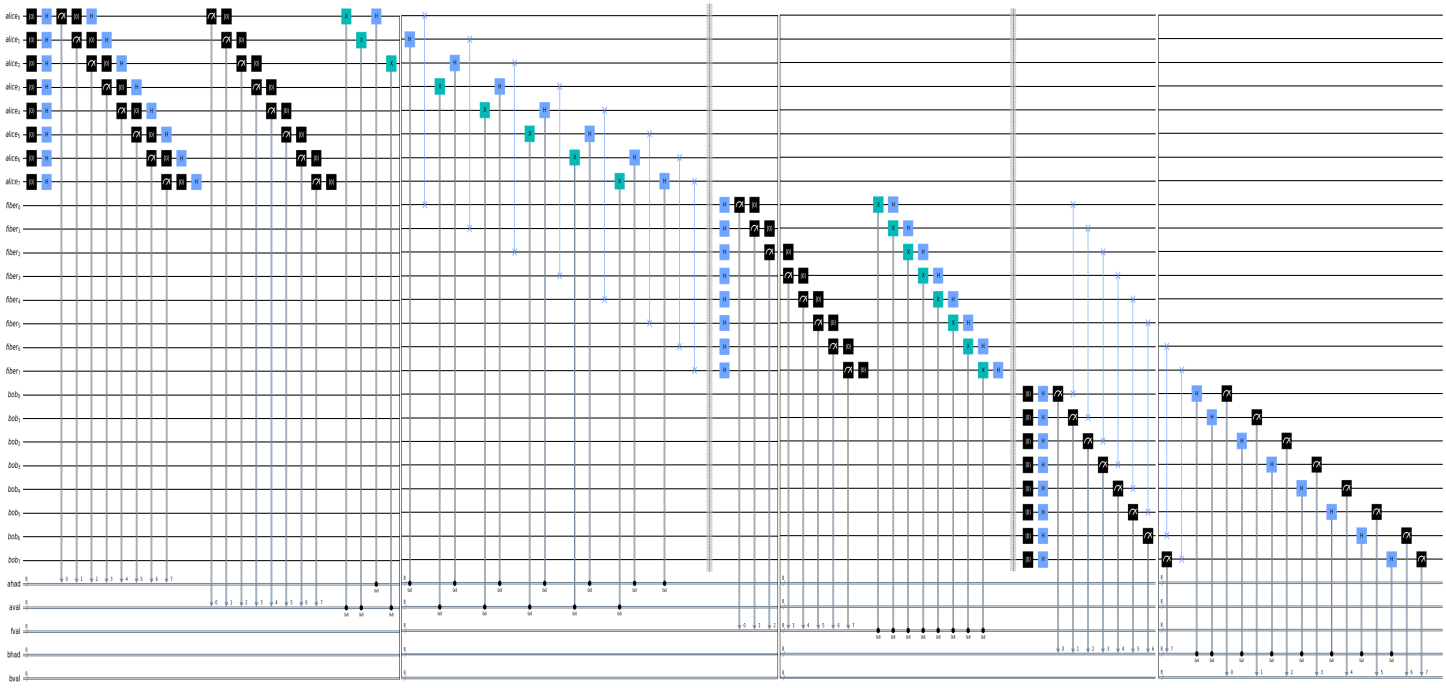


Figure 7: Illustration of spy hunter quantum program simulation using Qiskit

Algorithm 1: Spy Hunter

```
#Use Alice's QPU to prepare Alice's qubit
qc.measure(alice, alice_had);
qc.measure(alice, alice_val);
if (apply a HAD) then
    qc.h(alice)=alice_had
    else
        qc.x(alice)=alice_val
    end
end
# Send the qubit!
qc.swap(alice, fiber)
# Activate the spy
spy_is_present = True
if (spy_is_present) then
    qc.barrier()
end
spy_had = True
if (spy_had) then
    qc.h(fiber)
end
qc.measure(fiber, fiber_val); qc.reset(fiber);
if (fiber_val) then
    qc.x(fiber)
end
if (spy_had) then
    qc.h(fiber)
end
qc.barrier()
# Use Bob's QPU to prepare data...
# Receive the qubit!
qc.swap(fiber, bob)
if (bob_had) then
    qc.h(bob)
end
qc.measure(bob, bob_val)
# If the setting matches and the value
#does not, there's a spy!
counts = result.get_counts(qc);
print('counts:',counts)
caught = False
for key,val in counts.items() do
    alice_had,alice_val,f,bob_had,bob_val =(int(x)
end
for (x in key.split(' ')) do
    if (alice_had == bob_had) then
        if (alice_val != bob_val) then
            print('Caught a spy!') caught = True
            else
                not caught: print('No spies detected.')
            end
        end
    end
end
```

to apply everything we know as quantum algorithms, but we may state that this is only for certain applications that need fast solu-

tions, and anything we use in a conventional computer, for example, Word or all our existing applications cannot be replaced by quantum programs.

5.2 Quantum Image Security

In cryptography, encryption is the process of hiding information to make it unreadable without special knowledge. In steganography or in watermarking, the technique used to hide information seems to be more secure because it is not easily detected by hackers. However, its major problem is a large amount of information that can be hidden inside an image without distortion of its visual imperceptibility.

In 2012, the authors of [38] have proposed the first quantum image security protocol, by a scheme of watermarking and authentication of quantum images (WaQI), and was based on restricted geometric transformations on the images.

In 2013 [39], a novel watermarking scheme based on Quantum Wavelet Transform (QWT) was proposed , and in 2014 [40],another watermarking scheme based on Hadamard transform was proposed .

In [41], the authors have proposed an improvement of the watermarking protocol proposed in [39].

In the field of quantum image encryption, methods are classified into spatial domain-based strategies or frequency domain-based strategies, on which quantum image encryption algorithms focus [42].

5.3 Multilevel Quantum Systems

In many works in literature, various models of quantum image representation have been proposed, and many of the proposed models are based on a binary quantum system.

Instead of a binary quantum system, a ternary quantum system is a new tendency to represent and process the quantum image.

A ternary quantum system is a 3-level quantum system having three mutually orthogonal states $|0\rangle$, $|1\rangle$ and $|2\rangle$ that form a basis in the Hilbert space $\mathcal{H}^3$ called qutrit.

As in a quantum binary system for a qubit, the superposition state of a qutrit can be formulated as:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle + \gamma|2\rangle \tag{5}$$

with $|\alpha|^2 + |\beta|^2 + |\gamma|^2 = 1$

In [43], the authors have proposed an RGB color image represented and stored on a ternary (3-levels) quantum system. They postulate that "images on quantum computers can be represented in the form of a normalized state which captures information about colors ($|c\rangle$) and their corresponding positions ($|p\rangle$) in the images". The image can be represented as:

$$|I\rangle = |c\rangle \otimes |p\rangle \tag{6}$$

$$|I\rangle = \frac{1}{2^n}\sum_{i=0}^{2^n-1}(cos\theta_i|0\rangle + sin\theta_i|1\rangle) \otimes |p\rangle \tag{7}$$

where, $\theta_i \in [0, \pi]$, i= 0, 1, 2,. . . , 2^{2n} - 1 and $\theta = \theta_0, \theta_1, \ldots, \theta_{2^{2n}-1}$ is the vector of angles encoding colors.

In the case of the qutrits-based quantum system, According to Klimov qutrit phase model [44] can be represented as:

$$|I\rangle = \frac{1}{3^n}\Big(\sum_{i=0}^{3^{2n}-1} sin(\frac{\xi}{2})cos(\frac{\theta}{2})|0\rangle + \exp^{i\phi_{01}} sin(\frac{\xi}{2})cos(\frac{\theta}{2})|1\rangle + \exp^{i\phi_{02}} cos(\frac{\xi}{2})|2\rangle\Big) \otimes |p\rangle \tag{8}$$

where, θ and ξ represent the magnitudes of the components of $|\phi\rangle$, and ϕ_{01} is interpreted as the phase of $|0\rangle$ relative to $|1\rangle$ and analogously for ϕ_{02}.

The multilevel quantum system was implemented and designed in various works for different image processing purposes. In their work [45], the authors designed a circuit-level implementation of the quantum multilevel threshold-based color image segmentation technique.

Using multilevel quantum system have more advantages compared to binary quantum system in the most cases that can be stated for example:

> With the same amount of physical resources, the use of higher-dimensional quantum states increases exponentially the available Hilbert space.
>
> An n-qutrit quantum system can be represented by a superposition of 3^n basis states, thus a quantum register of size n can hold 3^n values simultaneously, in the other side an n-qubits register can only hold 2^n values.
>
> More efficient ternary logical gates implementation are used in multilevel quantum system.
>
> For representing, storing and processing the color images, ternary quantum system is more beneficial than binary quantum system.

6 Conclusion

In this paper, we describe how we can profit from the quantum computer to strengthen the security of our data on the Internet using quantum physics, based on the principles of quantum physics that we do not have in classical computers.

The unequalled power of the quantum computer led us to think of using it to represent and process big-size images. We have presented the new tendency of using ternary quantum systems compared to binary ones.

In future work, we will use the idea of sending a scrambled image with a ternary quantum system, which will guarantee us better security and an exponential gain in processing time and memory.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The authors acknowledge the use of IBM Quantum services in this article. The views expressed are those of the authors, and do not reflect the official policy or position of IBM or the IBM Quantum team. They also acknowledge the support of the DGRSDT of the Ministry of Higher Education and Scientific Research of the Algerian Government.

References

[1] T. Nouioua, A. H. Belbachir, "The Quantum Computer and the Security of Information Systems," in 2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI), 1–9, 2021, doi:10.1109/ICRAMI52622.2021.9585929.

[2] R. P. Feynman, "Simulating physics with computers," Int. J. Theor. Phys., **21**, 467–488, 1982.

[3] R. P. Feynman, "Quantum mechanical computers," Foundations of Physics, **16**, 507–531, 1986.

[4] P. W. Shor, "Algorithms for Quantum Computation: Discrete Logarithms and Factoring," in in Proceedings of the 35th Annual Symposium on the Foundations of Computer Science, 124–133, 2021.

[5] P. W. Shor, "Polynomial-time algorithms for prime factorization anddiscrete logarithms on a quantum computer," SIAM review, **41**, 303–332, 2020.

[6] D. Deutsch, R. Jozsa, "Rapid solution of problems by quantum computation," SIAM review, **439**, 553–558, 2020.

[7] L. K. Grover, "A fast quantum mechanical algorithm for database search," in In Proceedings of the twenty-eighth annual ACM symposium on Theory of computing, 212–219, 1996.

[8] J. A. Buchmann, D. Butin, F. Göpfert, A. Petzoldt, "Post-Quantum Cryptography: State of the Art," Lecture Notes in Computer Science, **9100**, 2016, doi:10.13140/RG.2.2.29502.18243.

[9] M. A. Nielsen, I. L. Chuang, Quantum Computation and Quantum Information, Cambridge University Press, 2011.

[10] S. M. Pranav, M. Ritwik, "A Comprehensive but not Complicated Survey on Quantum Computing," in In Proceedings of the twenty-eighth annual ACM symposium on Theory of computing, 144–152, 2014, doi:10.1016/j.ieri.2014.09.069.

[11] D. Arighna, D. Gerhard W, W. Robert, "Towards exploring the potential of alternative quantum computing architectures," in Proceedings of the 23rd Conference on Design, Automation and Test in EuropeMarch, 682–685, 2020.

[12] M. Vasileios, V. Kamer, M. D. Zych, A. Jøsang, "The Impact of Quantum Computing on Present Cryptography," International Journal of Advanced Computer Science and Applications, **9**, 2018.

[13] Y. Dulek, F. Speelman, "Quantum Ciphertext Authentication and Key Recycling with the Trap Code," Leibniz International Proceedings in Informatics, 1–17, 2020.

[14] P. S. Emani, J. Warrell, A. Anticevic, "Quantum computing at the frontiers of biological sciences," Nat Methods, **18**, 701–709, 2021, doi:10.1038/s41592-020-01004-3.

[15] D. C. Doimari, "computing versus classical computing in the field of Biological Sciences," E-ZINE OF BIOLOGICAL SCIENCES, 2020, doi:10.1038/s41592-020-01004-3.

[16] S. Chakraborty, S. B. Mandal, S. H. Shaikh, "Quantum image processing: challenges and future research issues," International Journal of Information Technology, 1–15, 2018.

[17] F. Yan, S. E. Venegas-Andraca, K. Hirota, "Toward implementing efficient image processing algorithms on quantum computers," Soft Computing, 1–13, 2022.

[18] P. Christopher, R. Renato, "Renato,Security in quantum cryptography," Rev. Mod. Phys, 2022, doi:10.1103/RevModPhys.94.025008.

[19] Bennett, H. Charles, G. Brassard, A. K. Ekert, "Quantum Cryptography," Scientific American 267, **4**, 50–57, 1992.

[20] G. Beach, C. Lomont, C. Cohen, "Quantum image processing (QuIP)," in 32nd Applied Imagery Pattern Recognition Workshop, Procedongs, 39–44, 2003, doi:10.1109/AIPR.2003.1284246.

[21] Y. Liu, Z. Qi, Q. Li, "Comparison of the similarity between two quantum images," Sci Rep 12, 50–57, 2022, doi:10.1038/s41598-022-11863-9.

[22] M. A. Nielsen, I. L. Chuang, Quantum Computation and Quantum Information, Cambridge University Press, 2010.

[23] C. Kiefer, E. Joos, "Quantum Future From Volta and Como to the Present and Beyond," Lecture Notes in Physics, 1999, doi:10.1007/BFb0105342.

[24] D. P. DiVincenzo, "The Physical Implementation of Quantum Computation," Fortschr. Phys, **48**, 771–783, 2000, doi:10.1002/1521-3978(200009)48:9/11⟨771::AID-PROP771⟩3.0.CO;2-E.

[25] M. Brune, . Hagley, J. Dreyer, X. Maître, A. Maali, C. Wunderlich, J. M. Raimond, S. Haroche, "Observing the Progressive Decoherence of the "Meter" in a Quantum Measurement," Phys. Rev. Lett, **77**, 301, 1996.

[26] N. Mikio, O. Tetsuo, Quantum Computing - From Linear Algebra to Physical Realizations, CRC Press, 2010.

[27] A. D. O'Connell, M. Hofheinz, M. Ansmann, B. Radoslaw, M. Lenander, L. Erik, M. Neeley, D. Sank, H. Wang, W. Martin, J. Wenner, M. John, A. Cleland, "A. Quantum ground state and single-phonon control of a mechanical resonator," Nature, **464**, 697–703, 2010, doi:10.1038/nature08967.

[28] T. Sauter, W. Neuhauser, R. Blatt, P. E. Toschekt, "Observation Of Quantum Jumps," PHYSICAL REVIEW LETTERS, **57**, 697–703, 1986.

[29] A. Einstein, B. Podolsky, N. Rosen, "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? ," PHYSICAL REVIEW, **47**, 1935.

[30] J. Yin, Y. Cao, Y.-H. Li, S.-K. Liao, L. Zhang, J.-G. Ren, W.-Q. Cai, W.-Y. Liu, B. Li, H. Dai, G.-B. Li, Q.-M. Lu, Y.-H. Gong, Y. Xu, S.-L. Li, F.-Z. Li, Y.-Y. Yin, Z.-Q. Jiang, M. Li, J.-J. Jia, G. Ren, D. He, Y.-L. Zhou, X.-X. Zhang, N. Wang, X. Chang, Z.-C. Zhu, N.-L. Liu, Y.-A. Chen, C.-Y. Lu, R. Shu, C.-Z. Peng, J.-Y. Wang, J.-W. Pan, "Satellite-Based Entanglement Distribution Over 1200 kilometers ," Science, **356**, 1140–1144, 2017, doi:10.48550/arXiv.1707.01339.

[31] A. Aspect, J. Dalibard, G. Roger, "Experimenta lTest of Bell's Inequalities Using Time-Varying Analyzers ," Physical Review Letters, **49**, 1982.

[32] "quantum algorithm zoo," 2020.

[33] G. D. Campo, I. Gomez, G. Canada, L. Piovano, A. Santamaria, "Guidelines and criteria for selecting the optimal low-power wide-area network technology ," LPWAN Technologies for IoT and M2M Applications, 281–305, 2020, doi:10.1016/b978-0-12-818880-4.00.

[34] N. Shireen, F. Mohammed, "RSA Public Key Cryptography Algorithm. A Review ," International Journal of Scientific and Technological Research, **6**, 187–191, 2017.

[35] X. Zhou, X. Tang, "Research and implementation of RSA algorithm for encryption and decryption," 2011, doi:10.1109/IFOST.2011.6021216.

[36] M. T. Madzik, S. Asaad, A. Youssry, B. Joecker, K. M. Rudinger, E. Nielsen, K. C. Young, T. J. Proctor, A. D. Baczewski, A. Laucht, V. Schmitt, F. E. Hudson, K. M. Itoh, A. M. Jakob, B. C. Johnson, D. N. Jamieson, A. S. Dzurak, C. Ferrie, R. Blume-Kohout, A. Morello, "Precision tomography of a three-qubit donor quantum processor in silicon ," Nature, **601**, 348, 2022, doi:10.1038/s41586-021-04292-7.

[37] E. R. Johnston, N. Harrigan, M. Gimeno-Segovia, Programming Quantum Computers: Essential Algorithms and Code Samples, O'Reilly Media, 2019.

[38] A. Iliyasu, P. Le, F. Dong, K. Hirota, "Watermarking and authentication of quantum images based on restricted geometric transformations," Inf. Sci., **186**, 126–149, 2012.

[39] X. Song, S. Wang, S. Liu, A. A. El-latif, X. Niu, "A dynamic watermarking scheme for quantum images using quantum wavelet transform," Quantum Information Processing, **12**, 3689–3706, 2013.

[40] X. Song, S. Wang, A. A. El-latif, X. Niu, "Dynamic watermarking scheme for quantum images based on Hadamard transform," Multimedia Systems, **20**, 379–388, 2014.

[41] Y. Yang, P. Xu, J. Tian, H. Zhang, "Analysis and improvement of the dynamic watermarking scheme for quantum images using quantum wavelet transform," Quantum Information Processing, **13**, 1931–1936, 2014.

[42] F. Yan, A. M. Iliyasu, P. Q. Le, "Quantum image processing: A review of advances in its security technologies," International Journal of Quantum Information, **15**, 1730001, 2017.

[43] S. Chakraborty, S. B. Mandal, , S. H. Shaikh, L. Dey, "Ternary quantum circuit for color image representation," In Advanced Computing and Systems for Security, 95–108, 2017.

[44] A. B. Klimov, L. L. Sánchez-Soto, H. de Guise, G. Bjork, "Quantum phases of a qutrit," Journal of Physics A, **37**, 4097–4106, 2000.

[45] S. Chakraborty, S. B. Mandal, S. H. Shaikh, "Design and implementation of a multivalued quantum circuit for threshold based color image segmentation," Intell. Decis. Technol., **12**, 251–264, 2018.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 5, 79-90 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

DEM Models Calibration and Application to Simulate the Phosphate Ore Clogging

Bouassale Nasr-Eddine*, Sallaou Mohamed, Aittaleb Abdelmajid, Benaissa Elfahim

Innovation and Systems Engineering Laboratory, M2I Team, ENSAM Meknès Moulay Ismail University, Mèknes, Morocco

ARTICLE INFO

Article history:
Received: 30 July, 2022
Accepted: 21 September, 2022
Online: 10 October, 2022

Keywords:
Phosphate ore
Calibration
DEM
Contact model
Angle of Repose
EDEM®

ABSTRACT

In different areas of industry (mining, food processing, pharmaceutical, manufacturing...), the problem of grain, aggregate and clusters flow arises during the handling activities especially loading and unloading. Thus, the study and control of the parameters that govern the flow of the granular medium and its interaction with its environment are key parameters to achieve the desired operational excellence and performance of these activities. The adhesion of granular materials on various surfaces of equipments (trucks surfaces, hoppers, silos…) is one of the major problems facing mining companies. In this paper we presented the Calibration of the Discrete Element Method (DEM) parameters for modeling phosphate ore built on the identification of the repose angle. This will help us to specify the correct inputs parameters that will be introduced for the modeling of adhesion phenomena. First we introduced the model contact which allow us to well define contact between phosphate particles-particles (or clusters-clusters) and phosphate particles (clusters) with tipper surface taking into consideration the cohesive and plastic nature of the contact. Secondly, we presented a calibration method based on the determination of the repose angle of the phosphate ore. This method allowed us, using fractional factorial designs and Box Behnken designs, to determine the optimal parameters for a more accurate simulation. Thirdly, we studied the impact of particle velocities on the tipper surface during charging and discharging of phosphate ore. This study allowed us to predict the areas most affected by the abrasion-erosion phenomenon caused by the impact of particles on the tipper surface. This calibration method allowed us to identify the optimal values for the key parameters that will be used later in the modelling of the phosphate clogging phenomenon on the surfaces of transport truck tippers in the mines.

1. Introduction

In recent years, the global fertilizer market has been characterized by a significant increase in consumption (1.7% /year since 2011), supported mainly by Asia which dominates the market and concentrates about 2/3 of global production and consumption [1]. The lowest possible purchase cost has become an important competitive factor which has led to the relocation of fertilizer production to areas with significant resources at low exploitation costs (Russia, Qatar, Iran, North Africa and probably in the long term the USA with the exploitation of shale gas). Since fertilisation contributes to both the quantity and quality of agricultural production on which the world's food system is based, phosphate mining has become the linchpin of the world food security. Studies show how the consumption of phosphate rock, the origin of phosphorus which is a key element in the manufacture of fertilisers, will continue to increase over the coming years. In Morocco, it is estimated that phosphate production will reach 262 Mt in 2050 [1–4].

Clogging is one of the main problems that reduce the efficiency of phosphate extraction. This phenomenon results from the adhesion of the phosphate to different surfaces of the transport truck tipper. This blocked portion intensifies during the winter period, especially in the bottom and sides of the tipper [5]. This problem is called granular clogging.

The phosphate rock, used as raw material, comes from Benguerir mine. From a stratigraphic point of view, the Benguerir phosphate series is formed by an alternation of phosphate rock levels and sterile silico-carbonate levels (intercaler) (Figure 1) and the ensemble is capped by a carbonate slab called the Thersity slab.

It is very difficul to explore the charging and discharging for extraction in situ because of high-risk work conditions [6,7]. Thus,

*Corresponding Author: Bouassale Nasr-Eddine, n.bouassale@umi.ac.ma

to minimise the cost of experiments and save time, a number of Discrete Element Method simulations (DEM) have been developed [8–12]. With DEM, we can investigate the mechanical and kinematic behaviour of granular materials. In recent years, researchers have carried out DEM simulations with grain shapes and sizes that are complex enough and large enough to be closer to reality, given that DEM was invented much earlier.[13]. the discrete elements method remains the most used calculation tool for the simulation of the various phenomena that appear during the handling of granular materials [14–21]. In the literature, and up to today, we have found three principal approaches used in the simulationt of problems of granular systems: the Monte Carlo approach [22-27], the cellular automation approach [28-31] and the DEM approache [32-34].The cell automation method and the Monte Carlo approaches have enabled the simulation of problems related to granular materials by taking into account only the geometric and statistical aspect, but without considering the real mechanical interactions between grains.. In contrast, the DEM method takes into account the interaction between particles by introducing different contact models. Thus it accurately reflects the mechanical phenomenon of the grain packing. It can generate differents particle size distributions, with varying geometric characteristics and surfaces, as well as various material ratios from the analyzed sample.

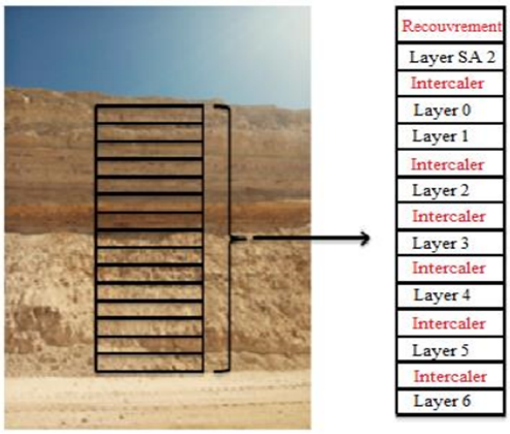


Figure 1: Layered decomposition of Benguerir phosphate [5]

However, what influences the precision of the simulation is the parameter settings [22]. Precise simulations can only be performed if the parameters of the granular materials are well defined. The settings of the process parameters, the nature of the contacts that govern the phenomenon and the behaviour of the materials influences the simulation results [23].

A crucial phase in the development of a discrete element method is the identification of the simulation parameters, namely the specific granular material parameters (Young's modulus,bulk density and Poisson Coefficient) and the parameters related to the nature of the contacts (the particle-wall and particle-particle friction coefficients and the coefficients of restitution). As a result, a calibration of the setup parameters is strongly required to obtain accurate values for the DEM simulation [24–27].

Simulation on software in general requires values for the parameters of the model. For this purpose, there are two main methods for the DEM software to acquire these parameters of contact. The first one is the direct measurement method (DMM), and the second one is the bulk calibration method (BCM) [22,28–31]. The DMM approach is based on measuring values directly on the particle or contact level. Some properties are easily measured while others remain very difficult, depending on the size of the grains. Almost all attempts have been made at the millimetre level or above [32,33]. Although the values are being directly measured with a certain precision, this does not automatically imply that the DEM model will have the same level of precision at the global level [34–36]. This DMM would be precise if the contact is modelled in an accurate way that captures the physical behaviour of the contact and if both shape and size of the grains are precisely modelled [37,38]. Precise modelling of grain shape and size is too complex when modelling real industrial cases. the calculation time and power limit of the simulators tend to lead to the selection of reasonably large grain sizes, while the shape of the particles cannot be accurately modelled [39–41]. The BCM uses an approach where field measurements or laboratory tests are performed to quantify a specific parameter of the bulk granular material. Experiences are reproduced on software as closely as possible to the configuration and procedures in the laboratory or in the terrain. Simulation values are continuously iterated till the intended response corresponds to the measured one. The difference between the real shape and the virtual model of the grains requires a calibration of the contact parameters [30,31,42,43].

In the DEM simulation, the contact model is a crucial factor that affects the precision level of the simulation. In order to get closer to the actual situation, different contact models appeared in a number of several applications [44–53]. In this study, we used two contact models for our simulation. The Johnson-Kendall-Roberts (JKR) is used for the particle-tipper contact, and the Edinburgh-Elasto-Plastic-Adhesion (EEPA) is used for the particle-particle contact [45]. Furthermore, The modelling of the real shape is important in a discrete element simulation, and the more the geometrical characteristics of the real particles are related to the numerical particles, the closer the results of the numerical simulation are to the real behaviour of the material [54–59]. In the ideal case the grain geometry is retrieved using a rotary scanner as illustrated in the work [58] which allows to obtain the contours of a particle with precision.

In this work we calibrated DEM for phosphate ore in order to define the parameters of the contact models between the phosphate particles themselves and the surface of the tipper of the ore transport truck. This work will allow us to model with DEM the phenomenon of phosphate adhesion during unloading. We presented the contact models we used, followed by the calibration method adopted during this work which is based on the calibration of the angle of repose. The size, shape and distribution of the phosphate particles used in the calibration were studied. We used the fractional factorial designs and Box Behnken designs to determine the influence of the different calibration parameters on the angle of repose.

2. Edinburgh-Elasto-Plastic-Adhesion (EEPA) model for the particle-particle contact

The Edinburgh Elasto-plastic adhesion EEPA model was introduced by Morrissey [60], and implemented in EDEM® as an option [53], to model cohesive materials with elasto-plastic behaviour such as phosphate and clay. It is an extension of the

hysteretic spring model but this time with non-linear behaviour. This model is able to highlight the history of stress dependence of cohesive granular materials with uniaxial consolidation experiments [53]. In [74], the author highlighted the ability of EEPA to generate behaviour that is dependent on the history of previous loads, which is not the case for Hertz-Mindlin, since it's an elastic model. For this purpose, we chose the EEPA to model particle-particle contact of phosphate ore.

2.1. EEPA Force: Normal Direction

This model is an extension of the linear hysteretic model of Walton and Braun [61] , allowing tensile forces to develop and also allowing non-linear force-displacement behaviour (overlap) as shown in Figures 2 and 3. The load-displacement curve is defined by the constant pull-off force F_0, the loading branch stiffness k_1, the loading-unloading branch stiffness k_2 , the minimum force F_{min} , the adhesion branch stiffness k_a ,and the plastic overlap (deformation) δ_p.

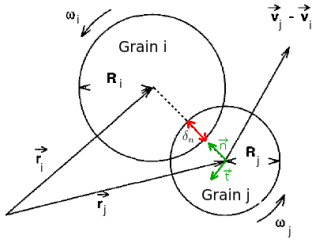


Figure 2 : Formalism used during a contact [62]

The normal force F_n^{EEPA} is composed of two forces, which show its plastic character, namely the hysteretic spring force F_{hys} and the damping force F_{nd} .

$$F_n^{EEPA} = (F_{hys} + F_{nd})\boldsymbol{n} \tag{1}$$

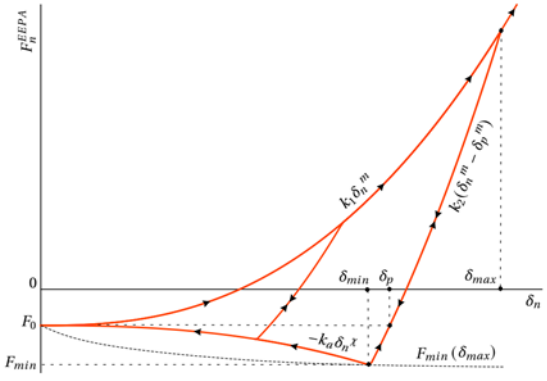


Figure 3 :EEPA model force-displacement relation in the normal direction[63]

where $\boldsymbol{n}$ is the normal unit vector at the point of contact at the centre of the two particles.

$$F_n^{EEPA} = \begin{cases} F_0 + k_1 * \delta_n^m & if\ k_2 * (\delta_n^m - \delta_p^m) \geq k_1 * \delta_n^m \\ F_0 + k_2 * (\delta_n^m - \delta_p^m) & if\ k_1 * \delta_n^m > k_2 * (\delta_n^m - \delta_p^m) \\ F_0 - k_{adh} * \delta_n^\gamma & if\ -k_{adh} * \delta_n^\gamma \geq k_2 * (\delta_n^m - \delta_p^m) \end{cases} > -k_{adh} * \delta_n^\gamma \tag{2}$$

where F_0 is the adhesion force constant, δ_n^m and δ_p^m are respectively the total normal overlap and the plastic overlap. The loading stiffness k_1 is given by:

$$k_1 = \begin{cases} 2E^*R^* & if\ m = 1 \\ \frac{4}{3}\sqrt{R^*}E^* & if\ m > 1 \end{cases} \tag{3}$$

where E^* and R^* are the equivalent Young's modulus and radius.

$$R^* = \frac{R_1 R_2}{R_1 + R_2} \tag{4}$$

R_1, R_2 are radius of particule 1 and 2.

$$E^* = \left(\frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2}\right)^{-1} \tag{5}$$

$E_1, E_2$are Young's modulus of particule 1 and 2.

ν_1, ν_2 are Poisson's ratio of particule 1 and 2.

For our study we took 1.5 as the value of the non-linear slope exponent n. λ_p is defined as the contact plasticity ratio and is given by:

$$\lambda_p = \left(1 - \frac{k_1}{k_2}\right) \tag{6}$$

The damping force Fnd can be calculated by:

$$F_{nd} = -\beta_n * \nu_n \tag{7}$$

where $\beta_n$and ν_n are respectively the normal damping coefficient and the relative normal velocity.

$$\beta_n = \left(\frac{4m^*k_1}{1+(\pi/\ln e)^2}\right)^{1/2} \tag{8}$$

where e is the coefficient of restitution and m^* is the equivalent mass of the two particles withe:

$$m^* = \frac{m_1 m_2}{m_1 + m_2} \tag{9}$$

2.2. EEPA Force: Force in the Shear Direction

The tangential force F_s is composed of two forces namely tangential damping force F_s^d and tangential spring force F_s^{EEPA} .

$$F_s = F_s^{EEPA} + F_s^d \tag{10}$$

The EEPA force vector is incrementally updated :

$$F_s^{EEPA} = F_s^{EEPA,(dt-1)} - k_t \nu_t \Delta t \tag{11}$$

The tangential stiffness is given by :

$$k_t = \begin{cases} k_1\ if & n = 1 \\ 8\ G^*\sqrt{R^*\delta_n^m}\ if & n > 1 \end{cases} \tag{12}$$

where G^* is the effective shear modulus,

$$G^* = \left(\frac{2-\nu_1}{G_1} + \frac{2-\nu_2}{G_2}\right)^{-1} \tag{13}$$

and k_{sf} is the tangential stiffness multiplier. The tangential damping force F_s^d is calculated by :

$$F_s^d = -\beta_t \nu_t \tag{14}$$

β_t is given by:

$$\beta_t = \sqrt{\frac{4m^*k_t}{1+(\pi/\ln e)^2}} \tag{15}$$

For the stability of the model, tangential force take a critical value f_{ct} :

$$f_{ct} \leq \mu\left|F_{hys} + k_{adh}\delta_n^m - F_0\right| \quad (16)$$

where μ is the friction parameter.

3. Johnson-Kendall-Roberts (JKR) Contact Model

Johnson, Kendall and Roberts theory (JKR) [64] introduced adhesion between different surfaces, induced by their surface energies. The (JKR) model is often used to model adherence when caused by capillary forces [60,65–67]. This phenomenon of adhesion has the effect of increasing the contact surface, independently of the load, and justifies the need for a negative force for the release of the surfaces. The theory of the JKR model has been detailed in [68].

For the collision detection, the Sweep and Prune algorithm remains the most famous algorithm used in contact detection when applying the discrete element method [69,70]. This method gives each particle a rectangle parallelepiped that contains it exactly, and the boundaries of this box define on each x, y and z axis a moving interval [x_{min} , x_{max}], [y_{min} , y_{max}] and [z_{min} , z_{max}] respectively (figure 4). The collision of the particles occurred if there is an intersection between the intervals on one of the three axes.

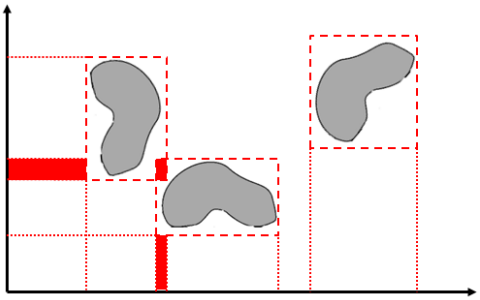


Figure 4: Principle of the Sweep and Prune algorithm

4. Calibration methodology and parameter analysis

4.1. Angle of repose calibration

To generate results close to reality based on simulations of real industrial processes, the inputs relating to the granular materials must be as accurate as possible. The difficulties encountered in the use of DEM lie in the experimental determination of the grain properties (shapes, sizes, coefficient of restitution, rolling friction, grain-grain friction and grain-wall friction, density of gains). Consequently, it is necessary to calibrate the input parameters to achieve satisfactory results [31]. The calibration procedure consists of determining the value of an input parameter by direct measurement such as AoR, bulk density and particle size distribution [71], we then proceed with a series of adjustments until the coincidence between the measured values and the simulated values. A necessary condition is that the tests and measurement method must be reproducible in the laboratory so that the coefficients of interaction can be easily determined.

The angle of repose During a flow, a material forms a natural pile whose slope, defined with the horizontal, forms the natural angle of repose (AoR). It is used to describe the mechanical properties of granular materials. . The natural angle of repose is constant for a given material, but varies with its moisture content [72]. It is used when studying collapsing, stratifiing and segregating. The AoR does not depend only on the mechanical properties of the particles, like density, friction static coefficient and friction dynamic coefficient, but also on the intrinsic properties like shape, geometry and the size of the particles as well as the bulk forming methods [72–75]. When the input values are set, the identification of interaction parameters such as static friction, rolling friction and restitution coefficients must be carried out. These parameters have the ability to considerably affect the nature of the flow. Thus they are very often used as calibration parameters [22, 31,76, 77].

The cylinder lifting (Figure 5) is a common experiment for obtaining the AoR of granular materials [78,79]. The device is basically similar to the one commonly used in litterature, whose principle consists essentially in translating a cylinder suspended upwards [79,80].

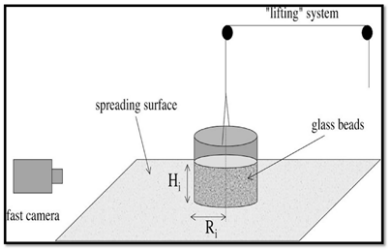


Figure 5: Controlled hollow cylinder test [78]

In [29,73], the author investigated the influence of the dimensions of the test rig and the particle size distribution on the calibration test based on the angle of rest test using a lifting cylinder and they showed that the results remain invariant . The translation speed of the cylinder is considered as a parameter influencing the AoR. Thus the experiment has provided invariant results for granular materials subjected to static stresses when lifting speeds are low.

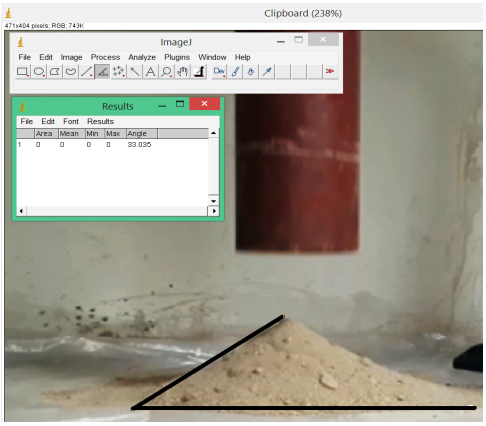


Figure 6: Phosphate AoR test

For our test, 15 kg of mixed phosphate layers (SA2, C1, C2, C3, C4, C5 and C6) is used to carry out the Talus angle test. Inside a cylinder with an internal diameter of 200 mm and a height of 30 cm, a quantity of the Phosphate's particules are glued in such a way that the flow does not interact with the cylinder material. Similarly, on the support on which the phosphate mass will stand, a quantity of phosphate grains is glued to avoid interaction between the particles and the support (Figure 6). Consequently, the interaction

of the particles with the material of the cylinder and the support will be neglected. The cylinder was lifted at a speed of 0.05 m/s [29].

The AoR test was designed to investigate the influence of grain-grain interaction parameters on output variables (angle of repose in our case). We repeated this physical test several times, to have an average of the AoR responses. This average will constitute the target value to be reached during the calibration of the interaction parameters in the EDEM® software. This angle is calculated by measuring the base diameter of the phosphate cone and its height (Figure 7,8).

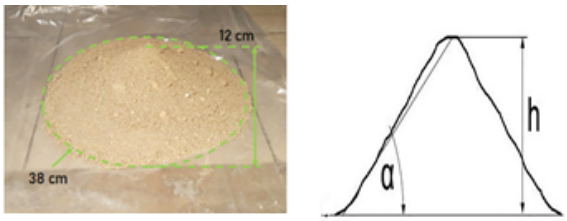


Figure 7: Angle of repose of Phosphate

$$\alpha = tan^{-1}\left(\frac{2h}{d}\right) = tan^{-1}\left(\frac{2\times12}{38}\right) = 33^{\circ} \quad (19)$$

h, d are respectively the height and the diameter of the phosphate cone. Thus the characteristic angle of repose of the phosphate is 33°. For the verification we used the software ImageJ [81] and the results are shown in Figure 6. This value will therefore be the calibration target in the following.

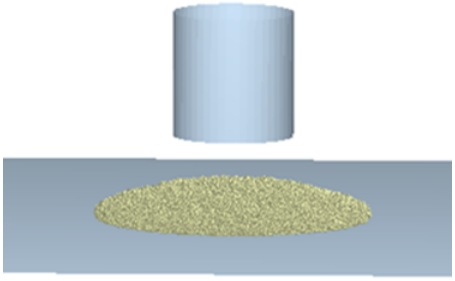


Figure 8: Simulation of the AoR test with EDEM® software

4.2 Particle size distribution and shape

4.2.1 Particle size and distribution

The dimensional distribution of the grains in percentage is determined using dry particle size analysis of samples of phosphate of mass M=1.5 Kg (10 samples). This test is the oldest, most common and most economical test described in the NP 18-560 standard. It consists in classifying the different particles constituting the sample by a series of nylon sieves nested one inside the other, with decreasing openings from top to bottom. The phosphate, dried for 24 hours in an oven at T=120°C to avoid agglomeration of the grains due to the presence of water, is placed at the top and the classification is then obtained by vibrating all the sieves for 5 min. At the end of the test, the rejection from each sieve is weighed (Table1) and the percentage of each size is presented in Figure 9. This study showed that most of the overall mass is concentrated in the 150 µm to 500 µm range.

Table 1: Results of the particle size analysis of phosphate

Sieve size µm	Retained (g)	Retai-ned (%)	Retained cumulat-ive (g)	Retained cumulative (%)	Cumulati-ve sieve (%)
2000	49.8	3.32	49.8	3.32	96.68
800	172.05	11.47	221.85	14.79	85.21
500	399.45	26.63	621.3	41.42	58.58
250	399	26.6	1020.3	68.02	31.98
150	280.5	18.71	1300.95	86.73	13.27
80	150.3	10.02	1451.25	96.75	3.25
< 80	45.3	3.02	1496.55	99.77	0.23

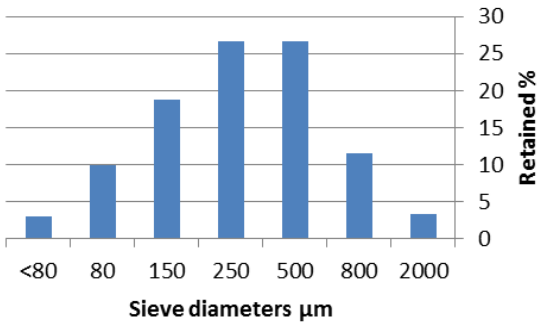


Figure 9: Grain size distribution of phosphate

For the simulation we are asked to enter the particle size distribution of the granular material. For this purpose a statistical test of normality will be carried out to check whether the data in Table 1 follow a normal distribution. The distribution is symmetrical, bell-shaped. It would seem that it can be modelled by a normal distribution. To confirm this, we construct Henry's line. Henry's line is used to approximate the average and standard deviation of the observations of a variable when only the frequencies of its values are available (Figure 10). The distribution can therefore be considered Gaussian. The estimated population average is 390.8µm and the standard deviation is $\sigma = 476.19\mu m$. The software results show that the coefficient of determination R^2 = 0.907 and the corrected coefficient of determination R^2_{corr}= 0.91. These two values indicate that our model is very reliable and can be applied to describe the influence of factors on the response.

4.2.2 Particule shape

The modelling of the real shape is important in a discrete element simulation, and the more the geometrical characteristics of the real particles are related to the numerical particles, the closer the results of the numerical simulation are to the real behaviour of the material [57,59,82]. In the ideal case the grain geometry is retrieved using a rotary scanner [58] which allows to obtain the contours of a particle with precision. In this work, and in order to carry out a three-dimensional modelling of the geometry of an irregular Phosphate particle, samples of each layer of the Benguerir Phosphate are taken to view the structure of their grains under an electronic microscope (Figure 11). Several particle images were taken to select the most representative in terms of size and angularity in order to produce an STL file of the outer shape of a Phosphate particle using CATIA V5 software and which has a geometry roughly gathering almost all the details of the real particles.

The EDEM® software allows users to specify the particle geometry through the so-called "multi-sphere" model, which selects spheres of different sizes to fill the inner volume of the numerical particle geometry, and with the non-linear least squares optimisation the solutions are refined to to reach the minimum error value for mass and shape. When the STL file is entered into the software, different types of particles can be generated in such a way that the shape and size of the particle can be modified by changing the number of bonded spheres and their sizes. However, the calculation time increases exponentially with the number of bonded spheres.

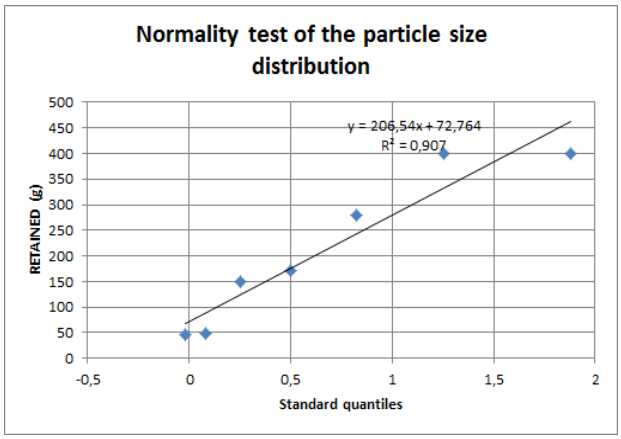


Figure 10: Henry line of particle size distribution

In order to minimise time of simulation [83], consider that combining a number of spheres greater than eight can give satisfactory simulation results (Figure 12).

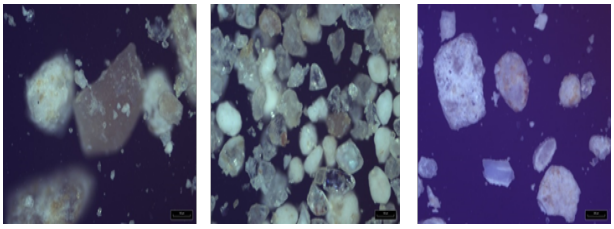


Figure 11: Phosphate grain of sample C1, C3, C6 taken by electronic microscope

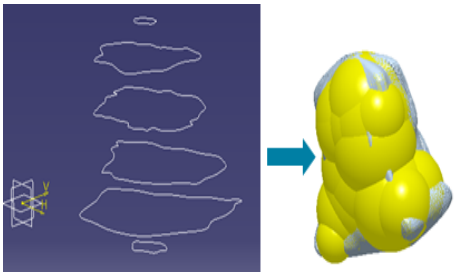


Figure 12: Particle constructed using the 'multi-sphere' method from STL files

5. Test results according to the fractional factorial design

For our simulation process, the DEM was applied, and the EDEM® software was chosen in this research [84]. The EEPA contact model was adopted , with a value of 9.81 m/s^2 for the gravitational acceleration. The inputs parameteres applied in our simulation are presented in Table 2,3,4.

Table 2: Values of some simulation parameters

	Young's modulus E (Pa)	Shear modulus G (Pa)	Poisson Coef. (ν)	Average density $\rho\ (Kg/m^3)$
Phosphate	10^6	2.5 10^6	0.25	2404

Table 3: Results of the shear cell test

Reference layer	C (KPa)	φ (°)
C3	30	33

Table 4: Results of the odometer test

Layer reference	shear index (Cc)	Swell index (Cs)
C6	0.193	0.010

The developers of the EDEM® software recommend using a experimental plan to adjust the properties of the granular material. Thus, a fractional factorial design [85,86] with two levels of resolution and five factors will be used to study the five inter-granular interaction factors in sixteen experiments (Figure 13). For this purpose, an initial range of these interaction parameters was determined on the basis of the literature [5,87,88] and on the basis of the GEMM integrated material data library in the EDEM® software.

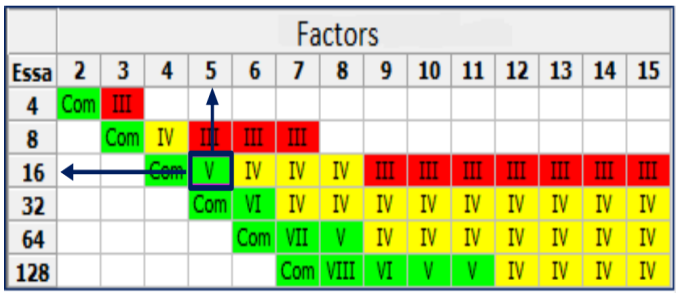


Figure 13: Choice of experimental plan

Table 5: Model parameter ranges

Parameter	Identifier	Low level	High level
Coefficient of restitution	A	0.15	0.75
Static friction coefficient	B	0.2	1
Rolling resistance coefficient	C	0	0.3
Surface energy (J/m^2)	D	0.3	1.5
Pull-off constant	E	-0.009	0

This library helps the software user through a set of questions (field of application, density, experimental angle of repose) to set likely values of the parameters to model the physical phenomenon (Table 5). The fractional factorial design will be used first to analyse the relationships between the parameters potentially influencing the AoR. Once the influencing factors have been

determined, a response surface design analysis will be carried out to accurately determine the parameter values of the model.

The results of the test according to the fractional factorial design 2^{5-1} are presented in Figure 14. It can be seen that the angle of repose has a large range of variation with a maximum and minimum value of 47.92° and 14.3° respectively, which suggests the influence of certain simulation parameters on the angle of repose.

We conducted a more in-depth analysis of the model results reported on the graph (Figure 15). The significance of the factors can be seen intuitively and is in the following decreasing order: coefficient of static friction > surface energy > coefficient of rolling resistance > coefficient of restitution > pull-off constant.

Based on Figure 16 we state that the four parameters: coefficient of rolling resistance, surface energy and static friction coefficient of the particles have a great influence on the AoR of the phosphate mass, Unlike the other parameters, with an order of influence as follows: coefficient of static friction > surface energy > coefficient of rolling resistance.

Coef. of restitution	Static friction coefficient	Rolling resistance coefficient	Surface energy	Pull-off constant	Angle of Repose
0,15	0,2	0,0	0,3	0,000	14,30
0,75	0,2	0,0	0,3	-0,009	14,30
0,15	1,0	0,0	0,3	-0,009	33,00
0,75	1,0	0,0	0,3	0,000	33,00
0,15	0,2	0,3	0,3	-0,009	20,14
0,75	0,2	0,3	0,3	0,000	20,99
0,15	1,0	0,3	0,3	0,000	36,01
0,75	1,0	0,3	0,3	-0,009	36,01
0,15	0,2	0,0	1,5	-0,009	34,96
0,75	0,2	0,0	1,5	0,000	23,86
0,15	1,0	0,0	1,5	0,000	40,25
0,75	1,0	0,0	1,5	-0,009	40,25
0,15	0,2	0,3	1,5	0,000	27,96
0,75	0,2	0,3	1,5	-0,009	27,96
0,15	1,0	0,3	1,5	-0,009	44,60
0,75	1,0	0,3	1,5	0,000	47,92

Figure 14: Simulation results following the fractional factorial design

As the parameters contact coefficient of restitution and pull-off constant have no significant effect on the response of the angle of repose of the phosphate mass, their values will be set at the low levels shown in Table 6.

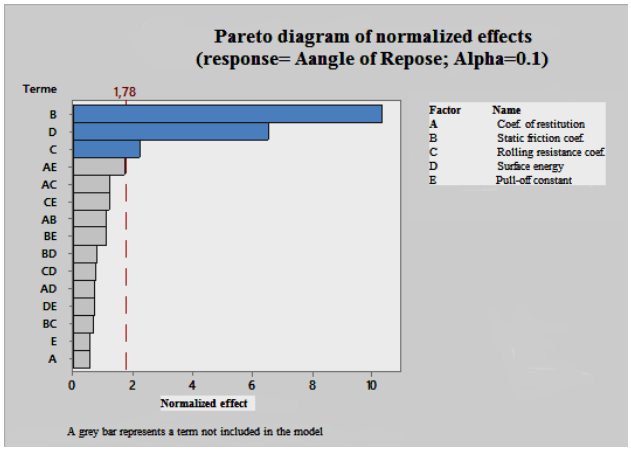


Figure 15: Significance of study factors

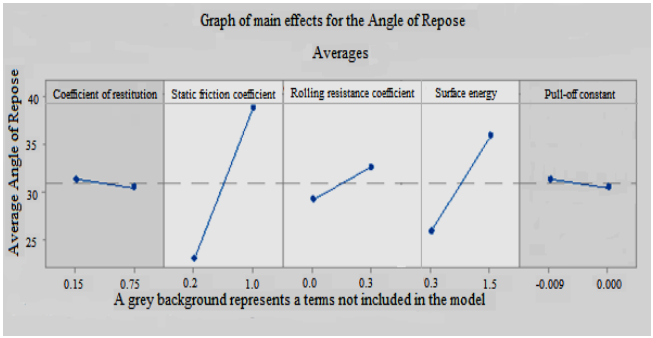


Figure 16: Factors influencing the Angle of Repose

Table 6: Values of non-significant parameters

Non-significant parameter	Value
Coefficient of restitution	0.15
Pull-off constant	-0.009

6. Results of the simulation according to Box Behnken's planTest [89–91]

After the determination of the significative factors for the angle of repose through the fractional factorial design it is necessary to find the exact values of these parameters retained to arrive at a correct modelling of the behaviour of the phosphate. For this purpose the Box Behnken design of experiments are used to find the quadratic relation that links these parameters, considered continuous this time, to the 'angle of repose' response. A three-factor response surface design is chosen and will be studied in 15 trials as recommended by the Minitab software (Figure 17).

Plan		Continuous factors								
		2	3	4	5	6	7	8	9	10
Centred composite (complete)	not divided into blocks	13	20	31	52	90	152			
	divided into blocks	14	20	30	54	90	160			
Centred composite (1/2)	not divided into blocks				32	53	88	154		
	divided into blocks				33	54	90	160		
Centred composite (1/4)	not divided into blocks							90	156	
	divided into blocks							90	160	
Centred composite (1/8)	not divided into blocks									158
	divided into blocks									160
Box-Behnken	not divided into blocks		15	27	46	54	62		130	170
	divided into blocks			27	46	54	62		130	170

Figure 17: Choice of plan of experiments

From an analysis of the responses of simulation the angle of repose in Figure 14, it can be seen that the (AoR) approaches the experimental value for ranges of coefficient of friction, coefficient of rolling resistance and surface energy in the intervals [0.4; 1], [0; 0.4] and [0.8; 1.2] respectively. Thus these intervals will form the basis of the study using Box Behnken plan. The results of the numerical simulation following the test matrix are shown in Figure 18 , and it can be seen that the responses are close to the results of the previous fractional plan. In addition, a quadratic regression model is established by the software MiniTab (Equation 20) which has a coefficient of determination of 98.58% close to 1 (i.e.) it has a high reliability.

$$AoR = 2.89 + 50.81\,B + 15.66\,C + 6.5\,D - 23.25\,B^2 - 20.43\,C^2 \quad (20)$$

Static friction coef.	Rolling resistance coef.	Surface energy	AoR
-1	-1	0	26,36
1	-1	0	37,16
-1	1	0	29,66
1	1	0	39,20
-1	0	-1	26,52
1	0	-1	38,77
-1	0	1	29,05
1	0	1	40,29
0	-1	-1	31,26
0	1	-1	35,43
0	-1	1	35,28
0	1	1	37,76
0	0	0	36,09
0	0	0	35,58
0	0	0	36,47

Figure 18: Simulation results following the Box Behnken plan

7. Analysis of the interaction effect of the regression model

In this test, the AoR of the phosphate mass is used as an indicator for evaluating the regression model. Response surfaces and contour diagram of the pairwise interaction of the different parameters that affect the angle of repose are shown in Figures 19 and 20. An analysis of the parabolas that exist on the response surface curve shows the existence of a non-linear dependence of the AoR on the coefficient of static friction and the coefficient of rolling resistance. These quadratic dependencies are illustrated in the model of Equation 20 with the square term of the coefficient of static friction and rolling resistance. As shown in Figures 19 and 20, the angle of repose increases with the increase in surface energy. When this energy increases, it will lead to a stronger adhesion between the particles, which will lead to the creation of new aggregates that will give more stability to the granular material. This will make particle collapse more difficult and therefore the angle of rest will increase. Similarly the AoR increases with the increase of the coefficient of static friction since the friction contradicts the movement of the particles on each other and thus the angle of the phosphate pile increases. The figures 21 and 22 show that while the rolling resistance coefficient of the particles increases the repose angle of the phosphate volume increases. This is because the rolling resistance prevents the particles from rolling down the slope of the granular cone, which contributes to the increase in the angle of repose.

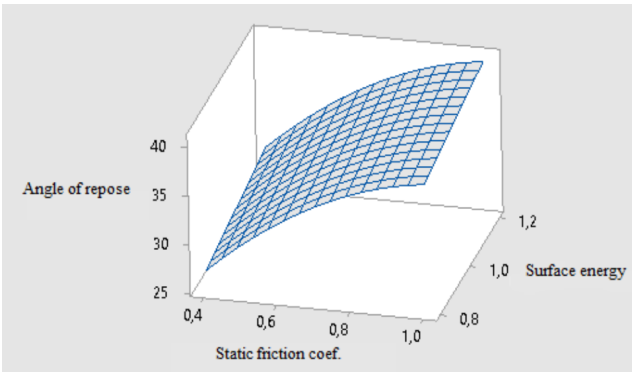


Figure 19: Variation of the angle of repose with surface energy and the static friction coefficient

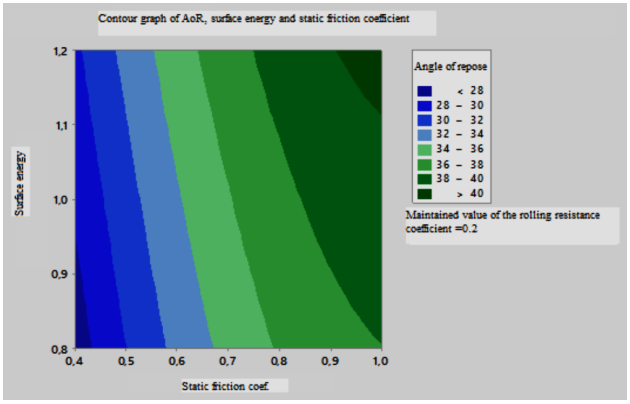


Figure 20 :Contour graph of AoR, surface energy and static friction coefficient

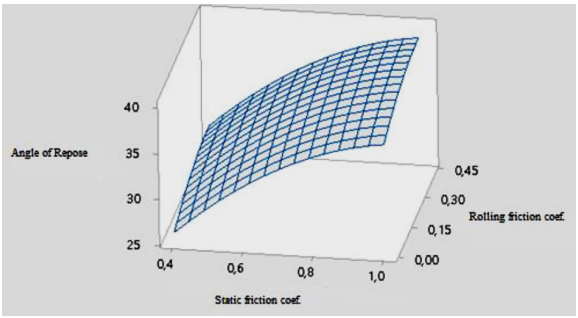


Figure 21 :Variation of the AoR as a function of the surface energy and the rolling resistance coefficient

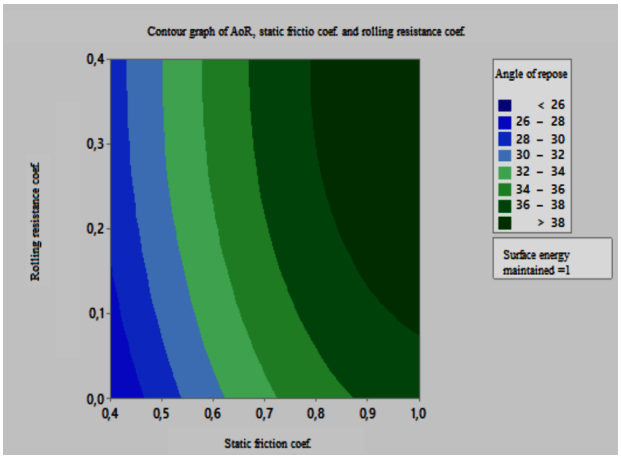


Figure 22: Contour graph of the AoR as a function of the surface energy and the rolling resistance coefficient

8. Results and discussion

The repose angle AoR of the phosphate mass measured by field test was 33°. We taked this real angle as the response value and used the contours above and take as values of significant parameters the following values:

Table 7: Optimised values of the simulation parameters

Parameters	Values
Static friction coef.	0.7
Rolling resistance coef.	0.3
Surface energy	1j/m^2

A simulation under EDEM® was performed under the optimised parameters and the angle of repose for phosphate ore was 34.38°. We compared the test results with the simulation results. The relative error was 4.19%. This error value shows that the optimised parameters can be utilised as a reference for the characteristics of the phosphate.

9. Particle velocities

In this section we will investigate the impact of particle velocity on the contact with the tipper surface during loading and unloading. During the first few moments of loading of the tipper by the loader bucket, the simulation results show that the area most affected by the falling grain is in the front part of the tipper as shown in Figure 23. This configuration is quite natural since the particles leaving the bucket are accelerated by gravity to find this area on the tipper first, and then the grains that follow will be oriented towards the sides once the angle of repose exceeds the AoR of the phosphate.

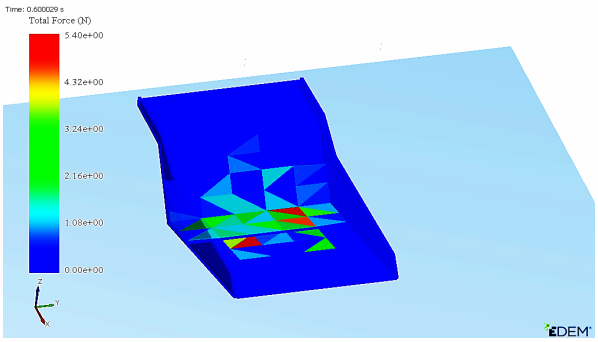


Figure 23: Area of first impact on the tipper

Just before the collision,the particles have a maximum kinetic energy, as they are accelerated by gravity to reach a maximum speed of 3.92 m/s during the first contacts with the tipper (Figure 24). This maximum falling speed starts to decrease as the granular material accumulates in the tipper since the height of fall decreases (figure 25).

Of course, this area of first impact can move on the tipper depending on the position of the loader bucket. Thus, the amplitude of the falling speed will depend on the height of the bucket. Figure 26 shows the decrease in the value of the maximum particle velocity (2.75 m/s) in the case of the reduction of the drop height. However, the position of the filler was fixed according to multiple observations of the truck loading operation in the field or on website videos.

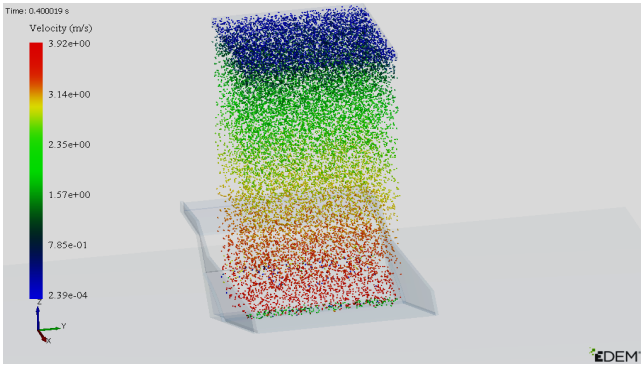


Figure 24: Maximum impact speed

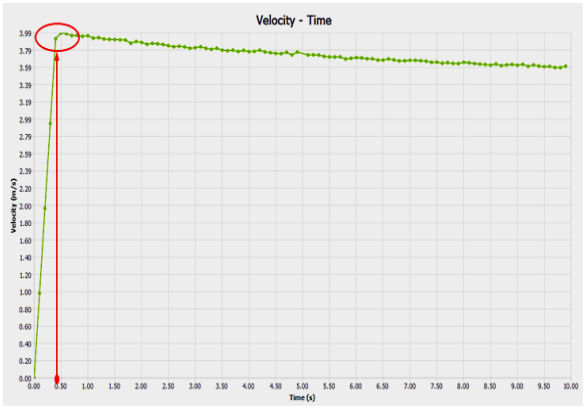


Figure 25: Variation of maximum particle velocities

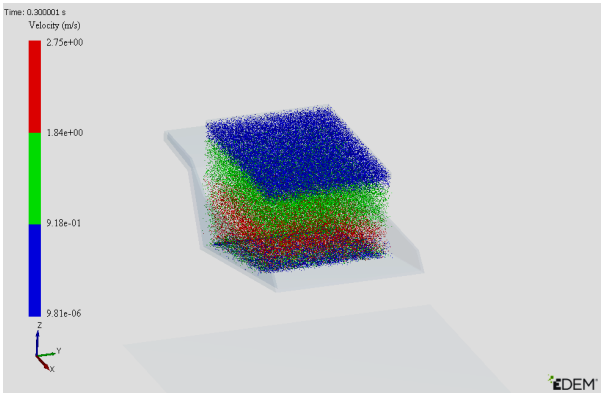


Figure 26: Decrease in particle velocity with decreasing of drop height

This data predicts that the front part of the tipper will be more exposed to abrasive erosion due to the high rates of direct collision of the particles against the tipper during the early loading phase. This is confirmed by the presence of pits and cavities in the same area on the tippers . These pits will increase the flow angle and consequently trap the phosphate grains and give rise to the nucleation of the clogging phenomenon.

10. Conclusion

In this work, we studied phosphate ore using the discrete element method. The study is devoted to the calibration of the DEM based on the angle of repose. The simulation parameters are calibrated and optimized by combining the results from the various physical tests and the DEM simulation with the EDEM® software. We took the value of the angle of repose as the target for the simulation response values, and using the Minitab software, we were able to analyze and optimize the simulation parameters using the response surface method. These physical parameters that significantly affect the AoRt were selected, and the interaction between them was analyzed to determine optimal values. We also applied the fractional factorial design test, and the results are analysed to determine the influence of each of the factors and it was found that the angle of repose depends only on the coefficient of static friction, the surface energy and the rolling resistance coefficient. In addition we used Box-Behnken response surface methode. The results were analysed and we established the regression model of the significant factors on the angle of repose. This model predicts an increase in the angle of repose as the significant factors increase. The results show that when the surface energy, static friction coefficient and rolling resistance coefficient

take the values: 1 J/m^2, 0.7 and 0.3 respectively, the simulation results are consistent with the experimental test result and the error in the angle of repose obtained is 4.19%.

A study was conducted on the impact of particle velocity on the surface of the tipper during loading and unloading. This study allowed us to predict that the front part of the tipper will be more exposed to abrasive erosion due to the high speed of direct collision of the particles against the tipper during the beginning of the loading phase, which corresponds perfectly to the reality in the field.

This work constitutes the first part for the study of phosphate or clogging. Indeed, the identification of the different parameters that define phosphate by the calibration method will allow us to better model the phenomenon with the DEM method. The study of the impact velocities will be used to model the stresses and strains on the tipper surface during the loading and unloading of the phosphate in our next work.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] FAO, Marché mondial des engrais: bilan du resserrement actuel du marché ., ROME, 2022.

[2] FAO, World fertilizer trends and outlook to 2020, 2017.

[3] N. Alexandratos, J.B. Fao, "WORLD AGRICULTURE TOWARDS 2030 / 2050 The 2012 Revision ESA Working Paper No . 12-03."

[4] P. Christmanna, "A forward look into rare earth supply and demand: A role for sedimentary phosphate deposits?," Procedia Engineering, **83**, 19–26, 2014, doi:10.1016/j.proeng.2014.09.005.

[5] M. Medkour, Y. Najih, J. Bengourram, A. El Ghmari, M. Bachaoui, H. Latrache, M. Mabrouki, "Mechanical and thermal characterization of natural phosphate: Clogging understanding (Part 1)," Materials Today: Proceedings, **51**, 2032–2039, 2022, doi:10.1016/j.matpr.2021.06.261.

[6] I. Louw, "Potential radiological impact of the phosphate industry in South Africa on the public and the environment (Paper 1)," Journal of Environmental Radioactivity, **217**(February), 106214, 2020, doi:10.1016/j.jenvrad.2020.106214.

[7] J. Duarte, A.T. Marques, Occupational Accidents in the Mining Industry — A Short Review, Springer International Publishing: 61–69, 2019, doi:10.1007/978-3-030-14730-3.

[8] J. Shen, A. Roberts, C. Wheeler, "Dem simulations on gate loads and bin storage characteristics before discharge," Powder Technology, **383**, 280–291, 2021, doi:10.1016/j.powtec.2021.01.028.

[9] M.J. Mohajeri, W. De Kluijver, R.L.J. Helmons, C. Van Rhee, D.L. Schott, "A validated co-simulation of grab and moist iron ore cargo : Replicating the cohesive and stress-history dependent behaviour of bulk solids," Advanced Powder Technology, **32**(4), 1157–1169, 2021, doi:10.1016/j.apt.2021.02.017.

[10] J. Sun, H. Chen, J. Duan, Z. Liu, Q. Zhu, "Mechanical properties of the grooved-wheel drilling particles under multivariate interaction influenced based on 3D printing and EDEM simulation," Computers and Electronics in Agriculture, **172**(July 2019), 105329, 2020, doi:10.1016/j.compag.2020.105329.

[11] M. Haustein, A. Gladkyy, R. Schwarze, "Discrete element modeling of deformable particles in YADE," SoftwareX, **6**, 118–123, 2017, doi:10.1016/j.softx.2017.05.001.

[12] C. Ramírez-Aragón, J. Ordieres-Meré, F. Alba-Elías, A. González-Marcos, "Comparison of Cohesive models in EDEM and LIGGGHTS for simulating powder compaction," Materials, **11**(11), 1–17, 2018, doi:10.3390/ma11112341.

[13] B.C. Burman, P.A. Cundall, O.D.L. Strack, "A discrete numerical model for granular assemblies," Geotechnique, **30**(3), 331–336, 1980, doi:10.1680/geot.1980.30.3.331.

[14] J. Jahnke, M. Burger, "PARAMETER IDENTIFICATION FOR SOIL SIMULATION BASED ON THE DISCRETE ELEMENT METHOD AND APPLICATION TO SMALL SCALE SHALLOW PENETRATION TESTS," 332–342, 2019.

[15] Y. Zhu, J. Gong, Z. Nie, "Particuology Shear behaviours of cohesionless mixed soils using the DEM : The influence of coarse particle shape," Particuology, **55**, 151–165, 2021, doi:10.1016/j.partic.2020.07.002.

[16] F.-V.D. Luc Scholtès, "A DEM model for soft and hard rocks : Role of grain interlocking on strength," Journal of the Mechanics and Physics of Solids, **61**, 352–369, 2013, doi:10.1016/j.jmps.2012.10.005.

[17] D. Yan, J. Yu, Y. Wang, L. Zhou, K. Sun, Y. Tian, "A Review of the Application of Discrete Element Method in Agricultural Engineering : A Case Study of Soybean," 1–18, 2022.

[18] L.K. Nordell, "Particle flow modeling: transfer chutes & other applications," International Materials Handling Conference (BELTCON 9), Johannesburg, South Africa, 1–16, 1997.

[19] D.B.B. Hastie, a. P.P. Grima, P.W.W. Wypych, "Validation of particle flow through a conveyor transfer spoon via particle velocity analysis," Bulk Europe 2008, (March 2015), 1–5, 2008.

[20] T. Gröger, A. Katterfeld, "On the numerical calibration of discrete element models for the simulation of bulk solids," Computer Aided Chemical Engineering, **21**(C), 533–538, 2006, doi:10.1016/S1570-7946(06)80100-8.

[21] M.D. Sinnott, P.W. Cleary, "The effect of particle shape on mixing in a high shear mixer," Computational Particle Mechanics, **3**(4), 477–504, 2016, doi:10.1007/s40571-015-0065-4.

[22] S. Zhang, M.Z. Tekeste, Y. Li, A. Gaul, D. Zhu, J. Liao, "Scaled-up rice grain modelling for DEM calibration and the validation of hopper flow," Biosystems Engineering, **194**, 196–212, 2020, doi:10.1016/j.biosystemseng.2020.03.018.

[23] B. Soltanbeigi, A. Podlozhnyuk, C. Kloss, S. Pirker, J.Y. Ooi, S.A. Papanicolopulos, "Influence of various DEM shape representation methods on packing and shearing of granular assemblies," Granular Matter, **23**(2), 1–16, 2021, doi:10.1007/s10035-020-01078-y.

[24] S. Huang, J.X. Huang, Y.S. Mo, Z.F. Zhang, "Study on Wear Properties of the Flow Parts in a Centrifugal Pump based on EDEM-Fluent Coupling," Reneng Dongli Gongcheng/Journal of Engineering for Thermal Energy and Power, **35**(1), 62–69, 2020, doi:10.16146/j.cnki.rndlgc.2020.01.010.

[25] B.M. Ghodki, M. Patel, R. Namdeo, G. Carpenter, "Calibration of discrete element model parameters: soybeans," Computational Particle Mechanics, **6**(1), 3–10, 2019, doi:10.1007/s40571-018-0194-7.

[26] Y. Tan, Y. Yu, J. Fottner, S. Kessler, "Automated measurement of the numerical angle of repose (aMAoR) of biomass particles in EDEM with a novel algorithm," Powder Technology, **388**, 462–473, 2021, doi:10.1016/j.powtec.2021.04.062.

[27] M. Mousaviraad, M. Tekeste, K.A. Rosentrater, "CALIBRATION AND VALIDATION OF A DISCRETE ELEMENT MODEL OF CORN USING GRAIN FLOW SIMULATION IN A COMMERCIAL SCREW GRAIN AUGER," **60**(4), 1403–1415, 2017.

[28] M.J. Mohajeri, H.Q. Do, D.L. Schott, "DEM calibration of cohesive material in the ring shear test by applying a genetic algorithm framework," Advanced Powder Technology, **31**(5), 1838–1850, 2020, doi:10.1016/j.apt.2020.02.019.

[29] T. Roessler, A. Katterfeld, "DEM parameter calibration of cohesive bulk materials using a simple angle of repose test," Particuology, **45**, 105–115, 2019, doi:10.1016/j.partic.2018.08.005.

[30] A. Boikov, R. Savelev, V. Payor, A. Potapov, "Universal Approach for DEM Parameters Calibration of Bulk Materials," 2021.

[31] C.J. Coetzee, "Review: Calibration of the discrete element method," Powder Technology, **310**, 104–142, 2017, doi:10.1016/j.powtec.2017.01.015.

[32] N. Kassotakis, V. Sarhosis, M. V Peppa, J. Mills, "Quantifying the effect of geometric uncertainty on the structural behaviour of arches developed from direct measurement and Structure-from-Motion (SfM) photogrammetry," Engineering Structures, **230**(December 2020), 111710, 2021, doi:10.1016/j.engstruct.2020.111710.

[33] M. Marigo, E.H. Stitt, "Discrete element method (DEM) for industrial applications: Comments on calibration and validation for the modelling of cylindrical pellets," KONA Powder and Particle Journal, **32**(32), 236–252, 2015, doi:10.14356/kona.2015016.

[34] S. Kirsch, "Avoiding ambiguity in DEM in-situ calibration for dry bulk materials," Minerals Engineering, **145**(July 2019), 106094, 2020, doi:10.1016/j.mineng.2019.106094.

[35] J. Quist, M. Evertsson, "Framework for DEM Model Calibration and Validation," Proceedings of the 14th European Symposium on Comminution and Classification, (September), 103–108, 2015.

[36] T.A.H. Simons, R. Weiler, S. Strege, S. Bensmann, M. Schilling, A. Kwade, "A ring shear tester as calibration experiment for DEM simulations in agitated mixers - A sensitivity study," Procedia Engineering, **102**, 741–748, 2015, doi:10.1016/j.proeng.2015.01.178.

[37] G. Ma, Z. Sun, H. Ma, P. Li, "Calibration of Contact Parameters for Moist Bulk of Shotcrete Based on EDEM," Advances in Materials Science and Engineering, **2022**, 2022, doi:10.1155/2022/6072303.
[38] G.K.P. Barrios, R.M. de Carvalho, A. Kwade, L.M. Tavares, "Contact parameter estimation for DEM simulation of iron ore pellet handling," Powder Technology, **248**, 84–93, 2013, doi:10.1016/j.powtec.2013.01.063.
[39] C. Chen, J. Gu, Z. Peng, X. Dai, Q. Liu, G.Q. Zhu, "Discrete element modeling of particles sphericity effect on sand direct shear performance," Scientific Reports, **12**(1), 1–14, 2022, doi:10.1038/s41598-022-09543-9.
[40] Z. Zhao, G. Jiang, R. Mao, "Effects of particle sizes of rock phosphate on immobilizing heavy metals in lead zinc mine soils," **14**(2), 258–266, 2014.
[41] Z. Wang, Y. Fang, W. Feng, X. Tian, J. Lin, "Comparative Study on Particle Size Effect of Crushable Granular Soils through DEM Simulations," **2022**, 2022, doi:10.2113/2022/1608454/5558342/1608454.pdf.
[42] S. Ben Turkia, D.N. Wilke, P. Pizette, N. Govender, N.E. Abriak, "Benefits of virtual calibration for discrete element parameter estimation from bulk experiments," Granular Matter, 1–16, 2019, doi:10.1007/s10035-019-0962-y.
[43] E.X.Z.. P.Z.. Z.H. Jiang, "Discrete Element Method of Coke Accumulation : Calibration of the Contact Parameter," IFAC-PapersOnLine, 241–245, 2018, doi:10.1016/j.ifacol.2018.09.425.
[44] . A.V.B.. R.V.S.A.P., "DEM Calibration Approach : Implementing Contact Model," Journal of Physics: Conf. Series, 2018.
[45] Z. Wu, X. Wang, D. Liu, F. Xie, L. George, Z. Zhang, Q. Tang, "Calibration of discrete element parameters and experimental verification for modelling subsurface soils," Biosystems Engineering, **212**, 215–227, 2021, doi:10.1016/j.biosystemseng.2021.10.012.
[46] Y. Su, Y. Xu, T. Cui, X. Gao, G. Xia, Y. Li, "Determination and interpretation of bonded- particle model parameters for simulation of maize kernels," Biosystems Engineering, **210**, 193–205, 2021, doi:10.1016/j.biosystemseng.2021.08.022.
[47] Z. Zhao, H. Li, J. Liu, S.X. Yang, "Control method of seedbed compactness based on fragment soil compaction dynamic characteristics," Soil and Tillage Research, **198**(January), 104551, 2020, doi:10.1016/j.still.2019.104551.
[48] L. Zhang, C.Q. Ru, "A refined JKR model for adhesion of a rigid sphere on a soft elastic substrate," Journal of Applied Mechanics, Transactions ASME, **86**(5), 1–11, 2019, doi:10.1115/1.4042574.
[49] J. Hellwig, R.M.P. Karlsson, L. Wågberg, T. Pettersson, "Measuring elasticity of wet cellulose beads with an AFM colloidal probe using a linearized DMT model," Analytical Methods, **9**(27), 4019–4022, 2017, doi:10.1039/c7ay01219e.
[50] A. Papangelo, R. Lovino, M. Ciavarella, "Electroadhesive sphere-flat contact problem: A comparison between DMT and full iterative finite element solutions," Tribology International, **152**(May), 106542, 2020, doi:10.1016/j.triboint.2020.106542.
[51] S. Richesson, M. Sahimi, "Hertz-Mindlin Theory of Contacting Grains and the Effective-Medium Approximation for the Permeability of Deforming Porous Media," Geophysical Research Letters, **46**(14), 8039–8045, 2019, doi:10.1029/2019GL083727.
[52] M.J. Mohajeri, R.L.J. Helmons, C. van Rhee, D.L. Schott, "A hybrid particle-geometric scaling approach for elasto-plastic adhesive DEM contact models," Powder Technology, **369**, 72–87, 2020, doi:10.1016/j.powtec.2020.05.012.
[53] S.C. Thakur, J.P. Morrissey, J. Sun, J.F. Chen, J.Y. Ooi, "Micromechanical analysis of cohesive granular materials using the discrete element method with an adhesive elasto-plastic contact model," Granular Matter, **16**(3), 383–400, 2014, doi:10.1007/s10035-014-0506-4.
[54] A. Hafez, Q. Liu, T. Finkbeiner, R.A. Alouhali, T.E. Moellendick, J.C. Santamarina, "The effect of particle shape on discharge and clogging," Scientific Reports, **11**(1), 1–12, 2021, doi:10.1038/s41598-021-82744-w.
[55] P. Vangla, N. Roy, M.L. Gali, "Image based shape characterization of granular materials and its effect on kinematics of particle motion," Granular Matter, **20**(1), 2018, doi:10.1007/s10035-017-0776-8.
[56] S.S. Ahmed, A. Martinez, "Effects of Particle Shape on the Shear Wave Velocity and Shear Modulus of 3D Printed Sand Analogs," Open Geomechanics, **3**(1), 1–18, 2022, doi:10.5802/ogeo.9.
[57] D. Höhner, S. Wirtz, V. Scherer, "A numerical study on the influence of particle shape on hopper discharge within the polyhedral and multi-sphere discrete element method," Powder Technology, **226**, 16–28, 2012, doi:10.1016/j.powtec.2012.03.041.
[58] Y. Chen, Z. Chu, X. Yu, "Research of The Influence of The Particle Geometry on The Accumulation of Repose Angle Based on DEM," **154**(Icmia), 38–46, 2017, doi:10.2991/icmia-17.2017.8.
[59] A. Khazeni, Z. Mansourpour, "Influence of non-spherical shape approximation on DEM simulation accuracy by multi-sphere method," Powder Technology, **332**, 265–278, 2018, doi:10.1016/j.powtec.2018.03.030.
[60] J.P. Morrissey, "Discrete Element Modelling of Iron Ore Fines to Include the Effects of Moisture and Fines," 1–395, 2013.
[61] O.R. Walton, R.L. Braun, "Viscosity, granular-temperature, and stress calculations for shearing assemblies of inelastic, frictional disks," Journal of Rheology, **30**(5), 949–980, 1986, doi:10.1122/1.549893.
[62] M.J. BOUTELOUP, Simulation numérique de la dynamique d'un lit granulaire cisaillé par un fluide visqueux, Institut National Polytechnique de Toulouse (INP Toulouse), 2017.
[63] C.J. Coetzee, Edinburgh-Elasto-Plastic-Adhesion (EEPA) Contact Model Implementation in PFC, Version 1.1, 2022, doi:10.13140/RG.2.2.13659.18724.
[64] K.L. Johnson, K. Kendall, A.D. Roberts, "Surface energy and the contact of elastic solids," Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences, **324**(1558), 301–313, 1971, doi:10.1098/rspa.1971.0141.
[65] J. Hærvig, U. Kleinhans, C. Wieland, H. Spliethoff, A.L. Jensen, K. Sørensen, T.J. Condra, "On the adhesive JKR contact and rolling models for reduced particle stiffness discrete element simulations," Powder Technology, **319**, 472–482, 2017, doi:10.1016/j.powtec.2017.07.006.
[66] S.A. Carr, J. Liu, A.G. Tesoro, "Transport and fate of microplastic particles in wastewater treatment plants," Water Research, **91**, 174–182, 2016, doi:10.1016/j.watres.2016.01.002.
[67] R. Xia, B. Li, X. Wang, T. Li, Z. Yang, "Measurement and calibration of the discrete element parameters of wet bulk coal," Measurement: Journal of the International Measurement Confederation, **142**, 84–95, 2019, doi:10.1016/j.measurement.2019.04.069.
[68] J. Zhou, L. Zhang, C. Hu, Z. Li, J. Tang, K. Mao, X. Wang, "Calibration of wet sand and gravel particles based on JKR contact model," Powder Technology, **397**, 117005, 2022, doi:10.1016/j.powtec.2021.11.049.
[69] T. Capannini, Gabriele and Larsson, "Adaptive Collision Culling for Massive Simulations by a Parallel and Context-Aware Sweep and Prune Algorithm," IEEE Transactions on Visualization and Computer Graphics, **24**(7), 2064–2077, 2018, doi:10.1109/TVCG.2017.2709313.
[70] J.C. Monan Wang, "A review of collision detection for deformable objects," Computer Animation and Virtual Worlds, **32**(5), 2021, doi:https://doi.org/10.1002/cav.1987.
[71] A. Grima, D. Hastie, D. Curry, P. Wypych, R. La Roche, A. Grima, D. Hastie, L. Roche, "The beginning of a new era in design: Calibrated discrete element modelling Publication Details," Australian Bulk Handling Review, **16**(6), 14–21, 2011.
[72] H.M. Beakawi Al-Hashemi, O.S. Baghabra Al-Amoudi, "A review on the angle of repose of granular materials," Powder Technology, **330**, 397–417, 2018, doi:10.1016/j.powtec.2018.02.003.
[73] T. Roessler, A. Katterfeld, "Scaling of the angle of repose test and its influence on the calibration of DEM parameters using upscaled particles," Powder Technology, **330**, 58–66, 2018, doi:10.1016/j.powtec.2018.01.044.
[74] H. Nakashima, Y. Shioji, T. Kobayashi, S. Aoki, H. Shimizu, J. Miyasaka, K. Ohdoi, "Determining the angle of repose of sand under low-gravity conditions using discrete element method," Journal of Terramechanics, **48**(1), 17–26, 2011, doi:10.1016/j.jterra.2010.09.002.
[75] M. Khanal, M. Elmouttie, D. Adhikary, "Effects of particle shapes to achieve angle of repose and force displacement behaviour on granular assembly," Advanced Powder Technology, **28**(8), 1972–1976, 2017, doi:10.1016/j.apt.2017.04.016.
[76] H. Zhou, Z. Hu, J. Chen, X. Lv, N. Xie, "Calibration of DEM models for irregular particles based on experimental design method and bulk experiments," Powder Technology, **332**, 210–223, 2018, doi:10.1016/j.powtec.2018.03.064.
[77] F. Ye, C. Wheeler, B. Chen, J. Hu, K. Chen, W. Chen, "Calibration and verification of DEM parameters for dynamic particle flow conditions using a backpropagation neural network," Advanced Powder Technology, **30**(2), 292–301, 2019, doi:10.1016/j.apt.2018.11.005.
[78] E. Lajeunesse, A. Mangeney-Castelnau, J.P. Vilotte, "Spreading of a granular mass on a horizontal plane," Physics of Fluids, **16**(7), 2371–2381, 2004, doi:10.1063/1.1736611.
[79] R. Cabiscol, J.H. Finke, A. Kwade, "Calibration and interpretation of DEM parameters for simulations of cylindrical tablets with multi-sphere approach," Powder Technology, **327**, 232–245, 2018, doi:10.1016/j.powtec.2017.12.041.
[80] H.T. Chou, C.F. Lee, Y.C. Chung, S.S. Hsiau, "Discrete element modelling and experimental validation for the falling process of dry granular steps," Powder Technology, **231**, 122–134, 2012,

doi:10.1016/j.powtec.2012.08.001.
[81] W. Rasband, ImageJ https://imagej.net/downloads.
[82] D. Höhner, S. Wirtz, V. Scherer, "A study on the influence of particle shape and shape approximation on particle mechanics in a rotating drum using the discrete element method," Powder Technology, **253**, 256–265, 2014, doi:10.1016/j.powtec.2013.11.023.
[83] G.R. McDowell, O. Falagush, H.S. Yu, "A particle refinement method for simulating DEM of cone penetration testing in granular materials," Geotechnique Letters, **2**(7–9), 141–147, 2012, doi:10.1680/geolett.12.00036.
[84] N. Stoimenov, J. Ruzic, "Analysis of the particle motion during mechanical alloying using EDEM software," IFAC-PapersOnLine, **52**(25), 462–466, 2019, doi:10.1016/j.ifacol.2019.12.583.
[85] X. Li, Y. Du, L. Liu, Y. Zhang, D. Guo, "Parameter calibration of corncob based on DEM," Advanced Powder Technology, **33**(8), 103699, 2022, doi:10.1016/j.apt.2022.103699.
[86] W. Tinsson, Plans d'expérience: constructions et analyses statistiques (Mathématiques et Applications, 67), 2011.
[87] E. El Ouardia, "Étude de la calcination du phosphate clair de youssoufia (Maroc)," Afrique Science: Revue Internationale Des Sciences et Technologie, **4**(2), 199–211, 2010, doi:10.4314/afsci.v4i2.61676.
[88] N. Abdenouri, M. Hasnaoui, H. Mazouz, "Effective moisture diffusivity during the phosphate drying. modeling and experimental study," Materials Today: Proceedings, **51**, 2071–2079, 2022, doi:10.1016/j.matpr.2022.01.346.
[89] D.P. Maurice Pillet, François Louvet, "Plan d'expériences " Definitive Screening Design " , une alternative crédible aux plans composites centrés ?," in Qualita 2017, Bourges, France, 2017.
[90] C.C. Design, I. Variable, R. Surface, J.S. Rao, B. Kumar, "Box-Behnken Design 3D Blade root shape optimization Meta-Model Development," 2018.
[91] E. Daems, Frédéric and Lognay, Georges and Romnee, Jean-Michel and Froidmont, "Application du plan d'expérience de Box-Behnken et de la méthodologie des surfaces de réponses pour optimiser une méthode de quantification de phyto-œstrogènes dans les fourrages," 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 4, 79-91 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

On the Prediction of One-Year Ahead Energy Demand in Turkey using Metaheuristic Algorithms

Basharat Jamil[1,*], Lucía Serrano-Luján[1,2], José Manuel Colmenar[1]

[1]*Departamento de Ciencias de la Computación, Arquitectura de Computadores, Lenguajes y Sistemas Informáticos y Estadística e Investigación Operativa, Universidad Rey Juan Carlos, Móstoles-28933, Madrid, Spain*

[2]*Departamento de Electrónica, Tecnología de Computadoras y Proyectos, Universidad Politécnica de Cartagena, Campus Muralla del Mar. C/Doctor Fleming s/n, 30202 Cartagena, Spain*

ARTICLE INFO

Article history:
Received: 04 April, 2022
Accepted: 05 July, 2022
Online: 16 August, 2022

Keywords:
Energy Demand
Metaheuristics
Prediction
Evolutionary Algorithms
Turkey

ABSTRACT

Estimation of energy demand has important implications for economic and social stability leading to a more secure energy future. One-year-ahead energy demand estimation for Turkey has been proposed in this paper, using the metaheuristics method with GDP, the total population, and the quantities of imports and exports, as inputs variables. The records obtained from historical data were bifurcated into training and test datasets, where the training dataset is used by the algorithm in the process of generating models, while the test dataset was used to evaluate the performance of the algorithm. Here, two particular approaches have been proposed: Grammatical Evolution alone, and an ensemble of Grammatical Evolution with Differential Evolution. Under these four different forms are developed, viz, Grammatical Evolution with a recursive grammar (M1), an ensemble of Grammatical evolution executed on a linear grammar and Differential Evolution (M2), an ensemble of Grammatical evolution executed on a quadratic grammar and Differential Evolution (M3), and, Grammatical Evolution with a recursive grammar and Differential Evolution (M4). Moreover, the present approaches were also compared for estimation accuracy against the previously published DE models. It was substantiated that the M4 proposal exhibited the best performance towards estimation. It is therefore established that the current approach exhibits a better estimation capability (with RMSE of 2.2002), compared to the models previously available in the literature. M4 approach is then employed to predict the future energy demand using the same set of socio-economic inputs and the results demonstrated high prediction accuracy with an RMSE of 2.2278.

1. Introduction

Energy requirements of a country are met through proper planning, execution, and prediction based on the infrastructure and the availability of resources. It is no surprise that energy consumption has been rising aggressively across the globe with the increasing energy issues, while more attention must be focused on reducing environmental pollution, together with keeping up the economic growth [1,2]. Energy planning for the future is of utmost importance as it can have a significant impact on the economy of a country [3]. Broadly, energy demand has been mainly influenced by the price of energy and earnings, distributional effects, literacy of energy use among society, and the demand for energy in industry and transportation. Particularly the aspects that affect the demand for energy in the country are the climate of the region, urbanization rate [4], the sectorial energy consumption [5], use of technology [6], transition to renewable energy [7], and adherence and enforcement of environmental law and governmental policies [8].

The evolution of the energy sector and the uncertainties related to it have been a topic of interest to strategize energy secure future and layout policies for future energy demand. In this view, recently, a lot of focus has been realized on the modeling tools, frameworks, and assessment procedures that can provide more accurate insight into the future energy demand of a country [9]. Various approaches can be found in the literature dealing with the modeling energy demand in total of a country such as Turkey [10], China and India [11], Spain [12], etc. The recent ones involve the use of metaheuristics for estimating and predicting the energy

*Corresponding Author: Basharat Jamil, +34-631608231, basharat.jamil@urjc.es

https://dx.doi.org/10.25046/aj070411

demand for the future. We now look at various approaches presented in the literature for estimating and predicting country-wide energy demand.

In [11], the authors analyzed the performance of time series forecasting methods (using a grey model) to predict China's and India's future energy demand (1990-2016). The reported techniques were deemed to increase the prediction of the future energy demand of the two considered countries. Trend map, error measure, and fit method were used to analyze the accuracy and the mean absolute error of single-linear, hybrid-linear, and non-linear techniques were reported as 1.30-3.08%, 0.80-2.57%, and 2.06-2.19%, respectively. In [12], the authors presented a one-year-ahead estimation of energy demand based on historical data for the years 1980 to 2011 for 14 socio-economic indicators. Modified Harmony Search (HS) and an Extreme Learning Machine (ELM) were used and compared for prediction. Reported results by the proposed approaches made very accurate energy demand predictions with a mean absolute error of 3.21-4.63%. In a similar work [13], the authors analyzed the performance of the robust hybrid approach composed of a Basic Variable Neighbourhood Search (BVNS) algorithm with ELM. BVNS was employed to select the most suitable macroeconomic features from the large group of considered ones, whereas the ELM performed the energy demand predictions on the considered features. The proposed algorithm was reported to have the best Mean absolute error lower than 1.66% and an average Mean absolute error of 3.90%. In [14], the authors proposed a new framework for energy demand estimation by combining an adaptive Genetic Algorithm (GA) and a co-integration analysis for China. Model weights were optimized using Artificial Intelligence techniques (GA, ACO, and hybrid algorithms) together with co-integration analysis. It was reported that the proposed models have significantly better performance (Mean Absolute Percentage Error, of 8.68%). In [15], the authors proposed a hybrid algorithm formed by Bat Algorithm, Gaussian Perturbations, and Simulated Annealing Energy Demand (BAG-SA EDE) for China data. The analysis of the relationship between the energy demand and the input factors was carried out using a stationary test, co-integration test, and Granger causality test. It was reported that the quadratic model performs better prediction (mean absolute percentage error of 0.28%) of future energy demand than the multiple linear regression (mean absolute percentage error of 0.88%). Grammatical Evolution (GE) was proposed by [16] for developing new models using 14 macro-economic parameters for a year-ahead estimation of country-wide energy demand for Spain and France. The Differential Evolution (DE) algorithm was used to optimize each model's parameters. The proposed algorithms exhibited excellent accuracy (best Mean absolute error 1.9, and average Mean absolute error 3.33) for energy prediction. Bees Algorithm technique was suggested in [17] for estimating total energy demand in Iran based on population, GDP, import, and export data. Exponential and linear models were proposed for estimation and were also deployed to predict energy demand for up to the year 2030. It was concluded that the linear Bees Algorithm technique has the best prediction accuracy (relative error of 1.07%). In [18], the authors demonstrated a Mix-encoding Particle Swarm Optimization and Radial Basis Function (MPSO-RBF) based energy demand forecasting for China until the year 2020. The input data (for the years 1980-2009) consists of GDP, population, industry proportion, urbanization rate, and share of coal energy. It was reported that the proposed MPSO-RBF has four nodes of hidden layers and better accuracy in terms of the errors (Mean absolute percentage error of 0.78%) when compared to other Artificial Neural Networks (ANN)[19] based models. In [20], Particle Swarm Optimization and Genetic Algorithm optimal Energy Demand Estimating (PSO-GAEDE) model was proposed to improve the estimation efficiency for future projection in China. Linear, quadratic and exponential forms of models were proposed based on historical data from 1990 to 2009. The proposed PSO-GAEDE algorithm was reported to perform better than other algorithms with a Mean absolute percentage error of 0.54%. In [21], the authors deployed an ANN (with feed-forward multilayer perceptron model, coupled with an error back-propagation technique, FF-BP-ANN) to estimate the energy demand for Korea. Multiple linear regression models, exponential model, and ANN models [22] were analyzed and it was concluded that the ANN model outperforms the multiple linear regression models (linear and exponential) with accuracy in terms of Root Mean Squared Error, RMSE=5.7803. In [23], the authors forecasted the energy demand of China using a hierarchical Bayesian approach. Static and dynamic models were proposed with variable input parameters and forecasts were made for years up to 2030. It was reported that the hierarchical Bayesian approach has better performance (RMSE = 0.025) in model fitting than the fixed effects method (RMSE = 0.030). An analysis of energy structure and carbon emissions for China was performed in [24] using an optimized mixed data sampling model (ADL-MIDAS) model involving quarterly GDP, quarterly added value, and annual energy demand as the input variables. It was reported that compared to previous studies, the prediction error (RMSE) was in general below 0.1%, and the smallest error was reported to be 0.02%, thereby energy demand prediction was significantly improved. The forecast of energy demand for the Hunan province of China was presented in [25] between the years 2012 to 2030. Autoregressive integrated moving average and vector autoregressive models were employed for forecasting, and the resulting uncertainties analyzed using the Monte-Carlo method were reported to be under 15%. In [26], the authors employed a swarm intelligence-based Adaptive Firefly Algorithm (AFA) to improve the energy demand estimation in Turkey based on economic parameters. Linear and quadratic forms of models were proposed for estimation and historical data for Turkey. The proposed models were compared with Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO), each involving linear and quadratic models. It was concluded that the AFA-quadratic model resulted in the best accuracy with an accuracy of 99.24%. The determinants for energy demand in Turkey were discussed in [27]. ACO was used to optimize the energy demand problem. Three scenarios of energy demand growth were discussed and values of energy demand were determined for Turkey for 2006-2025. The quadratic equation optimized by ACO was suggested to be the most accurate with a deviation of 2.83%. Similar works were also reported by [28] and [29] for Turkey. Other authors used Differential Evolution (DE)[10], Artificial Algae Algorithm (AAA)[30], PSO [31], Ridge Regression (RR), and Partial Least Squares Regression (PLSR)[32] for estimating future energy demand in Turkey.

Based on previous research, it is elucidated that the long-term dependence of energy demand on economic growth factors certainly exists [33].

In this paper, a one-year-ahead energy demand prediction for Turkey is proposed involving a combination of evolutionary metaheuristic algorithms. These algorithms are based on Grammatical Evolution (GE) and Differential Evolution (DE), where GE has the capacity to develop flexible model forms based on the information provided in the grammar, while DE optimizes the model's parameters. Utilizing these two algorithms in combination results in better estimation accuracy and the predictions for energy demand can be made with greater reliability. The input variables selected are similar to the previous studies so that the comparison of algorithms is fair and the advantage of the present approach can be highlighted. These variables consist of the historical data of socio-economic parameters: GDP, population, import, and export quantities, which were used as inputs, while Energy demand was the target or output variable.

The novelty of the work is further justified by the comparison of algorithms and the grammars used. In the current scenario, two particular approaches have been proposed: Grammatical Evolution alone, and an ensemble of Grammatical Evolution with Differential evolution. Based on these algorithms, four different proposals were studied:

i. *M1*: Grammatical Evolution with a recursive grammar.
ii. *M2*: Ensemble of Grammatical Evolution with linear grammar and Differential Evolution.
iii. *M3*: Ensemble of Grammatical Evolution with quadratic grammar and Differential Evolution.
iv. *M4*: Grammatical Evolution with a recursive grammar and Differential Evolution.

The performance of these proposals is compared by assessing their estimation accuracy using RMSE as the objective function. The average error, R^2, absolute error, and relative error metrics were calculated to compare the performance of the proposals.

The present algorithm combination is also compared to the previous research works reported in literature involving Differential Evolution to validate their accuracy. Further, the application of these metaheuristics is utilized to predict the future energy demand for Turkey. One year ahead energy demand prediction is made using the combination of the previously described two algorithms (GE and DE). Thus, the approach with the combination of the Grammatical Evolution algorithm provided with a highly recursive grammar unified with the optimization strength of the Differential Evolution algorithm (M4) presents a unique and highly accurate approach to making energy demand predictions.

The rest of the paper is organized as follows. *Section 2* provides the outline of the methodology used consisting of the data obtained, problem definition, and the objective function. *Section 3* discusses the results obtained dealing with the estimation of energy demand and prediction of the year ahead of future energy demand for Turkey. Finally, *Section 4* presents the conclusions of the study.

2. Research Methodology and Algorithms

In this section, the proposed methodology is described. Firstly, the used dataset is described as well as the curation, training, and test selection processes. Then, the algorithmic proposals and quality metrics are detailed.

2.1. Data

The required data for the study consisted of the historical energy demand as well as the corresponding data for socio-economic parameters for Turkey, obtained from literature [10]. This data was compiled from MENR, the energy reports, and some of the previously reported studies in the literature [32]. The data consisted of Turkey's GDP ($\$10^9$), Population ($10^6$), and the amounts of Import ($\$10^9$) and Export ($\$10^9$) for the period from 1979 to 2011 (as annual values). These four elements are generally considered to have the biggest impact on the energy demand. Therefore these are considered as the input variables for the present study as well. The target variable for estimation (as well as prediction), is the energy demand (MTOE). Data obtained can be found in the appendix (Table A.1). Notice that the input variables are also named *X1* (GDP), *X2* (Population), *X3* (Quantity of Imports) to *X4* (Quantity of Exports) while the energy demand is denoted as *E*.

From the historical dataset of the energy demand for Turkey, an interesting trend can be deduced for the period studied. The energy demand growth rate ranges from a minimum of -6.34% (for the year 2001) to a maximum of 10.38% (for the year 1987) with an average growth rate of 4.26%. While the input parameters growth rate are: GDP (Average 8.41%, minimum -27.00% and maximum 39.81%), population (Average 1.57, minimum -3.26 and maximum 2.95), import (Average 14.81%, minimum -30.22% and maximum 56.02%), and export quantities (Average 14.58%, minimum -22.64% and maximum 61.51%). It will be, therefore, interesting to find out how these socio-economic parameters correlate to energy demand and further their impact on its future.

In order to avoid the influence on a model of the particular amount of each variable (due to the units used in each one of them), all the data have been normalized by the maximum value of each corresponding parameter. This way, every column has been transformed into the values that correspond to the ratio with the maximum value of the data:

$$\left(X_{i,n} = {X_{i,j}}\Big/{X_{i,max}}\right) \quad (1)$$

where i = 1, 2, 3, 4. j = 1, 2…n, n being the number of years, and *max* being the maximum value of the ith series.

Therefore, the maximum value of a given column is 1, while the rest of the values will be lower. This normalization is also done for the output variable. The data which spans 33 years were divided into two datasets i.e., *training* and *test*, where the training dataset consists of data for 17 years and the test dataset consists of the dataset for the remaining 16 years. The selection for each dataset was made randomly.

2.2. Problem definition and objective function

The search for a mathematical expression that captures the behavior of a target variable can be tackled as an optimization

problem. In particular, this approach aims to find the expression with the minimum error, which in this case, is modeling the energy demand.

To assess the quality of a model, the root mean squared error (RMSE) metric, as shown in Equation (2), is used as the objective function.

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}\left(E_{est,i} - E_{act,i}\right)^2\right]^{\frac{1}{2}} \quad (2)$$

where, $E_{act,i}$ and $E_{est,i}$ in Equation (2) are the actual and estimated values of energy demand.

While other statistics viz, average error (AE), coefficient of determination (R^2), absolute error (ABS), and relative error (RE) were also used to evaluate the accuracy as follows:

$$AE = \frac{1}{n}\sum_{i=1}^{n}(E_{est,i} - E_{act,i}) \quad (3)$$

$$R^2 = \frac{\sum_{i=1}^{n}(E_{est,i}-E_{est,avg})(E_{act,i}-E_{act,avg})}{\sqrt{\sum_{i=1}^{n}(E_{est,i}-E_{est,avg})^2 \sum_{i=1}^{n}\left(E_{act,i}-E_{act,avg}\right)^2}} \quad (4)$$

$$ABS = \frac{1}{n}\sum_{i=1}^{n}\left|E_{est,i} - E_{act,i}\right| \quad (5)$$

$$RE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|E_{est,i}-E_{act,i}\right|}{E_{act,i}} \quad (6)$$

2.3. Algorithmic Methods

In this work, the deployment of Grammatical Evolution (GE) and Differential Evolution (DE) is proposed for the estimation of energy demand. GE is a metaheuristic algorithm belonging to the family of Genetic Programming whose main advantage is the ability to direct the search of the algorithm using grammar [34]. This way, researchers may introduce knowledge about the problem into the grammar, reducing the search space. On the other hand, DE is a metaheuristic algorithm better suited for problems where some parameter values need to be found [35].

Despite the combination of GE and DE has been proven to be effective in the past for this problem [16],[36,37], one of the main contributions of this work is the comparison with different combinations of GE and DE determined by different grammars using a reduced number of input variables. The tool used in this work is WebGE, which is an open-source optimization tool that implements both GE and DE algorithms described before. We refer the reader to [38] for further information.

In particular, four different approaches are proposed in this work. The first one is the use of recursive grammar in GE with no DE intervention (M1). Figure 1 shows the grammar applied to this approach. Elements on the left-hand side of each "::=" symbol are non-terminal values, which must be decoded using any of the productions on the right-hand side, separated by "|" symbols. As seen in the figure, this grammar can generate mathematical expressions using the four input variables (X1 to X4), constant values, addition, subtraction, and product arithmetic operators, and the exponential, power, and logarithmic functions. It is important to note that the <param> non-terminal symbol is devoted to generating constant numbers in the range [0.00,99.99].

The second approach, named M2, is directed by the grammar shown in Figure 2. The main idea is to produce expressions where arithmetic combinations of a parameter multiplied by an input variable are generated. In this case, DE is in charge of finding out the best parameter values (w_i) for each expression generated by GE. Notice that the <digit> rule is not needed since the parameters, represented by *wi*, are found by DE.

The third approach is similar to the previous one but includes the use of the quadratic values of input variables. This approach is termed M3 and is guided by the grammar shown in Figure 3. Again, DE is in charge of finding out the values of the parameters.

```
# Expressions
<recExpr> ::= <expr> | <expr> <op> <recExpr>
<expr> ::= <param> <op> <var> | <param> <op> (<var>)^(<param>) |
exp(abs(<param> <op> <var>)) | log(abs(<param> <op> <var>))
# Parameters
<param> ::= <digit><digit>.<digit><digit>
<digit> ::= 0|1|2|3|4|5|6|7|8|9
# Input variables
<var> ::= X1|X2|X3|X4
# Operands
<op> ::= +|-|*
```

Figure 1: GE executed on a recursive grammar (M1)

```
# Expressions
<expr> ::= (<param> <op> <var>) | (<param> * <var>) <op> (<param> * <var>) |
(<param> * <var>) <op> (<param> * <var>) <op> (<param> <op> <var>) |
(<param> * <var>) <op> (<param> * <var>) <op> (<param> <op> <var>) <op>
(<param> <op> <var>)
# Parameters
<param> ::= w1 | w2 | w3 | w4
# Input variables
<var> ::= X1 | X2 | X3 | X4
# Operands
<op> ::= + | - | *
```

Figure 2: Ensemble of GE executed on a linear grammar and DE (M2)

```
# Expressions
<expr> ::= (<param> <op> <var>) | (<param> * <var>) <op> (<param> * <var>) |
(<param> * <var>) <op> (<param> * <var>) <op> (<param> <op> <var>) |
(<param> * <var>) <op> (<param> * <var>) <op> (<param> <op> <var>) <op>
(<param> <op> <var>) | (<param> * <var> * <var>) | (<param> * <var> * <var>)
<op> (<param> * <var> * <var>) | (<param> * <var> * <var>) <op> (<param> *
<var> * <var>) <op> (<param> * <var> * <var>) | (<param> * <var> * <var>) <op>
(<param> * <var> * <var>) <op> (<param> * <var> * <var>) <op> (<param> *
<var> * <var>) | (<param> * <var> * <var>) <op> (<param> * <var> * <var>) <op>
(<param> * <var> * <var>) <op> (<param> * <var> * <var>) <op> (<param> *
<var> * <var>) | (<param> * <var> * <var>) <op> (<param> * <var> * <var>) <op>
(<param> * <var> * <var>) <op> (<param> * <var> * <var>) <op> (<param> *
<var> * <var>) <op> (<param> * <var> * <var>)
# Parameters
<param> ::= w1 | w2 | w3 | w4
# Input variables
<var> ::= X1 | X2 | X3 | X4
# Operands
<op> ::= + | - | *
```

Figure 3: Ensemble of GE executed on quadratic grammar and DE (M3)

Finally, the last approach is the ensemble of GE and DE using recursive grammar. In this case, first, the GE approach is taken, but allowing DE to look for the parameter values. Figure 4 shows the proposed grammar for this approach.

```
# Expressions
<recExpr> ::= <expr> | <expr> <op> <recExpr>
<expr> ::= <param> <op> <var> | <param> <op> (<var>)^(<param>) |
exp(abs(<param> <op> <var>)) | log(abs(<param> <op> <var>))
# Parameters
<param> ::= w1|w2|w3|w4
# Input variables
<var> ::= X1|X2|X3|X4
# Operands
<op> ::= +|-|*
```

Figure 4: Ensemble of GE executed a recursive grammar and DE (M4)

3. Results and discussion

The experimental experience consisted of the execution of the proposed algorithms over the training dataset under the four configurations presented in the previous section. In particular, 20 runs were executed for each one of the four approaches over the training dataset, obtaining a total amount of 80 different models. Table 1 shows the values of the parameters for the GE and DE algorithms, selected after preliminary experimentation.

Table 1: Details of the experiments

GE Parameters	
Generations	50
Crossover Probability	0.65
Population	20
Mutation Probability	0.1
Max wraps	3
Number of Codons	100
Tournament	2
Number of Runs	20
DE Parameters	
Recombination Factor	0.88
Mutation Factor	0.47
Population Size	20

The results presented in the following sections first describe the estimation followed by the prediction. On one hand, for the estimation problem, the values of the input variables of a given year are used to estimate the energy demand of the same year. On the other hand for prediction, the models have been adapted for the prediction of future energy demand, where the values of the input variables of a given year are used to predict the energy demand of the following year.

3.1. Estimation of Energy Demand

The energy demand estimations were obtained for all the years from 1979 to 2011 using the models generated by the proposed algorithms. In particular, for each one of the four approaches developed under the present study (M1, M2, M3, and M4), an average estimation using the 20 generated models has been evaluated for the target period.

The results of the estimation for the whole dataset are presented in Table 2 considering the average prediction of the 20 models obtained for each one of the four proposed approaches. The previously reported [10] two models Linear (DEL) and Quadratic (DEQ) are also included in this comparison.

Table 2: Estimations of the previously reported differential algorithms based on linear and quadratic [10] with the four model forms of the present approach i.e., M1, M2, M3 and M4

Year	Actual Energy (MTOE)	Linear (DEL) [10]	Quadratic (DEQ) [10]	M1	M2	M3	M4
1979	30.71	32.28	34.48	32.07	30.06	31.88	32.11
1980	31.97	30.89	31.26	29.79	30.71	30.99	31.11
1981	32.05	32.52	33.33	32.09	32.99	32.54	32.67
1982	34.39	34.15	34.84	34.50	34.48	34.01	34.12
1983	35.7	36.53	35.85	36.95	36.29	36.24	36.28

1984	37.43	38.73	37.01	39.50	38.84	38.39	38.39
1985	39.4	40.95	40.10	42.11	40.89	40.43	40.45
1986	42.47	43.27	42.78	44.48	42.27	42.50	42.52
1987	46.88	45.30	45.28	46.89	45.47	44.60	44.63
1988	47.91	46.89	47.68	49.38	47.12	46.02	46.09
1989	50.71	49.76	50.43	51.91	49.40	48.75	48.82
1990	52.98	54.02	53.08	54.55	53.86	53.22	53.25
1991	54.27	55.33	54.49	56.95	54.77	54.25	54.34
1992	56.68	57.52	56.42	59.30	56.85	56.41	56.49
1993	60.26	61.79	61.00	61.71	60.53	61.00	60.92
1994	59.12	60.08	60.48	64.20	59.63	58.79	58.87
1995	63.68	65.28	65.29	66.76	65.26	64.80	64.70
1996	**69.86**	**69.71**	**70.00**	**69.41**	**68.90**	**69.69**	**69.41**
1997	**73.78**	**72.31**	**73.17**	**72.13**	**71.64**	**72.54**	**72.24**
1998	**74.71**	**73.30**	**74.92**	**74.95**	**72.50**	**73.05**	**72.92**
1999	**76.77**	**74.18**	**75.47**	**78.44**	**72.74**	**73.30**	**73.29**
2000	**80.5**	**80.71**	**81.04**	**80.89**	**77.40**	**80.84**	**80.43**
2001	**75.4**	**75.71**	**74.60**	**83.28**	**75.13**	**74.81**	**74.78**
2002	**78.33**	**79.13**	**80.32**	**85.65**	**79.05**	**78.89**	**78.86**
2003	**83.84**	**82.36**	**84.01**	**88.05**	**84.57**	**83.47**	**83.54**
2004	**87.82**	**87.19**	**88.04**	**90.45**	**91.57**	**90.62**	**90.68**
2005	**91.58**	**93.10**	**92.84**	**95.31**	**97.36**	**97.83**	**97.90**
2006	99.59	96.25	56.00	95.31	101.70	102.69	102.57
2007	107.63	92.76	4.97	88.98	103.98	102.43	102.66
2008	106.27	94.19	-136.98	90.40	109.84	106.49	107.15
2009	106.14	90.17	-495.07	96.02	103.81	96.02	95.47
2010	109.27	103.09	-2.75	99.43	110.96	113.23	113.12
2011	114.48	114.50	64.97	100.12	118.99	129.30	128.27

Highlighted values represent the period of consideration in the study conducted by Beskirli et a [10]

Notice that the previous comparison of algorithms presented in the state of the art was made for a small range of years, corresponding to the period from 1996 to 2005 as highlighted in Table 2. However, in this work, the comparison is made for the entire range of the years (1979-2011) as previously described. Correspondingly, Table 3 compares the estimation performance over the test dataset of the four proposed approaches with the two methods found in the state of the art: a linear model (DEL) and a quadratic model (DEQ). For all of them, 5 error metrics have been obtained: root mean squared error (RMSE), Average Error (AE), R^2, Absolute Error (ABS), and Relative Error (RE). Notice that the depicted values correspond to the average value of the 20 models in the case of the four proposed approaches.

Table 3: Error metrics for the compared models over the test data. Best results are depicted in bold fonts.

Year	Linear (DEL) [10]	Quadratic (DEQ) [10]	M1	M2	M3	M4
RMSE	4.6403	116.5278	6.1913	3.7255	3.6481	**2.2002**
AE	2.4685	35.5655	4.0700	2.0997	2.1008	**1.6850**
R^2	0.9771	0.9812	0.9504	0.9807	0.9812	**0.9931**
ABS	81.4605	1173.6609	134.1800	69.2909	69.3275	**55.6063**
RE	0.0316	0.3408	0.0520	0.0269	0.0270	**0.0238**

As seen in the table, all the proposed approaches except M1 (GE with no DE) obtain superior results than the algorithms from the state of the art in terms of the error metrics. Further as observed, the M4 proposal exhibits the top performance among the proposed algorithms with the lowest values of RMSE (2.2002), Average Error (1.6850), Absolute Error (55.6063), Relative Error (0.0238), and the highest value of R^2 (0.9931). This can be attributed to the fact that the flexible model structure of the recursive grammar combined with the DE efficiency can obtain better parameter values, resulting in a much better performance in comparison to those approaches where the model structure is fixed. Therefore, the ensemble of GE and DE with a recursive grammar (M4) is the most appropriate algorithm for the estimation of energy demand. Hence, this appropriate combination of metaheuristic algorithms (M4) will be used to predict the energy demand for the future.

3.2. Prediction of Future Energy Demand

The confirmation of the performance improvement achieved from the proposed methods has been made in the previous section. It is established that Model M4 which represents an ensemble of a recursive GE combined with DE has the most accurate outcomes when applied to energy demand problems.

We now move on to the problem of predicting the year-ahead energy demand. To this aim, the set of input parameters remains the same as they were previously discussed. However, the algorithm took the previous year's data as input variables to predict the energy demand for the current year. This way a year ahead energy demand approach is established. In a similar fashion, the training and test datasets were separated, where the training dataset was used to train the algorithms while the test dataset was used as independent data to test the performance of the algorithms.

The experimentation then followed the same pattern: 20 runs were executed for the GE with a recursive grammar together with DE combination (with the same properties as previously discussed in Table 1) over the training dataset which produced the 20 models (which can be found in Appendix Table A.2). The results of the performance of algorithms in predicting the future energy demand are depicted in figures 5 (a) and (b) for training and test dataset, respectively, as an average of the 20 runs executed for energy prediction. It must be again noted here that since the input to the algorithm is normalized the output that is received is also in the normalized form. This can also be observed in figures 5(a) and (b).

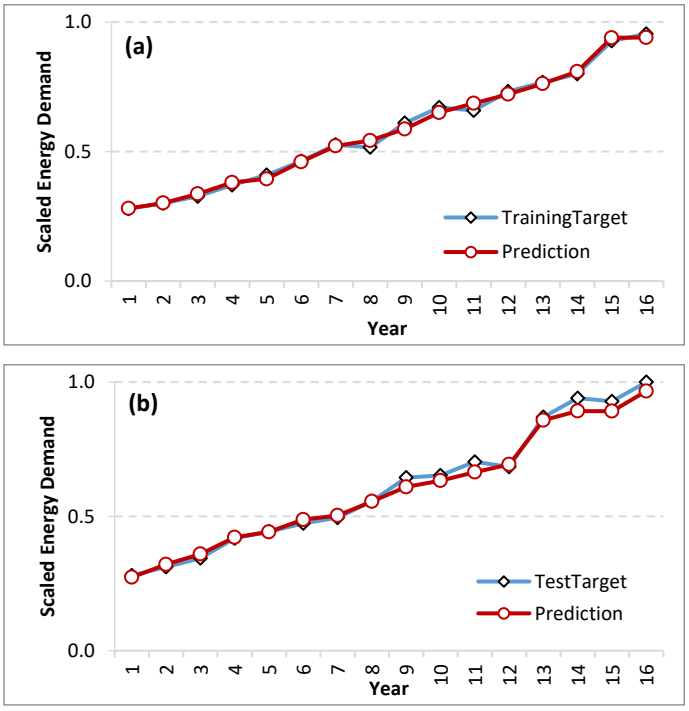


Figure 5: Predicted and actual energy demand (normalized) on (a) training and (b) test datasets. The X-axis denotes the ordinal number of the year in the dataset.

From the figure, it is observed that the predictions match the data very precisely and the errors are small. Table 4 shows the error metrics obtained for the prediction of a year ahead energy demand with M4. It is observed that the algorithm has high accuracy and performs well in terms of prediction accuracy as justified by the small errors produced on the test dataset. Also, since the training and test datasets were independent the issue of overfitting a model on the data is avoided, resulting in a more flexible and accurate model. The predictions of energy demand are presented below in Figure 6 with actual MTOE units for the complete dataset. As seen, the proposed method shows high accuracy on a one-year ahead prediction.

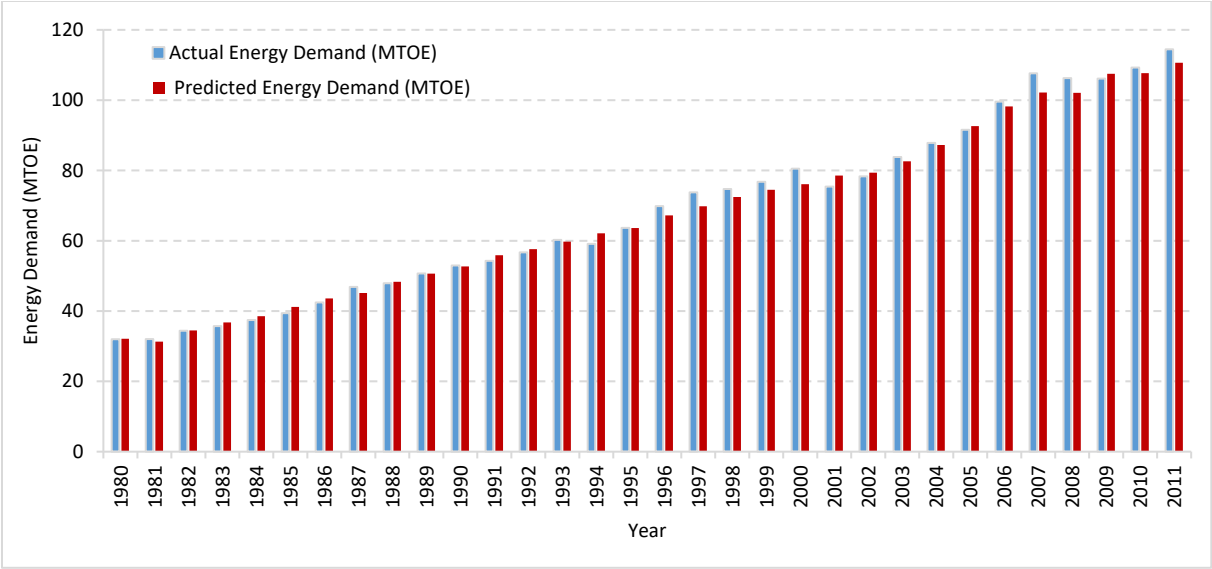


Figure 6: Evolution and comparison of actual and predicted energy demand (using the average of the 20 experimental runs)

The error computation is also made based on Figure. 6 and the obtained results were the following: RMSE=2.2278, AE=-0.6099, R^2= 0.9970, ABS=1.7263, and RE =0.0244. As seen, they are similar to the results with separated training and test data.

Table 4: Errors between predicted values from M4 and actual data for energy demand (normalized)

Error	Training	Test
RMSE	0.0146	0.0233
AE	0.0001	0.0107
R^2	0.9952	0.9959
ABS	0.1917	0.2908
RE	0.0210	0.0279

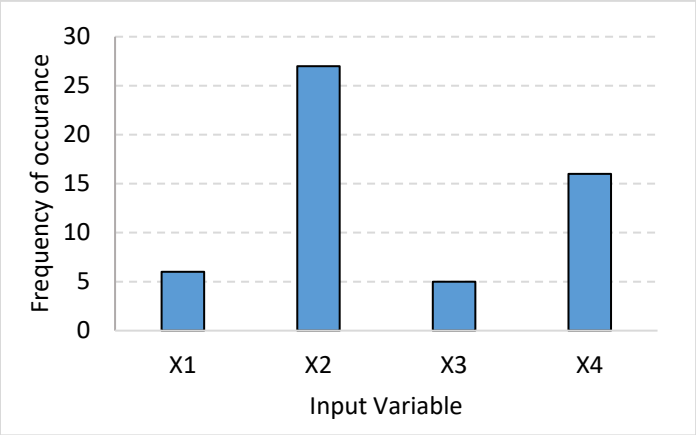


Figure 7: Histogram obtained from 20 runs of GE-DE

The details of prediction models are provided in the appendix (Table A.2). On further analysis of the obtained models, the number of terms appearing in the expressions ranges from 2 to 5, where some of the input parameters appear several times in the different model as well as different expressions. This leads us to find out the most influential parameter among the set of input parameters. To this aim, we calculate the frequency of occurrence of the four input parameters in the 20 models. The occurrence of the feature or the input variable in the equations is depicted in Figure 7. It can be observed that X2 (Population) appears the maximum number of times in the equations, followed by the X4 (quantity of export), the input X1 (GDP), and lastly the X3 (quantity of import). This means that energy demand primarily grows in correlation to the population of the country which in fact is imperative since the larger the population, the more is energy demand.

4. Conclusions

Country-wide energy demand predictions have been made considering socio-economic parameters based on metaheuristic algorithms for Turkey. Four methods were considered initially for the estimation of energy demand: (M1) Grammatical Evolution with a recursive grammar alone, (M2) Grammatical Evolution with a linear grammar together with Differential Evolution, (M3) Grammatical Evolution with a quadratic grammar together with Differential Evolution, and finally, (M4) Grammatical Evolution with a recursive grammar combined with Differential Evolution. The ensemble of GE with a recursive grammar and DE together (M4) performed very well in comparison to others with the least value of objective function RMSE=2.2002. Further, this M4 proposal was deployed to predict a year-ahead energy demand for Turkey. The algorithm again performed very well obtaining the predictions for a year ahead, and the results of prediction were deemed to be exceptionally well with the objective function to be RMSE=2.2278. Therefore, it is recommended to use the combination of Grammatical Evolution with recursive grammar to provide for more flexible model forms. Further, the use of Differential Evolution together with the recursive algorithm of GE provides optimization of the model weight leading to accurate predictions of the energy demand for a country. It was also inferred that population greatly affects the structure of the prediction models and thus is a significant parameter for energy demand

models. It is suggested that for future studies long-term causal relationships are established along with the elasticities of algorithms. This will be significant and reflect a more certain behavior to predict the energy demand. Further, it is advised to investigate the behavior of algorithms with more input parameters to elucidate influential and non-influential parameters for energy demand estimation/prediction. This can also implicate the direct and indirect relationship between the input parameters and the energy demand.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The work reported under the manuscript has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 754382. The authors would like to thank Prof. Abraham Duarte for extending his support towards the research work carried out in the article.

References

[1] S. Katircioglu, C. Köksal, S. Katircioglu, "The role of financial systems in energy demand: A comparison of developed and developing countries," Heliyon, **7**(6), e07323, 2021, doi:https://doi.org/10.1016/j.heliyon.2021.e07323.

[2] S. Di Leo, P. Caramuta, P. Curci, C. Cosmi, "Regression analysis for energy demand projection: An application to TIMES-Basilicata and TIMES-Italy energy models," Energy, **196**, 117058, 2020, doi:https://doi.org/10.1016/j.energy.2020.117058.

[3] C. Huang, Z. Zhang, N. Li, Y. Liu, X. Chen, F. Liu, "Estimating economic impacts from future energy demand changes due to climate change and economic development in China," Journal of Cleaner Production, **311**, 127576, 2021, doi:https://doi.org/10.1016/j.jclepro.2021.127576.

[4] Y. Yu, N. Zhang, J.D. Kim, "Impact of urbanization on energy demand: An empirical study of the Yangtze River Economic Belt in China," Energy Policy, **139**, 111354, 2020, doi:https://doi.org/10.1016/j.enpol.2020.111354.

[5] S. Qiu, T. Lei, J. Wu, S. Bi, "Energy demand and supply planning of China through 2060," Energy, **234**, 121193, 2021, doi:https://doi.org/10.1016/j.energy.2021.121193.

[6] J. Huang, H. Zhang, W. Peng, C. Hu, "Impact of energy technology and structural change on energy demand in China," Science of The Total Environment, **760**, 143345, 2021, doi:https://doi.org/10.1016/j.scitotenv.2020.143345.

[7] K. Oshiro, S. Fujimori, Y. Ochi, T. Ehara, "Enabling energy system transition toward decarbonization in Japan through energy service demand reduction," Energy, **227**, 120464, 2021, doi:https://doi.org/10.1016/j.energy.2021.120464.

[8] P. Späth, H. Rohracher, "Local Demonstrations for Global Transitions—Dynamics across Governance Levels Fostering Socio-Technical Regime Change Towards Sustainability," European Planning Studies, **20**(3), 461–479, 2012, doi:10.1080/09654313.2012.651800.

[9] M.A. Islam, H.S. Che, M. Hasanuzzaman, N.A. Rahim, Chapter 5 - Energy demand forecasting, Academic Press: 105–123, 2020, doi:https://doi.org/10.1016/B978-0-12-814645-3.00005-5.

[10] M. BESKIRLI, H. HAKLI, H. KODAZ, "The energy demand estimation for Turkey using differential evolution algorithm," Sādhanā, **42**(10), 1705–1715, 2017, doi:10.1007/s12046-017-0724-7.

[11] Q. Wang, S. Li, R. Li, "Forecasting energy demand in China and India: Using single-linear, hybrid-linear, and non-linear time series forecast techniques," Energy, **161**, 821–831, 2018, doi:https://doi.org/10.1016/j.energy.2018.07.168.

[12] S. Salcedo-Sanz, J. Muñoz-Bulnes, J.A. Portilla-Figueras, J. Del Ser, "One-year-ahead energy demand estimation from macroeconomic variables using computational intelligence algorithms," Energy Conversion and Management, **99**, 62–71, 2015, doi:https://doi.org/10.1016/j.enconman.2015.03.109.

[13] J. Sánchez-Oro, A. Duarte, S. Salcedo-Sanz, "Robust total energy demand estimation with a hybrid Variable Neighborhood Search – Extreme Learning Machine algorithm," Energy Conversion and Management, **123**, 445–452, 2016, doi:https://doi.org/10.1016/j.enconman.2016.06.050.

[14] J. Huang, Y. Tang, S. Chen, "Energy Demand Forecasting: Combining Cointegration Analysis and Artificial Intelligence Algorithm," Mathematical Problems in Engineering, **2018**(5194810), 1–13, 2018, doi:10.1155/2018/5194810.

[15] Q. Wu, C. Peng, "A hybrid BAG-SA optimal approach to estimate energy demand of China," Energy, **120**, 985–995, 2017, doi:https://doi.org/10.1016/j.energy.2016.12.002.

[16] J.M. Colmenar, J.I. Hidalgo, S. Salcedo-Sanz, "Automatic generation of models for energy demand estimation using Grammatical Evolution," Energy, **164**, 183–193, 2018, doi:https://doi.org/10.1016/j.energy.2018.08.199.

[17] M.A. Behrang, E. Assareh, M.R. Assari, A. Ghanbarzadeh, "Total Energy Demand Estimation in Iran Using Bees Algorithm," Energy Sources, Part B: Economics, Planning, and Policy, **6**(3), 294–303, 2011, doi:10.1080/15567240903502594.

[18] S. Yu, Y.-M. Wei, K. Wang, "China's primary energy demands in 2020: Predictions from an MPSO–RBF estimation model," Energy Conversion and Management, **61**, 59–66, 2012, doi:https://doi.org/10.1016/j.enconman.2012.03.016.

[19] X. Yin, Q. Zhang, H. Wang, Z. Ding, "RBFNN-Based Minimum Entropy Filtering for a Class of Stochastic Nonlinear Systems," IEEE Transactions on Automatic Control, **65**(1), 376–381, 2020, doi:10.1109/TAC.2019.2914257.

[20] S. Yu, Y.-M. Wei, K. Wang, "A PSO–GA optimal model to estimate primary energy demand of China," Energy Policy, **42**, 329–340, 2012, doi:https://doi.org/10.1016/j.enpol.2011.11.090.

[21] Z.W. Geem, W.E. Roper, "Energy demand estimation of South Korea using artificial neural network," Energy Policy, **37**(10), 4049–4054, 2009, doi:https://doi.org/10.1016/j.enpol.2009.04.049.

[22] J. Liu, Y. Liu, Q. Zhang, "A weight initialization method based on neural network with asymmetric activation function," Neurocomputing, **483**, 171–182, 2022, doi:https://doi.org/10.1016/j.neucom.2022.01.088.

[23] X.-C. Yuan, X. Sun, W. Zhao, Z. Mi, B. Wang, Y.-M. Wei, "Forecasting China's regional energy demand by 2030: A Bayesian approach," Resources, Conservation and Recycling, **127**, 85–95, 2017, doi:https://doi.org/10.1016/j.resconrec.2017.08.016.

[24] Y. He, B. Lin, "Forecasting China's total energy demand and its structure using ADL-MIDAS model," Energy, **151**, 420–429, 2018, doi:https://doi.org/10.1016/j.energy.2018.03.067.

[25] R. Chen, Z. Rao, G. Liu, Y. Chen, S. Liao, "The long-term forecast of energy demand and uncertainty evaluation with limited data for energy-imported cities in China: a case study in Hunan," Energy Procedia, **160**, 396–403, 2019, doi:https://doi.org/10.1016/j.egypro.2019.02.173.

[26] H. Wang, Z. Chen, W. Wang, Z. Wu, K. Wu, W. Li, "Improving Energy Demand Estimation Using an Adaptive Firefly Algorithm BT - Computational Intelligence and Intelligent Systems," in: Li, K., Li, W., Chen, Z., and Liu, Y., eds., Springer Singapore, Singapore: 171–181, 2018.

[27] M. Duran Toksarı, "Ant colony optimization approach to estimate energy demand of Turkey," Energy Policy, **35**(8), 3984–3990, 2007, doi:https://doi.org/10.1016/j.enpol.2007.01.028.

[28] A. Ünler, "Improvement of energy demand forecasts using swarm intelligence: The case of Turkey with projections to 2025," Energy Policy, **36**(6), 1937–1944, 2008, doi:https://doi.org/10.1016/j.enpol.2008.02.018.

[29] O. ERSEL CANYURT, H. CEYLAN, H. KEMAL OZTURK, A. HEPBASLI, "Energy Demand Estimation Based on Two-Different Genetic Algorithm Approaches," Energy Sources, **26**(14), 1313–1320, 2004, doi:10.1080/00908310490441610.

[30] A. Beşkirli, M. Beşkirli, H. Haklı, H. Uğuz, "Comparing Energy Demand Estimation Using Artificial Algae Algorithm : The Case of Turkey," Journal of Clean Energy Technologies, **6**(4), 349–352, 2018, doi:10.18178/jocet.2018.6.4.487.

[31] T. Paksoy, G. Weber, "Particle Swarm Optimization Approach for Estimation of Energy Demand of Turkey," Global Journal of Technology & Optimization, **3**(June), 1–9, 2012.

[32] Y.M. Bulut, Z. Yildiz, "Comparing energy demand estimation using various statistical methods: The case of Turkey," Gazi University Journal of Science, **29**(2), 237–244, 2016.

[33] A. Löschel, S. Managi, "Recent Advances in Energy Demand Analysis—Insights for Industry and Households," Resource and Energy Economics, **56**, 1–5, 2019, doi:https://doi.org/10.1016/j.reseneeco.2019.04.001.

[34] M. O'Neill, C. Ryan, "Grammatical evolution," IEEE Transactions on

Evolutionary Computation, **5**(4), 349–358, 2001, doi:10.1109/4235.942529.
[35] R. Storn, K. Price, "Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces," Journal of Global Optimization, **11**(4), 341–359, 1997, doi:10.1023/A:1008202821328.
[36] N. Lourenço, J.M. Colmenar, J.I. Hidalgo, S. Salcedo-Sanz, "Evolving energy demand estimation models over macroeconomic indicators," GECCO 2020 - Proceedings of the 2020 Genetic and Evolutionary Computation Conference, 1143–1149, 2020, doi:10.1145/3377930.3390153.
[37] B. Jamil, L. Serrano-Luján, J.M. Colmenar, "Modelling energy consumption in Spain with metaheuristic methods," in 2021 6th International Conference on Smart and Sustainable Technologies (SpliTech), 1–3, 2021, doi:10.23919/SpliTech52315.2021.9566391.
[38] J.M. Colmenar, R. Martín-Santamaría, J.I. Hidalgo, "WebGE: An Open-Source Tool for Symbolic Regression Using Grammatical Evolution BT - Applications of Evolutionary Computation," in: Jiménez Laredo, J. L., Hidalgo, J. I., and Babaagba, K. O., eds., Springer International Publishing, Cham: 269–282, 2022.

Appendix

Table A. 1 provides the historical data on energy demand and the socio-economic parameters used in the study. Table A.2 gives the models generated by the algorithm.

Table A.1: Data of Turkey's actual energy demand, GDP, population, imports, and exports [10]

Year	Energy Demand (MTOE)	GDP (10^9)	Population (10^6)	Import (10^9)	Export (10^9)
	E	X1	X2	X3	X4
1979	30.71	82	45.53	5.07	2.26
1980	31.97	68	44.44	7.91	2.91
1981	32.05	72	45.54	8.93	4.70
1982	34.39	64	46.69	8.84	5.75
1983	35.70	60	47.86	9.24	5.73
1984	37.43	59	49.07	10.76	7.13
1985	39.40	67	50.31	11.34	7.95
1986	42.47	75	51.43	11.10	7.46
1987	46.88	86	52.56	14.16	10.19
1988	47.91	90	53.72	14.34	11.66
1989	50.71	108	54.89	15.79	11.62
1990	52.98	151	56.10	22.30	12.96
1991	54.27	150	57.19	21.05	13.59
1992	56.68	158	58.25	22.87	14.72
1993	60.26	179	59.32	29.43	15.35
1994	59.12	132	60.42	23.27	18.11
1995	63.68	170	61.53	35.71	21.64
1996	69.86	184	62.67	43.63	23.22
1997	73.78	192	63.82	48.56	26.26
1998	74.71	207	65.00	45.92	26.97
1999	76.77	187	66.43	40.67	26.59
2000	80.50	200	67.42	54.50	27.78
2001	75.40	146	68.37	41.40	31.33

2002	78.33	181	69.30	51.55	36.06
2003	83.84	239	70.23	69.34	47.25
2004	87.82	299	71.15	97.54	63.17
2005	91.58	361	72.97	116.77	73.48
2006	99.59	483	72.97	139.58	85.54
2007	107.63	531	70.59	170.06	107.27
2008	106.27	648	71.13	201.96	132.03
2009	106.14	730	73.23	140.93	102.14
2010	109.27	615	74.47	185.54	113.88
2011	114.48	731	74.72	240.84	134.91

Table A.2: Equations produced from the execution of ensemble of GE executed on a recursive grammar and DE (M4) for prediction of energy demand.

S.No.	Model
1.	$\frac{E}{E_{max}} = \left[0.194\left(\frac{X4_i}{X4_{max}}\right)^{1.443}\right] + \left[-0.194 + \left(\frac{X2_i}{X2_{max}}\right)^{1.443}\right]$
2.	$\frac{E}{E_{max}} = \left[-0.047 + \left(\frac{X4_i}{X4_{max}}\right)^{0.180}\right] - \left[-0.047\left(\frac{X2_i}{X2_{max}}\right)\right] \times \left[-0.047 - \left(\frac{X4_i}{X4_{max}}\right)^{-0.047}\right] + \left[e^{\left\lvert -0.047\left(\frac{X4_i}{X4_{max}}\right)\right\rvert}\right]$
3.	$\frac{E}{E_{max}} = \left[-0.638 + \left(\frac{X2_i}{X2_{max}}\right)\right] + \left[-0.638 - \left(\frac{X1_i}{X1_{max}}\right)\right] \times \left[\log\left\lvert -0.638 - \left(\frac{X2_i}{X2_{max}}\right)\right\rvert\right] \times \left[-0.638 + \left(\frac{X4_i}{X4_{max}}\right)^{0.014}\right]$
4.	$\frac{E}{E_{max}} = \left[e^{\left\lvert 0.360 + \left(\frac{X2_i}{X2_{max}}\right)\right\rvert}\right] \times \left[0.360 + \left(\frac{X4_i}{X4_{max}}\right)^{0.152}\right] - \left[e^{\left\lvert 0.360\left(\frac{X2_i}{X2_{max}}\right)\right\rvert}\right]$
5.	$\frac{E}{E_{max}} = \left[-0.242 - \left(\frac{X4_i}{X4_{max}}\right)\right] \times \left[-0.242 + \left(\frac{X2_i}{X2_{max}}\right)^{1.283}\right]$
6.	$\frac{E}{E_{max}} = \left[1.417\left(\frac{X2_i}{X2_{max}}\right)^{0.508}\right] + \left[log\left\lvert 1.417\left(\frac{X2_i}{X2_{max}}\right)\right\rvert\right] \times \left[0.508\left(\frac{X4_i}{X4_{max}}\right)^{1.417}\right] + \left[log\left\lvert 0.508\left(\frac{X2_i}{X2_{max}}\right)\right\rvert\right] \times \left[log\left\lvert 1.417 + \left(\frac{X2_i}{X2_{max}}\right)\right\rvert\right]$
7.	$\frac{E}{E_{max}} = \left[0.999\left(\frac{X3_i}{X3_{max}}\right)^{0.199}\right] \times \left[0.999\left(\frac{X2_i}{X2_{max}}\right)\right]$
8.	$\frac{E}{E_{max}} = \left[0.547\left(\frac{X3_i}{X3_{max}}\right)^{0.390}\right] + \left[log\left\lvert 0.547 + \left(\frac{X2_i}{X2_{max}}\right)\right\rvert\right]$
9.	$\frac{E}{E_{max}} = \left[-2.443\left(\frac{X2_i}{X2_{max}}\right)^{0.187}\right] + \left[e^{\left\lvert 0.187 - \left(\frac{X2_i}{X2_{max}}\right)\right\rvert}\right] + \left[e^{\left\lvert 0.187\left(\frac{X3_i}{X3_{max}}\right)\right\rvert}\right]$
10.	$\frac{E}{E_{max}} = \left[4.946\left(\frac{X2_i}{X2_{max}}\right)\right] \times \left[0.203\left(\frac{X4_i}{X4_{max}}\right)^{0.203}\right]$
11.	$\frac{E}{E_{max}} = \left[0.356\left(\frac{X4_i}{X4_{max}}\right)^{0.626}\right] + \left[-0.356 + \left(\frac{X2_i}{X2_{max}}\right)\right]$
12.	$\frac{E}{E_{max}} = \left[0.201\left(\frac{X4_i}{X4_{max}}\right)^{0.201}\right] \times \left[log\left\lvert 0.887\left(\frac{X4_i}{X4_{max}}\right)\right\rvert\right] \times \left[0.887 + \left(\frac{X2_i}{X2_{max}}\right)\right]$

13.	$\frac{E}{E_{max}} = \left[-0.189\left(\frac{X4_i}{X4_{max}}\right)^{-0.189}\right] - \left[e^{\left\|0.716\left(\frac{X3_i}{X3_{max}}\right)\right\|}\right] \times \left[log\left\|0.716 + \left(\frac{X1_i}{X1_{max}}\right)\right\|\right] \times \left[-0.189 + \left(\frac{X2_i}{X2_{max}}\right)^{0.716}\right]$
14.	$\frac{E}{E_{max}} = \left[-0.427 + \left(\frac{X2_i}{X2_{max}}\right)\right] + \left[0.400\left(\frac{X4_i}{X4_{max}}\right)^{0.400}\right]$
15.	$\frac{E}{E_{max}} = \left[0.356\left(\frac{X4_i}{X4_{max}}\right)^{0.626}\right] - \left[0.356 + \left(\frac{X2_i}{X2_{max}}\right)\right]$
16.	$\frac{E}{E_{max}} = \left[-0.323 + \left(\frac{X2_i}{X2_{max}}\right)\right] - \left[-0.327\left(\frac{X1_i}{X1_{max}}\right)\right]$
17.	$\frac{E}{E_{max}} = \left[log\left\|2.222\left(\frac{X2_i}{X2_{max}}\right)\right\|\right] \times \left[log\left\|2.433 + \left(\frac{X1_i}{X1_{max}}\right)\right\|\right]$
18.	$\frac{E}{E_{max}} = \left[log\left\|2.159\left(\frac{X2_i}{X2_{max}}\right)\right\|\right] - \left[log\left\|0.291 - \left(\frac{X4_i}{X4_{max}}\right)\right\|\right] \times \left[2.159\left(\frac{X1_i}{X1_{max}}\right)\right] \times \left[0.291\left(\frac{X3_i}{X3_{max}}\right)^{2.159}\right]$
19.	$\frac{E}{E_{max}} = \left[-0.534\left(\frac{X2_i}{X2_{max}}\right)^{-0.264}\right] - \left[-0.264 - \left(\frac{X1_i}{X1_{max}}\right)\right] \times \left[-0.264 + \left(\frac{X2_i}{X2_{max}}\right)\right]$
20.	$\frac{E}{E_{max}} = \left[log\left\|2.584 + \left(\frac{X1_i}{X1_{max}}\right)\right\|\right] - \left[\left(-0.711 - \left(\frac{X2_i}{X2_{max}}\right)\right)^{-0.711}\right]$

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 87-102 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Cloud Telemedicine Platform Based on Workflow Management System: A Review of an Italian Case Study

Gianvito Mitrano[*,1], Antonio Caforio[2], Tobia Calogiuri[1], Chiara Colucci[3], Luca Mainetti[1], Roberto Paiano[1], Claudio Pascarelli[1]

[1]*University of Salento, Department of Engineering for Innovation, Via per Monteroni, Lecce, 73100, Italy*

[2]*A-thon s.r.l., Via dei Tufi 4, Monteroni di Lecce (Le), 73047, Italy*

[3]*National Interuniversity Consortium for Informatics, Via Ariosto 25, Roma, 00185, Italy*

ARTICLE INFO

Article history:
Received: 27 July, 2022
Accepted: 16 October, 2022
Online: 26 November, 2022

Keywords:
Telemedicine
Teleconsultation
Telemonitoring
Business Process Management
Workflow Management System

ABSTRACT

The paper aims to describe a new technological and organizational approach in order to manage teleconsultation and telemonitoring processes involving a Physician, who remotely interacts with one or more Specialists, in order to evaluate and discuss the specific clinical conditions of a patient, based primarily on the sharing of digital clinical data, reports and diagnostic images. In the HINT project (Healthcare INtegration in Telemedicine), a teleconsultation and telemonitoring cloud platform has been developed using a Hub and Spoke architecture, based on a Business Process Management System (BPMS). The specialized clinical centres (Hubs) operate in connection with the territorial hospital centres (Spokes), which receive specific diagnostic consultations and telemonitoring data from the appropriate Specialist, supported by advanced AI systems. The developed platform overcomes the concepts of a traditional and fragmented teleconsultation and consequently the static organization of Hubs and Spokes, evolving towards an integrated clinical workflow management. The project platform adopts international healthcare standards, such as HL7 FHIR, IHE (XDS and XDW) and DICOM for the acquisition and management of healthcare data and diagnostic images. A Workflow Management System implemented in the platform allows to manage multiple and contemporaneous processes through a single platform, correctly associating the tasks to the Physicians responsible for their execution, monitoring the status of the health activities and managing possible clinical issues.

1. Introduction

This paper is an extension of work originally presented in the *10th International Conference on Information Systems and Technologies (ICIST 2020)* entitled "*HINT project: a BPM teleconsultation and telemonitoring platform"* [1]. It is focused on the implementation of a telemedicine (teleconsultation and telemonitoring) platform, based on diagnostic imagines, which supports the *Hub and Spoke* clinical model.

In particular, optimized teleconsultation processes, supported by proper technological infrastructures, allow: i) to share data and diagnostic images between several professionals; ii) to have the first or second opinion evaluations in a very short time; iii) to prevent the patient from moving between Physicians for consultations; iv) to provide precise evaluations and diagnoses to the patient by reducing waiting times. Moreover, telemonitoring allows to control the specific vital parameters of each patient and to manage alerts if necessary [2], [3].

Over the years, many definitions of telemedicine have been given, based on the use of IT (Information and Telecommunication Systems) for remote clinical data exchange between the medical team (Physician, Nurses, Specialist) and patients, in order to provide the remote healthcare treatment. Some of these definitions are given below.

The American Telemedicine Association gives a definition of telemedicine as "The use of medical information exchanged from

[*]Corresponding Author: Gianvito Mitrano, University of Salento, Department of Engineering for Innovation, Via per Monteroni, 73100, Lecce, Italy - gianvito.mitrano@unisalento.it

https://dx.doi.org/10.25046/aj070610

one site to another via electronic communications to improve a patient's clinical health status" [4].

The World Health Organization (WHO) defines telemedicine as follows: "The delivery of healthcare services, where distance is a critical factor, by all healthcare professionals using information and communication technologies for the exchange of valid information for diagnosis, treatment and prevention of diseases and injuries, research and evaluation, and for the continuing education of healthcare providers, all in the interests of advancing the health of individuals and their communities" [5].

The outcomes of telemedicine services can be: i) accessibility to healthcare in remote areas, allowing consistent and reliable communication; ii) fast, affordable, and reachable care; iii) h24 diagnosis (also known as night-hawking); iv) efficiency in care delivery like telestroke scenario; v) cost and time optimization and saving; vi) better collaboration between Specialists [6].

Telemedicine can also reduce the risk of hospitalization, thanks to its ability to decrease in-clinic visits, promoting remote services for patients [7]. Another important outcome can be a more effective patient engagement and evolvement, which could help users to be informed overtime about their care plan and become more aware of their health. Telemedicine can also represent an effective way to treat depression and other mental health conditions, helping patients with mental disorders unwilling to seek any treatment in normal circumstances, and making use of video or remote controls.

Telemedicine services can be divided into synchronous and asynchronous services. In the former case, the service establishes an interaction and connection in real time (through audio or video-conferencing technologies) between the parties requiring their presence at the same time; in the latter case, clinical and medical data are temporarily stored and later transmitted to the Physician or medical Specialist for an offline assessment. Synchronous telehealth models can help to provide timely care, especially for emergency care or urgent situations. On the other hand, asynchronous models are ideal for evidence-based and not emergency care [8].

In healthcare, especially in telemedicine, some fundamental aspects concern the quality of data, clinical documents, diagnostic images and the modality to exchange. Physicians and healthcare professionals need to access and share all the patient's information and clinical data, also deriving from telemonitoring activities, and it is, therefore, necessary the interconnection between different Information Systems and the consequent interoperability based on clear and possibly open rules and standards [9], [10].

Interconnected and interoperable systems are defined as integrated if, thanks to the use of standard languages, protocols and formats, they allow to define the methods of transmission and reception of data, as well as their representation (syntax) and their semantics [11].

In the healthcare sector, there are several models and open standards for the information and clinical data management and exchange, among which, for the best of authors' knowledge, the most widely used are IHE (Integrating the Healthcare Enterprise), DICOM (Digital Imaging and Communications in Medicine) and HL7 (Health Level Seven) for the diagnostic images sharing [12].

These standards are used in order to have an effective interoperability between the Health Information Systems (HIS)such as hospital RIS (Radiology Information System), PACS (Picture Archiving and Communication System) and booking centers (e.g., the Italian CUP) [13].

Alongside the standards and new healthcare information technologies mentioned above, it is important to define and optimize clinical processes through the use of management approaches such as the BPM (Business Process Management), which allows to model, define, organize, execute and monitor the clinical activities of interest.

In this paper, after an initial overview of the motivation behind this work, section two provides a narrative literature review about telemonitoring and teleconsultation applications. The third section describes the HINT project, that has developed a cloud platform based on a *Hub and Spoke* architecture where telemedicine processes are managed by a BPM workflow system. The fourth section illustrates these processes modelled through the BPMN notation, with a detailed analysis of the following scenarios: Neuro-Radiological Teleconsultation, Dermatological Teleconsultation and Obstructive Sleep Apnoea Syndrome Telemonitoring. Finally, discussion and conclusions close the paper.

1.1 Motivation

Different telemedicine projects propose generic solutions with static Hub and Spoke architecture [14], [15], in which the teleconsultation and the exchange of clinical data and information take place through traditional clinical assessment, in absence of a single management platform and digitalized telemedicine pathways for the real time sharing of reports, diagnostic images and documents. To overcome a typical static organization, the HINT project envisages the use of a workflow engine for the complete routing and management of the consultation request and parameter monitoring to the appropriate and available Hub.

Moreover, the Hub Specialist can be helped in his/her clinic evaluation by Decision Support Systems (DSSs) based on Deep Learning, which can analyse and classify dermatological images and to manage out-of-threshold values with alerts. Specifically, the project proposal is an advanced teleconsultation and telemonitoring system, in which Physicians can: i) receive technological support through DSSs; ii) are aware of the consulting request status in every moment and iii) provide detailed clinical information to make assessments and final diagnoses. The HINT *Hub and Spoke* system allows the identification of the most appropriate available Specialist for the dermatological case to be treated, who responds on a single, integrated platform with his/her second opinion evaluations. It is particularly valuable in all the cases requiring high healthcare specializations, especially in remote areas where they are not immediately available.

Furthermore, the HINT platform wants to guarantee timely and constant monitoring of specific vital parameters in the treatment of sleep apnoea through UWB devices. It is supported by the implementation of a Workflow Management System for healthcare process digitalization through the application of the IHE standard (XDW) for the definition and traceability of all the steps, actors and documents produced in the clinical care processes.

2. Literature review

In this section, a literature study is carried out to analyse the different telemedicine services implemented in the HINT system,

which are teleradiology, teleconsultation and telemonitoring. A review on information systems and standards is also provided.

Teleradiology is implemented in different countries in order to face the following issues [1], [16]:

1. out-of-hour coverage by means of intramural services (at-home control);
2. out-of-hour coverage by means of extramural services;
3. rural or regional hospital coverage;
4. Specialist's second opinion.

In [16], the authors summarize some of the key aspects in image acquisition, transmission, and interpretation, with a discussion of the clinical practice of teleradiology.

A survey [17] states that 44% of the European radiologists use at-home control instead of call referral. In many cases this procedure is not based on a structured *Hub and Spoke* architecture, but only on the diagnostic image transmission to the Radiologist who sends the referral to the Physician by email or phone (79% of the cases).

It is also important to mention the Kaiser Permanente hospital [18] where, during the out-of-hours, there are only 2 radiologists sending medical reports for 11 different image acquisition sites, instead of having a specific Radiologist, called when needed, in each hospital.

In the U.S., it is estimated that over than 50% of the hospitals use out-of-hours reporting or night-hawking [19]. It is possible to gain an advantage from the different time zones by using extramural services. For example, in [20] radiologists from Sydney are called by a Swedish hospital for emergency; while in [21] an American hospital requests a consultation to radiologists located in India. Moreover, in [22] the authors report that radiology delays are shorter when teleradiology is used [1].

In [23], the authors delineate the main reasons for the medical imaging demand growing in Europe and U.S., explaining why teleradiology is critical to support this development in Europe. In particular, the paper presents three different teleradiology case studies regarding successful data sharing and innovative workflow models through single imaging implementations.

Furthermore, radiology is becoming increasingly specialised in recent years. The studies [24], [25] illustrate the improvement associated with producing a subspecialist report compared to a general one. In this context, teleradiology can be used to gain access to radiologists who are well trained in a particular field. Regarding this aspect, [26] underlines that teleradiology is an important component in radiology field for many aspects, one of which is the possibility to facilitate seeking a second opinion from an expert, improving in this way the efficiency and effectiveness of healthcare system. In particular, the paper presents a model based on second opinion teleradiology consultation service, which involves an Israel health insurance company and a premier medical center in San Francisco, showing in detail the operational procedure of this second opinion service [26].

In [27], the authors state that fast wireless network and mobile technologies allow to evaluate diagnostic images with a comparable accuracy to PACS workstations. In [28], the authors design and develop an economic web based platform for the diagnostic images referral by means of IT technologies. For more details about this approach, [29] gives a complete historical overview of the applications in the teleradiology sector [1].

Moreover, the tele-imaging application represents a significant chance for the future and it is a central practice for telemedicine. In particular, it includes the transmission of medical digital images and has an important role in all fields of telemedicine, such as expertise, consultation, teaching and research activities [30].

Another field of tele-imaging concerns teleconsultation, whose overall goal is to overlook geographical and functional distance between two or more geographically separated healthcare providers, using information and communication technology [31]. Surgery and general medicine are among the most common topics in adult teleconsultations. In particular, surgery teleconsultations are generally performed by sending radiology images (teleradiology), clinical pictures or videos, in order to reduce unnecessary patient transfers or travels with related costs, and also to support prompt decision making with high diagnostic accuracy, improving the future of the healthcare delivery [31].

In this context, teleconsultation is increasingly becoming an integral part of the hospital procedures. A European survey [14] estimates that 65% of all radiologists currently use teleradiology for the images sharing within the organisation and for on-call readings from home [32].

Moreover, some hospitals, especially in the United States, use teleconsultation to manage out-of-hour reporting by adopting a complete *Hub and Spoke* architecture. The architecture involves configurations which are centers of excellence opened 24 hours a day, providing highly complex diagnostic-therapeutic services (Hubs), and other configurations which aim to create a territorial network of services for the management of patients once overcome a certain clinical severity threshold (Spoke) [1], [33].

Additionally, teleconsultation is also applied in emergency or urgent situations, for example during pre-hospital critical care situations [34] or for emergency orthopaedic patients by mobile phone messages [35] or even for ophthalmic emergency consultation, through the use of a smartphone App or PC and a webcam [36].

There are also many studies about telemonitoring application; in particular, an analysis presented in [37] and conducted on 65 studies in the field of telemonitoring for cardiovascular, pulmonary, hypertension and diabetes diseases, describes the maturity and reliability of remote monitoring technologies [1]. Telemonitoring has also demonstrated the ability to identify changes in the condition of chronic patients at an early stage of their disease, obtaining fast intervention and avoiding complications [31]. An example is reported in [38], which refers to diabetic patients affected by a reduction in haemoglobin A1c and a significant blood glucose increase. For patients affected by hypertension, [39] reports how to respond through telemonitoring to a systolic and diastolic blood pressure reduction. In [40], [41], the authors report a high level of satisfaction of telemonitoring processes and [42], [43] emphasize the positive contribution given by telemonitoring to the decrease of hospital admissions for heart and pulmonary diseases.

Through telemonitoring it is also possible to analyse the Continuous Positive Airway Pressure (CPAP) in patients with Obstructive Sleep Apnoea (OSA) [44]. In particular, a study [45] based on [46], [47], shows that CPAP compliance is significantly higher in the Telemonitor (TM) care group compared with the usual care group, thanks to current technological advances and to the use of innovative medical devices which allow patients to control some vital parameters at home.

Telemonitoring is also applicable in the cardiology sector, for example through the development of a phone-based and rule-based expert system for heart failure telemonitoring, designed to provide remote clinical support for patients with cardiological problems [48], [49]. Another important field of application of telemonitoring concerns the control of elderly people [1], [50].

A complete review of telemedicine, is proposed into [51], which contains detailed information about: telecardiology, telestroke, teleradiology, teleoncology, teledermatology, telepathology and telemonitoring.

2.1. Health information systems and standards

An Information System (IS) can be defined as a technical and organizational system designed to collect, process, store, and exchange information and data [52].

In healthcare organizations (i.e. hospitals, radiology departments), IS manages all clinical data, diagnostic images, and reports. HIS (Hospital Information System), RIS (Radiology Information System), PACS (Picture Archiving and Communication System) are some of the most important clinical Information Systems applied in healthcare organizations [53], [54]:

- HIS: system based on IT used in healthcare to manage the administrative and clinical flows of a hospital;
- RIS: system used to manage the radiology data flow, with the purpose to collect, manage and display data and information produced in the Radiology Department;
- PACS: system used for archiving images acquired from modalities applications, managing worklists, and accessing images from reporting stations (workstations) [55].

The RIS is a HIS subsystem, while the PACS, an independent system integrated with the others, deals with the diagnostic images management.

The integration among the different systems and, in particular, between the RIS and the PACS within Radiology Department should lead to: i) use a single medical registry; ii) have a simultaneous display of images and reports; iii) facilitate the automatic distribution of images through departments [56].

Generally, every healthcare organization has a PACS which manages the various modes of digital image acquisition. A structured PACS comprises an image acquisition system and a secure network connecting radiology workstations (on-site or off-site) with a centralized archive [1]. The DICOM standard is used in order to simplify the interoperability of different devices in acquisition, transmission and saving of images.

Digital Imaging and Communications in Medicine (DICOM) was developed by the American College of Radiology and the National Electrical Manufacturers' Association [57], in order to use fully digital images with high resolution instead of physical X-ray films [58].

DICOM is an international standard for all clinical images and relevant metadata and concerns the handling, storing, printing and transmitting phases [59], [60]. This standard allows to manage different types of medical images with other data such as: patient name, reference number, study number, dates and reports [61].

The DICOM standard is widely adopted in radiology, cardiology and radiotherapy imaging, and it is progressively expanding in other medicinal areas, like ophthalmology and dentistry [62].

Health Level Seven (HL7) is another important standard which allows the standardisation of the exchange of electronic health information. It is focused on data communication and not on data storage and presents the following versions: HL7 v2, HL7 v3 and the most recent FHIR (Fast Healthcare Interoperability Resources) which uses real-time RESTful interfaces [1]. In particular, HL7 v2 is chosen from local hospital for the exchange of healthcare information, including electronic medical record information [63]. HL7 v3 is designed to be the successor to v2 and overcomes the v2's shortcomings. Finally, FHIR combines the features of HL7 v2 and v3 with the modern web technologies such as the REST architecture, in order to facilitate its implementation [63], [64]. In particular, differently from v2, FHIR is not based on the exchange of messages when certain events occur but on the availability of data and information exposed through a server, defining in this way the Resources (patient, encounter, condition, observation, procedures, care plan, goal and appointment), which are organized into modules, each representing a different functional area.

Modules are structured in 3 groups and 5 levels: infrastructure (levels 1 and 2), content (levels 3 and 4) and reasoning (level 5) [65].

In order to support the adoption of international standards such as DICOM or HL7, a consortium of professionals and industries named Integrating the Healthcare Enterprise (IHE) has been defined. IHE preserves a list of domains such as cardiology, radiology, pharmacy, IT system and patient care. In this context, it is important to mention the Cross Enterprise Basic eReferral Workflow Definition Profile (XBER-WD), which defines and standardises the workflow related to an eReferral Document, the actors involved and the digital documents related with the process [66]. XBER-WD is a standard based on a document type called Cross-Enterprise Document Workflow (XDW), in which the tasks completed in the workflow, the inputs and the outputs are stored [1].

The XDW standard supports participants in a complex environment to track and manage tasks related to clinical workflows, providing a common interoperability infrastructure to support them. All the documents can be exchanged between organisations through Cross-Enterprise Document Sharing (XDS) repositories, which facilitate the registration, sharing and storage across health enterprises of the patient electronic health records, in order to associate documents containing medical data to a specific workflow [67], [68].

3. HINT project

The HINT (Healthcare INtegration in Telemedicine)[1] project aims to implement a clinical support network both in cases requiring high specialization and to cover areas without specific expertise. The project develops a platform for the management of cases requiring access to diagnostic images, with the possibility of

[1] Intervention co-financed under the POR Puglia FESR-ESF 2014-2020 - Priority axis 1 - Research, technological development, innovation - Innonetwork Call

applying specific medical expertise even when no one is physically present where the images are produced. Specifically, the project envisages the realization of technological solutions aimed at optimizing the communication between the multi-disciplinary teams involved in a clinical case. Research activities also focus on defining tools which improve clinical data collection and healthcare information management, as well as on the clinical workflow efficiency, supporting the Physician in the decision-making process. HINT aims to realise a multi-target approach by integrating Physician support with the suggestion of diagnostic pathways, relying on a computational architecture capable of managing large data sets and algorithms for process mining and deep learning [1].

The research topics addressed by the project concern innovation in the area of:

- Cloud architectures in the clinical health field;
- Clinical Decision Support Systems and Computer Aided Diagnosis systems;
- Deep Learning models and algorithms;
- Integration and management standards for clinical processes;
- Secure data storage.

3.1 The BPM approach and lifecycle applied to the project

Looking at the approaches used in recent years, it is possible to note that there have been health and telemedicine projects which have used Workflow Management Systems based on the Business Process Management (BPM) approach and on the Business Process Model and Notation (BPMN) [69]. The BPM and BPMN tools allow to optimize clinical processes, tracking the key health parameters of an individual patient, organizing and monitoring all the healthcare activities, in order to support health organizations [70], [71].

The business processes are intended as "logically related tasks performed to achieve a defined business outcome" [72]. The BPM approach embodies methods and technologies which enable organizations to analyse, manage and optimize business processes [1], [73].

In [74], the authors analyse three BPM lifecycle proposals: the BPM lifecycle defined by Van der Aalst, the BPM lifecycle proposed by Netjes, and the BPM lifecycle proposed by Weske.

The first author proposes the BPM lifecycle with four steps: i) process design; ii) system configuration; iii) process enactment; iv) diagnosis phase, concerning processes' monitoring and analysis [75].

The second approach includes five steps: i) design; ii) configuration; iii) execution; iv) control; v) diagnosis, which provides information in order to identify opportunities for improvement, workflow bottlenecks and any other critical points [76].

The last approach presents a cycle consisting of the following four steps: i) design and analysis; ii) configuration; iii) enactment; iv) evaluation of the business processes' performance [77].

Taking the cue from the models proposed, the lifecycle of the health processes managed in the project (starting from modelling to optimization) are described and shown in the following revised phases and in Figure 1 [1], [78]:

1. Process analysis and modelling;
2. Process implementation;
3. Process execution;
4. Process monitoring.

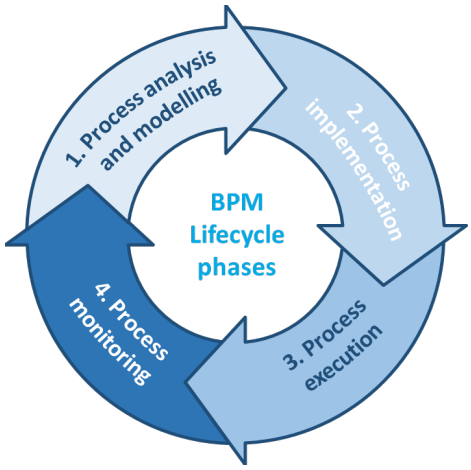


Figure 1: BPM lifecycle phases in HINT project

In the analysis phase, a business problem is highlighted and studied, in order to identify and analyse the related processes. The outcome of the process analysis is a new or updated process model which makes available a complete overview of the organization processes. In the modelling step, the current state of processes and the related activities are designed and modelled using tools and notations as BPMN, DMN (Decision Model and Notation), CMMN (Case Management Model and Notation) [1].

Then, in the implementation phase, the changes required to shift from the As-Is process to the To-Be process are defined and performed [63]. In particular, this phase involves two different aspects: the organizational change management and the process automation. Once the process has been mapped and illustrated, during the execution phase, it is run and tested, allowing its adaptation to the users' needs. Then, a series of logs and measurements are collected and used in the monitoring phase, in order to analyse the process performance through specific Key Performance Indicators (KPI) and targets [1]. The notations mentioned above can be used for the process modelling: BPMN, CMMN and DMN, which can support the clinical process management.

In particular, an application example can be seen in [79], which recreates a guideline for stable ischemic heart disease, and considers the advantages and limitations of representing clinical pathways with these three notations. Moreover, [80] presents a clinical pathway for living donor liver transplantation using a combined approach which involves the event processing (BPMN) and the case management (CMMN) [81]. Another applicative case is presented in [82], which reports a study in healthcare sector, using DMN and BPMN approaches in an integrated manner, analysing a pediatric surgery process in a mid-size hospital. Finally, [83] combines the BPMN and DMN notations for modelling care pathways in a standardised manner.

As previously analysed, the BPM proved its usefulness to improve the design and therefore the telemedicine processes' implementation. Indeed, there are some cases of the BPM application in the field of telemedicine projects, regarding, for example, de-hospitalization services with focus on e-Health

solutions [84], or related to chronic pathologies such as Chronic Obstructive Pulmonary Disease (COPD) with a patient-centered approach based on telemedicine devices [85]. There is also a study concerning the use of the BPM methodology in healthcare with reference to the end-user interaction with EHR interface [86], as well as a study involving patients with hypertension whose healthcare processes are modelled using the BPM paradigm in order to identify weaknesses and mitigate them, for a better patient empowerment [87]. These studies underline that BPM applied to telemedicine is an efficient approach which can improve the management of the all-internal processes, the work of healthcare operators, with a positive impact on the quality of care and life of the patients.

In fact, in recent years, different clinical organisations have applied a BPM-based technology approach in order to improve the process management in healthcare sector.

In this way the quality and efficiency of processes can be increased, ensuring a better and continuous relationship with the patient. The use of BPM can also support medical practices by monitoring care processes, managing and analysing the clinical workflow and specific activities. A structured knowledge of the business context could be useful for the analyst in these activities [88], [89].

Table 1 summarizes and describes the BPM lifecycle phases in telemedicine processes, outlines the possible actors involved (Physicians, Specialists, Developers, Analysts) and identifies the clinical profiles, the standards (i.e. HL7, IHE, DICOM) and notations for process modelling and rule management [1].

3.2 HINT architecture

The HINT project intends to develop a cloud platform with *Hub and Spoke* architecture for telemedicine based on imaging, which, by using international standards (DICOM, HL7, IHE), allows the integration of processes, health data and diagnostic images [1].

The cloud platform developed includes components which support Physicians and healthcare workers in the clinical cases management, enabling them to authenticate and authorise users, to manage roles, to search and consult the patient list, to manage examinations, to consult documents in the patient medical record, to open a clinical case, to record data in the required formats and to initiate and manage predefined diagnosis and treatment paths through the Workflow Manager functions whose interrelations will be highlighted later.

The core components developed, and described below, are: Portal, *Hub and Spoke* applications to manage Telemedicine services, DSS (Decision Support System), Workflow Management System, ESB (Enterprise Service Bus), VA (Virtual Archive) (Figure 2).

These components ensure [1]:

- user access to the different applications;
- the availability of basic record management services;
- access to integration services, allowing messages to be converted into the standard formats used in the health field;
- access to the necessary terminology management services;
- standard access to document archiving services.

Table 1 – BPM lifecycle phases, with related actors and clinical standards applied in the HINT project

BMP Lifecycle	**Description**	**Actors**	**Standards**
1. Business process analysis and modelling	The phase involves the identification and study of the relevant business processes and the As-Is process definition	Business/Process Analyst Clinical domain experts (General Practitioner, Specialist, Healthcare Manager)	BPMN CMMN DMN
2. Business process implementation	The phase involves changes preparation needed to move from the As-Is process to the To-Be process in terms of process automation and management	Software Developer Process Developer	CDA HL7 DICOM XBER-WD (IHE) XDS (IHE) DSUB (IHE)
3. Business process execution	The phase involves To-Be process running and testing, with measurements and logs collection needed in the monitoring phase	Physician and Specialist Users RIS/PACS systems DSS UWB devices	DICOM XDW (IHE) FHIR
4. Business process monitoring	The phase involves the collection of relevant data in order to define the process performance in relation to its KPIs	Clinical Data Analyst Clinical Risk Manager	XDW (IHE) AUDIT LOG

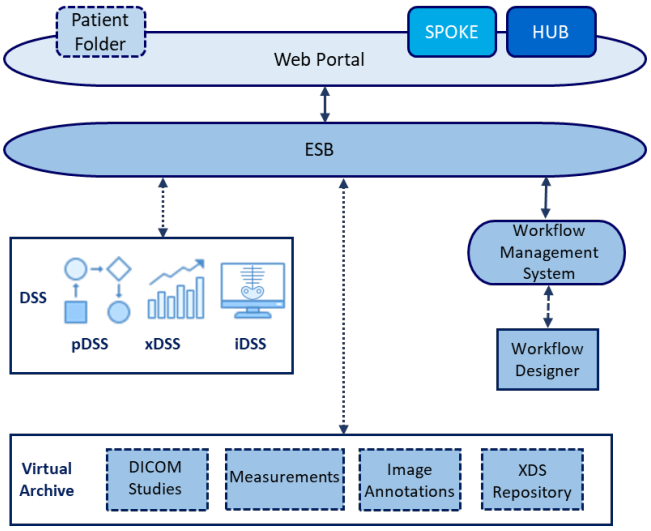


Figure 2 – HINT project architecture components

Portal

The HINT Portal provides the general services of the platform. It consists of a web interface hosting all the user applications (Web Container), such as the Patient Folder, the agenda management, the visit management, etc. and allows access to the specific applications for *Hub and Spoke* telemedicine processes described in the scenarios.
The Portal allows basic consultation to the patient medical record, in which it is possible to examine the patient's documents, the active pathways and the visit management through acceptance and discharge procedures.
The Portal offers a layer of back-end services to the other applications of the platform and has a local archiving component. This layer mainly offers:

- Single Sign-On services (SSO);
- services for the univocal patient data management (Master Patient Index) and for the patients' identification;
- EHR services.

Spoke

The Spoke represents a set of platform applications which enable the implementation of specific functionalities of a particular clinical-health task. Basically, a Spoke application identifies, for a requesting structure, the hardware and software equipment used to digitally record and archive the patient's physiological data, documents and clinical studies, i.e. to make use of that set of information to perform the task and obtain a diagnosis. For HINT, Spoke functionality also includes the generation of requests for examinations and consultations (Neuro-Radiological and Dermatological teleconsultation).
The Spoke hardware and software equipment can be represented by a specific software application and by the following devices: medical devices (including UWB devices - Ultra Wide Band), diagnostic apparatus, RIS/PACS system.

Hub

The Hub represents a set of applications of the platform which allows to receive the notification and proceed to the completion of a particular clinical-health task. Basically, a Hub application identifies, for the supplying structure, the system used to consult data, documents and images and to record the diagnosis in a medical report. The Hub may be the display component of a teleconsultation, telemonitoring or teleassistance system, or even a RIS/PACS system.

ESB

The integration services of the HINT platform enable communication between the different components by sorting and transforming messages into formats supported by the final components.
The Enterprise Service Bus (ESB) component implements services dealing with the exchange of information between the front-end of the Portal together with the user applications (including *Hub and Spoke*) and the back-end components of the platform.
These services implement, when required, the transformation of the messages sent by the clients in JSON (XML format) for data and documents, and in JPEG format, for images, into the health standard formats (HL7, CDA, DICOM).

DSS

The Decision Support System (DSS) component allows the following activities: feature extraction; image classification and diagnostic support by means of a neural network (iDSS); diagnostic logs analysis for the identification of the diagnostic process patterns (pDSS); control of physiological data and alerts reporting (xDSS) [1].

Workflow Management System

The Workflow Management System is responsible for managing the execution of clinical processes modelled in BPMN, which comply with the specifications of the workflow definition profile standard defined by XBER-WD and for the storage of workflow documents in the XDS.
The Workflow Management System consists of these sub-systems:

- Workflow Engine and related services;
- Workflow Document Management and storage system.

The Workflow Engine manages the execution of processes and controls activities involving actors and applications. It receives communications which allow tasks to be updated through services compliant with the FHIR standard. Moreover, it manages the assignment of tasks, the progress of processes and the notifications to external components.

Virtual Archive

The Virtual Archive implements repository services for archiving clinical-health data, documents and images. The VA defines a unique interface layer for the portal and identifies the applications which need to archive data and documents generated during the execution of the processes supported by the platform.

4. HINT telemedicine processes

4.1 Neuro-Radiological Teleconsultation

This process describes the remote consulting activity between Physicians (Requesting Physicians and Clinical Specialists) who, being in two different places, can communicate with each other via ICT systems. In the specific case, the specialized organization (Hub), where the Clinical Specialist provides health services, gives, when required, medical consultations to the local organization (Spoke), which has available clinicians providing only some basic health services.

The principal activity regards the evaluation of the patient's condition for first or second opinion by a Clinical Specialist, based on shared diagnostic images and clinical reports.
The entire process can be divided into the following phases [1]:

- Creation of a Referral Request;
- Consultation Scheduling;
- Clinical data control and evaluation by the Clinical Specialist (creation of report).

The HINT portal can manage the initial step concerning the admission of the patient to the clinical organization, the acquisition of the exam reports (MRI, PET, CT, diagnostic tests) carried out and the subsequent neurological evaluation during the first control visit. If the Requesting Physician of the Spoke Organization needs to perform a neuroradiology consultation by contacting the Hub Clinical Specialist, he/she can use the HINT system based on a workflow engine to create a *Teleconsultation Referral Request*. The Requesting Physician enters all the data regarding the teleconsulting request in the first free Hub.

The data to be entered in order to define the Request are: the patient's personal data, request type (e.g., first neuro-radiological control, diagnostic valuation for second opinion), clinical history (e.g. trauma, potential stroke, dementia), urgency level, the list of past clinical exams (reports and digital images).

The *Scheduling* task creates a list of Hub structures (and related Clinical Specialists) which can take charge of the teleconsultation execution. The Requesting Physician can then select one of them from the list. A decision table based on specific rules (for example Physician specialization, his/her time-sheet and available days and times, intervention mode type) will create the list of potential Hub organizations which can take responsibility for the clinical services. The request is repeated until the referral has been assigned and executed by the available Hub.

When the Hub organization takes charge of the teleconsultation request, the Clinical Specialist performs the consultation and sends the file reference containing the diagnosis and the related report. If the Hub does not take charge of the request within a predefined time interval based on the case urgency, it is sent to another organization. The Clinical Specialist can acquire and display all the patient's data and information regarding his/her clinical case.

After a thorough study of the specific case, the Clinical Specialist provides his/her consultation by first or second opinion. The Spoke application which has started the process receives a notification of final diagnosis, so the related Requesting Physician can analyse the consultation report provided by the colleague [1]. The BPMN model of the teleconsultation process, realised with the open-source Signavio software, is represented in Figure 3 and Figure 4, the latter represents the related sub-process.

4.2 Dermatological Teleconsultation

The periodic check of the moles is fundamental for prevention and early detection of the presence of any skin cancers or melanomas, and it can improve significantly the chances of cure. In fact, early diagnosis is an important factor to treat this type of skin pathologies in time.

Another teleconsultation process analysed and implemented in the project, regards specifically the melanoma pathologies, which is similar to the one described for the neuro-Radiological teleconsultation. In fact, the process steps are based on these phases: Dermatological image acquisition, Request for teleconsultation referral and its Scheduling, in order to find a specific Dermatology Clinical Specialist and the clinical evaluation through artificial neural network with the final report.
In this process, the platform must acquire the patient's data in order to perform analyses through a neural network and to provide predictions referring to the type of skin lesions and related statistics.

Unlike the previous scenario, a DSS is used by the Dermatology Clinical Specialist to support the evaluation of the clinical situation. It is focused on the analysis of images assimilated in previous diagnostic cases of melanoma. The imaging DSS (iDSS) is based on specific images fingerprint extraction (features) and it permits a diagnostic image classification through the neural network [1].

The application can acquire all biometrics data (phenotype, ethnicity), personal data (region and province of residence, age, sex) and dermoscopic images (png, jpeg, jpg or DICOM) of a patient. In this way, it is possible to carry out analyses based on these data through a neural network, to return the predictions referred to the typology of the cutaneous lesions and to characterize any melanomas.

In this clinical scenario, the Spoke Requesting Physician visits the patient and enters in the platform all his/her personal data and past clinical information (exams reports, dermatological images). The Requesting Physician may indicate the diagnosis regarding the specific case, if the clinical situation is clear after the first visit.
If necessary, the Requesting Physician can send a teleconsultation request to the Dermatology Clinical Specialist of a Hub for a first opinion (when the initial diagnosis has not been made and the case is sent to the Dermatology Clinical Specialist immediately) or for a second opinion after the diagnosis. The Dermatology Clinical Specialist who takes over the case can be supported in his/her analyses by a neural network which can:

- request the prediction of specific skin pathology evaluation returned as the most likely disease class;
- check the 5 similar cases retrieved from the dataset and compared with the image linked to the patient.

In this way, the Requesting Physician is supported in detecting the presence of possible melanomas. Thanks to the neural network (DSS), the Dermatology Clinical Specialist has the opportunity to compare the new dermoscopic images acquired regarding the suspected lesions or moles of a patient with those present in the system dataset. A comparison is then made to identify any similarities and consequently to provide the Requesting Physician

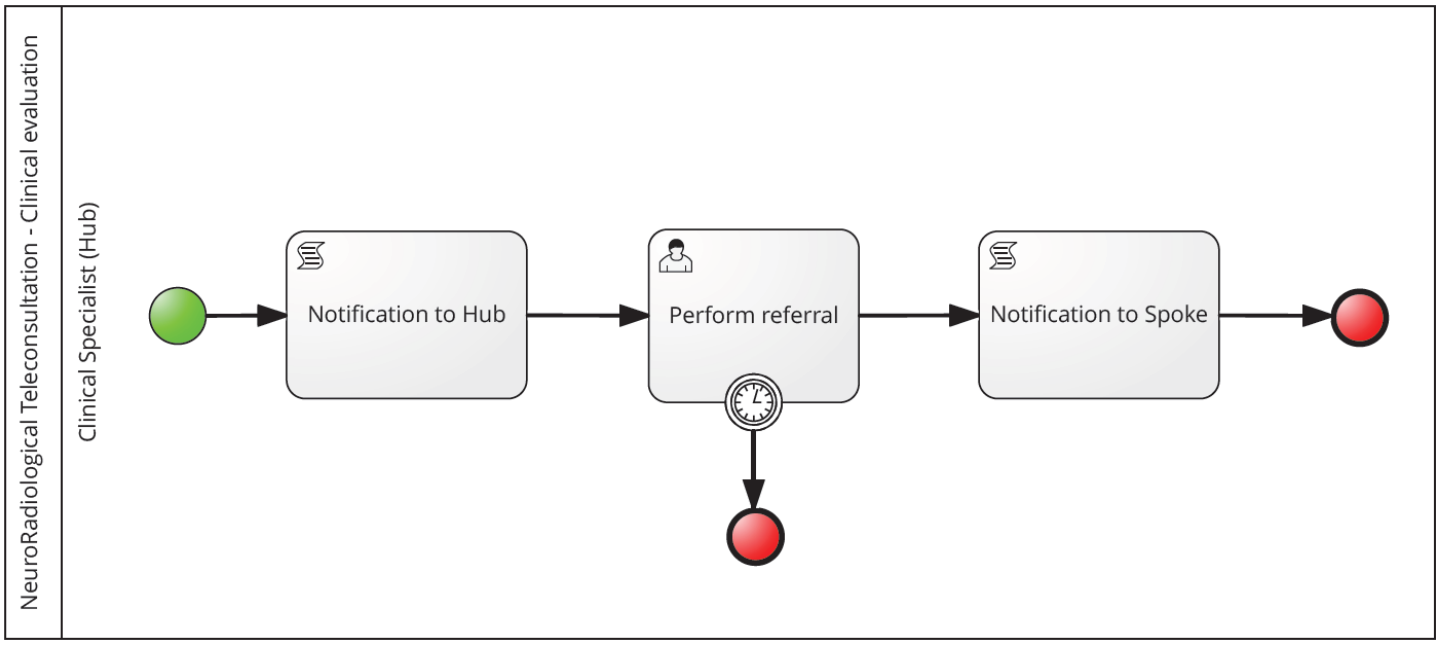


Figure 3 – NeuroRadiological Teleconsultation

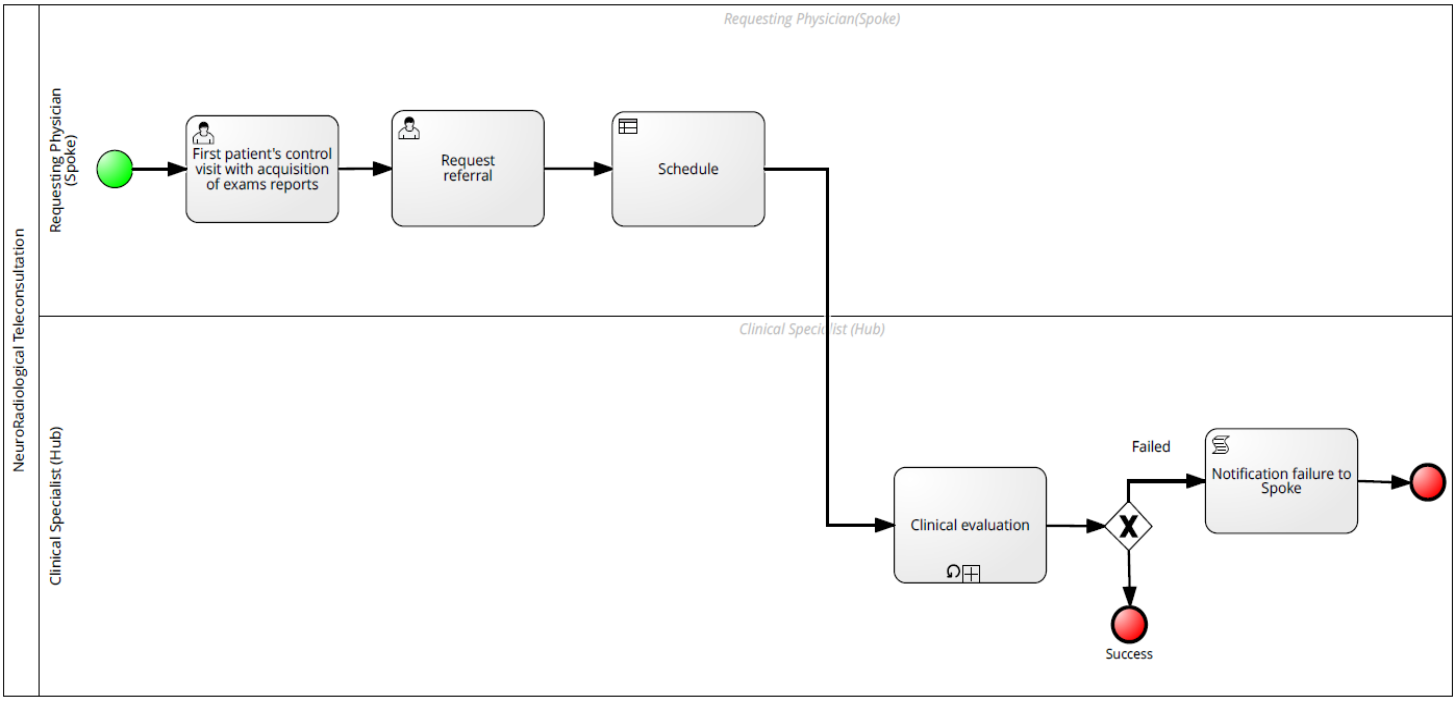


Figure 4: NeuroRadiological Teleconsultation Subprocess: Clinical evaluation

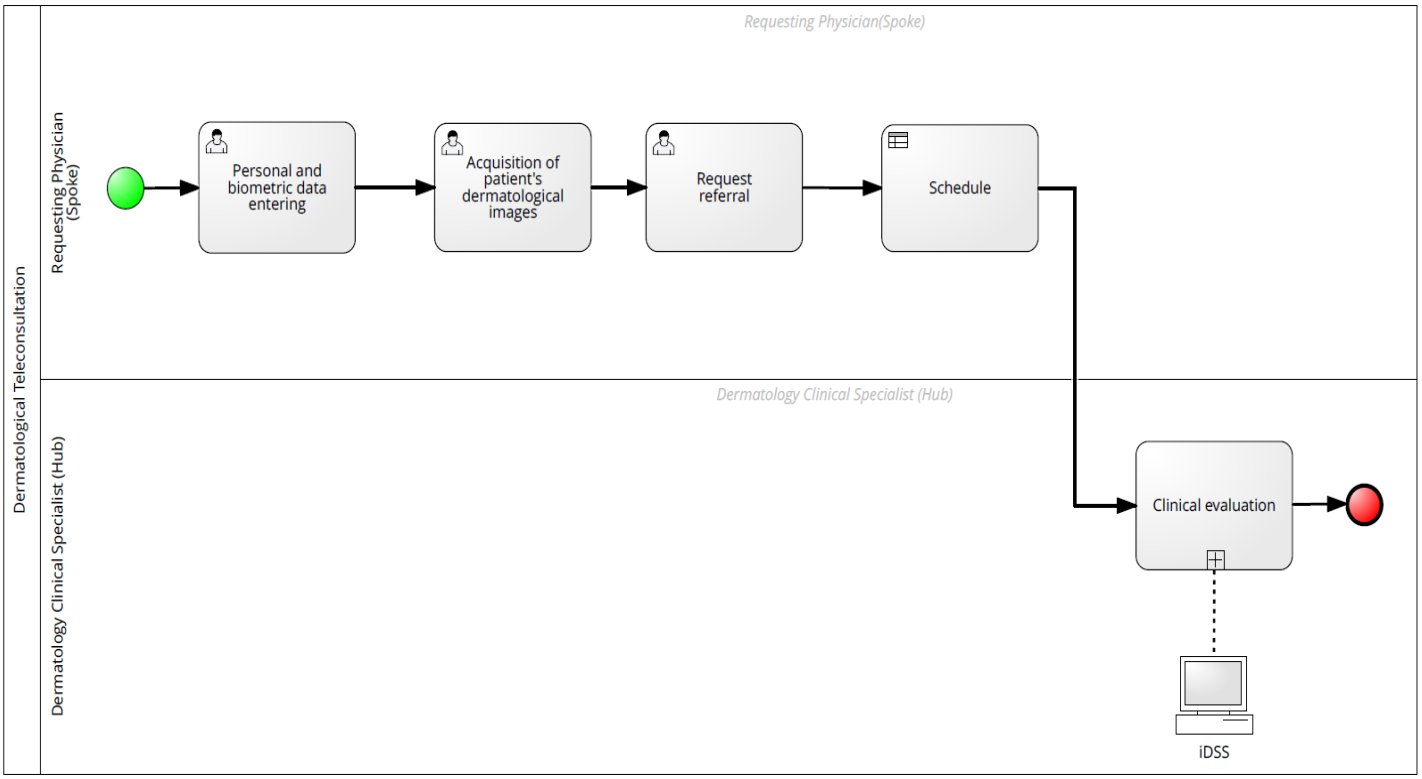


Figure 5: Dermatological Teleconsultation

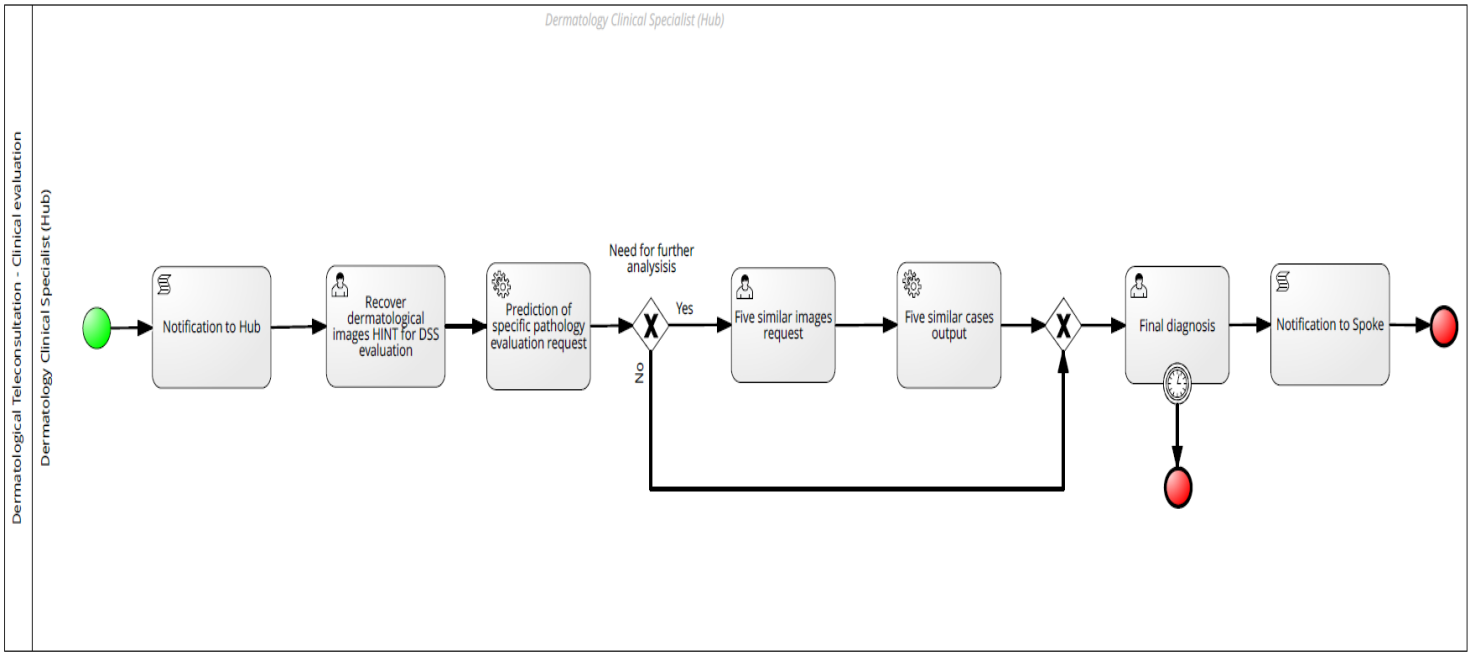


Figure 6: Dermatological Teleconsultation Subprocess: Clinical evaluation

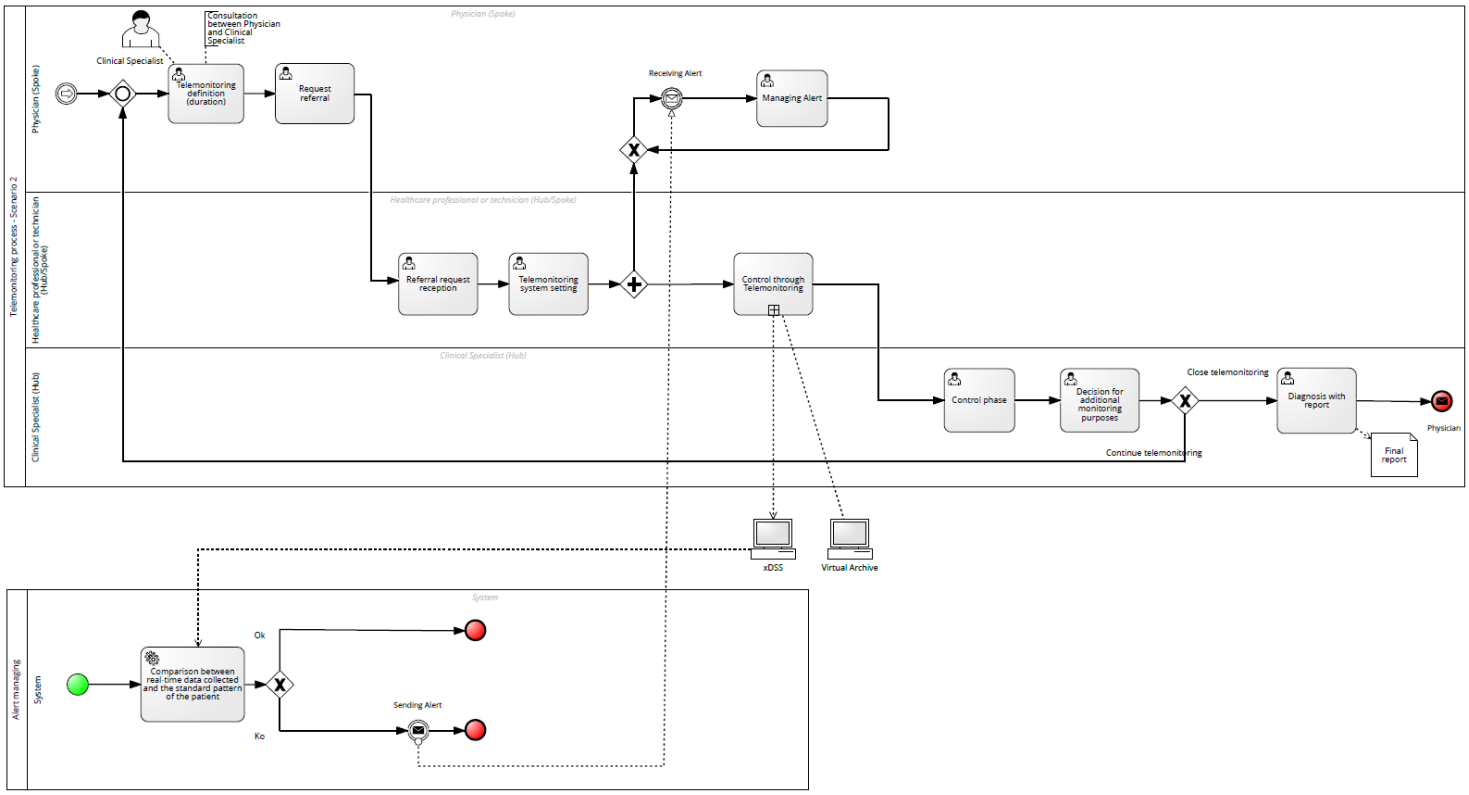


Figure 7: Telemonitoring process - Scenario 1

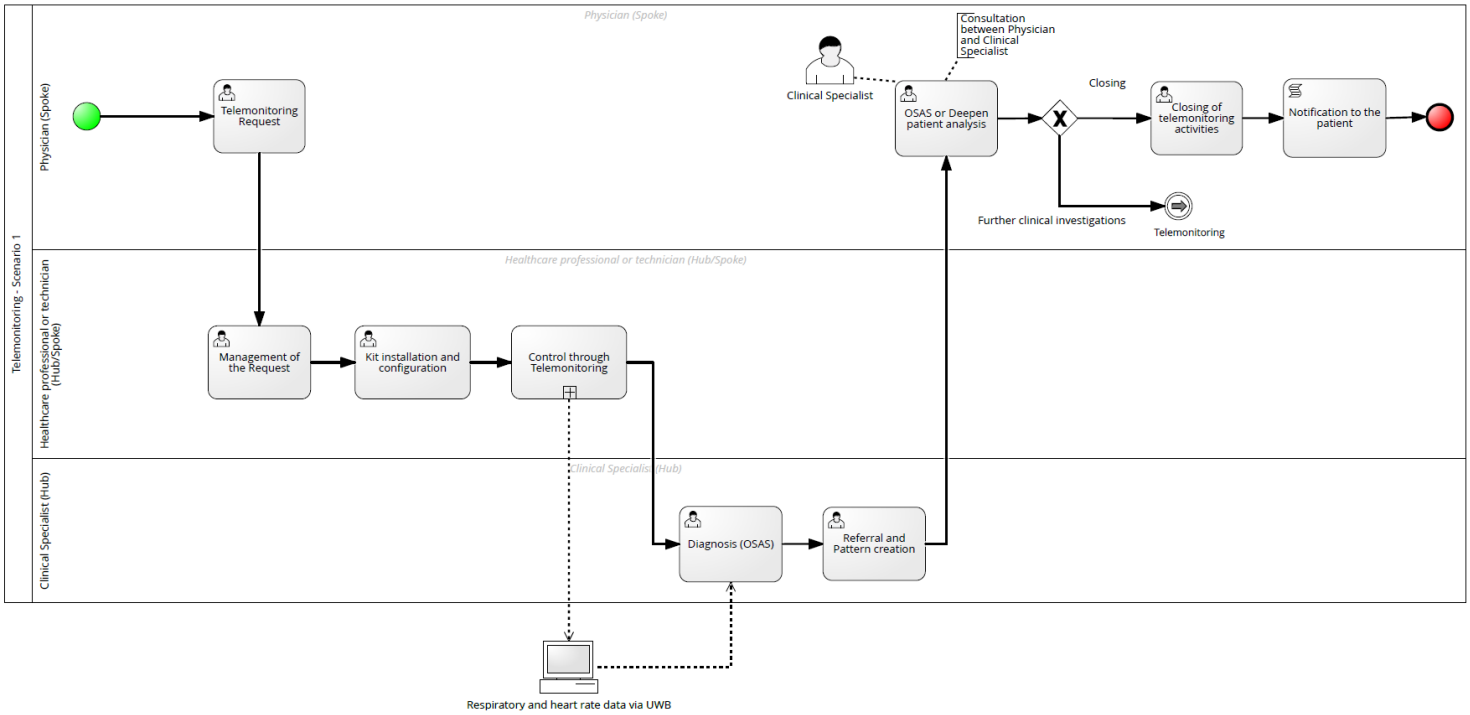


Figure 8 – Telemonitoring process - Scenario 2

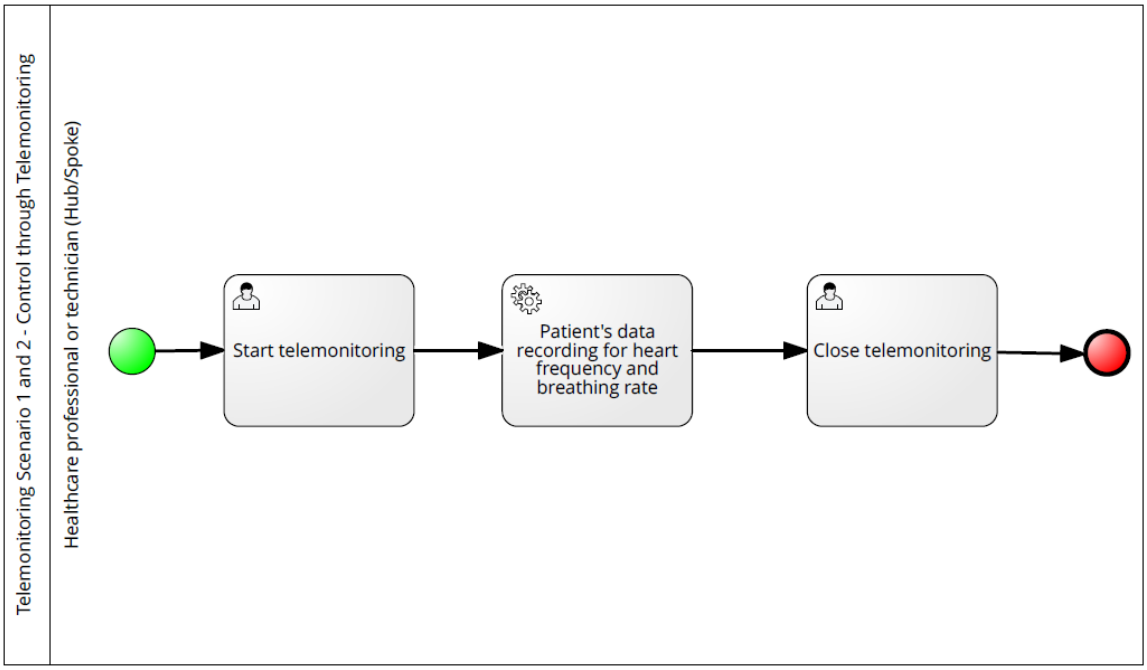


Figure 9 – Telemonitoring process Subprocess: Control through telemonitoring

with specific classes and coefficients of cosine similarity (fingerprints comparison).

The steps of the Dermatology Clinical Specialist for the evaluation of a clinical case via neural network and 5 similar cases are:

- the Dermatology Clinical Specialist selects the clinical case for which he/she wants to obtain a prediction regarding the acquisitions of data and images he/she has made;
- the neural network returns the predicted class with the probabilities of all other similar classes;
- the Dermatology Clinical Specialist can also request the 5 most similar acquisitions;
- the neural network returns the 5 images acquisition having the greatest correspondence with the patient's image, in order to be evaluated on the basis of a coefficient of similarity.

The Requesting Physician may use the results provided by the system to write the Dermatology Clinical Specialist diagnosis regarding the selected clinical case. In addition, statistics can be extracted by analysing biometric and personal data using DSS techniques. Also in this scenario, the Spoke organization which has initiated the process, receives a notification of completed diagnosis made by the Hub organization. The BPMN model of the dermatological teleconsultation process is represented with the open-source Signavio software in Figure 5 and Figure 6, the latter represents the related sub-process.

4.3. Telemonitoring – OSAS Syndrome Diagnosis

This process analyses the telemonitoring of the patient's heart and respiratory rate during sleep time to diagnose the Obstructive Sleep Apnoea Syndrome (OSAS). The system uses a non-invasive Ultra-Wide Band (UWB) technology, thanks to which it is possible to acquire data related to the patient's heart and breathing rate during sleep time in a non-invasive way.

The Physician of a health organization (public or private hospital) forwards the request for sleep monitoring in order to analyse the clinical situation of the patient, while a specific health operator of the Hub/Spoke assigned to the service, responds to this request. The service can also be provided at patient's home by activating a monitoring system based on digital devices. This home system is able to acquire, in contactless mode, clinical data from the patient during sleep. In addition, there is a tablet acting as a gateway, which receives the recorded data from the UWB device and retransmits them to the system [1].

The processes for this telemonitoring case are basically two and both involve the Physicians, the Clinical Specialists, the healthcare operators and the patients as follows:

1. the diagnosis formulation based on sleep data collected in a predefined period;
2. the subsequent real-time sleep telemonitoring finalized to continue the evaluation of the clinical case over time.

Formulation of diagnosis by telemonitoring

The steps of the first process leading to the formulation of the diagnosis by telemonitoring and to the subsequent report definition are specified below:

- Creation of pre-diagnosis sleep telemonitoring request;
- Receiving pre-diagnosis telemonitoring request;
- Configuration and installation of sleep telemonitoring kit;
- Start and end of data collection;
- Sleep monitoring data recovery for the patient;
- Diagnosis and report formulation.

After the request has been sent, the system assigns the task to the most suitable Hub/Spoke operator, who can manage the activity and supervise the process until its conclusion, with the production of the final report by the Hub Clinical Specialist.

The Physician who started the telemonitoring request, can recover the patient's monitoring data in order to evaluate his/her clinical condition. In the first phase regarding the *Pre-diagnosis*, the initial clinical information about the patient's condition and the data collected from the monitoring for a given time interval (time series of hearth and breath rate) are recorded in the local database and indicated as reference model to the patient.

After this monitoring phase, the Clinical Specialist can examine the patient's pattern to diagnose the eventual Obstructive Sleep Apnoea Syndrome and produce the associated report which will be saved in the system. Once the diagnosis has been defined, the Physician, consulting with the Clinical Specialist, can decide to perform a successive telemonitoring phase to record the patient's sleep apnoea episodes for an extended period. If further telemonitoring activity is not necessary, the Physician communicates the final diagnosis to the patient (Figure 7 and Figure 9, the latter represents the related sub-process) [1].

Telemonitoring and real-time evaluation

If the patient is afflicted with OSAS or the Physician needs to further investigate the patient's clinical situation, he/she can decide to set up a telemonitoring process, in order to control the evolution of the medical case in a medium or long period.

In this second scenario (Figure 8 and Figure 9, the latter represents the related sub-process), the activities regarding the telemonitoring and real-time evaluation are defined. They are aimed at detecting any critical situation on the basis of clinical data acquired which show a deviation from normal values (reference pattern) and, consequently, provide indications for a timely intervention and a possible change of care plan.

The steps of this second process are:

- Telemonitoring plan definition;
- Creation of sleep telemonitoring request;
- Receiving of telemonitoring request;
- Telemonitoring kit configuration;
- Start of telemonitoring activities and data collection;
- Data processing from xDSS;
- Management of possible alert sent to the system and their visualization by the Physician;
- Control phase for a possible extension of telemonitoring service at the end of the predefined period;
- Final clinical evaluation by the Specialist.

In this situation, there is an initial phase of telemonitoring programme definition, followed by its subsequent referral to the Hub/Spoke operator, who then configures the telemonitoring kit.

After the telemonitoring process has been defined and started, the real time data collected from the patient's UWB device, with the standard model data collected during the pre-diagnosis monitoring, are given as input to the platform's xDSS. The xDSS allows the analysis of patient data from the activated monitoring and, in case of significant deviations between the data collected in the pre-diagnosis and the data acquired through real-time telemonitoring, an alert message is sent by the system to the deputy Physicians. In this way, the system processes the real-time data and compares them with the patient's standard pattern, promptly notifying the presence of its deviation, such as an ongoing night apnoea episode [1]. The potential alert generated by the platform can, then, be viewed and managed by the Physician. At the end of the process, the Specialist makes an assessment of the OSAS data from the telemonitoring and makes it available and viewable to the requesting Physician.

5. Discussion and Conclusions

The use of telemedicine systems, supported by an efficient territorial health service, can reduce the number of admissions and inappropriate hospitalization, making home monitoring and treatments more effective [90]. This approach can allow a better interaction with Specialists and an accurate control of patients' clinical situation through the use of medical IoT devices.

The HINT project fits perfectly this context, as it has provided a comprehensive and optimized framework based on a flexible and workflow-focused teleconsultation and telemonitoring platform.

The new *Hub and Spoke* architecture proposed in this project allow Physicians (Hub) to find Specialists (Spoke) for a consultation even in remote areas, where medical specialists are not physically present, thanks to innovative teleradiology, teleconsultation and telemonitoring solutions.

An important innovation in the HINT project concerned the introduction of a Decision Support System for dermatological teleconsultation within a cloud telemedicine platform, which allows the classification of dermoscopic images of skin lesions, thanks to Artificial Intelligence and Deep Learning techniques.

Another important innovation, in the field of Obstructive Sleep Apnoea Syndrome, concerned the introduction of advanced monitoring devices.

Moreover, a workflow management component within the HINT platform, based on BPM, allowed the automated management of the clinical processes' activities and related data sharing. The steps of the BPM lifecycle were implemented specifying, for each phase, the healthcare standards used and the actors involved, in order to re-engineer the clinical processes analysed (see Table 1).

The technological and organisational innovations led to the definition of a prototype, whose components and functionalities have been tested but not in a real clinical setting. However, the HINT technological applications have been successfully used in the Italian LHA (Local Health Authority) by some partners after the project conclusion.

Many potential advantages deriving from HINT telemedicine solutions, involve different end-users. In particular, Physicians can be supported in finding the most suitable patient's diagnostic process, through telemonitoring and telehealth appointments.

Furthermore, in emergency situations, telemedicine can represent a guidance for Specialist Consultants (SPs) and for hospital admission, since, thanks to telemonitoring systems combined with possible teleconsultations with other Physicians, it is possible to know the patient's clinical situation before he/she arrives at the hospital. For private or public healthcare organisations, telemedicine systems can give the possibility to ensure an effective coverage of the largest number of clinical cases as well as a service homogeneity offered on the territory, thanks to

the support of integrated platforms enabling advanced care systems. The support in the diagnosis phase through IT systems based on telemedicine services can help the Specialist Physicians, who can work also remotely in a more efficient and effective way [1].

The advantages for patients can concern the availability of more specialist services within the hospital or territory healthcare processes. They can also benefit from the empowerment process ensured by telemedicine services, thanks to which they can be more involved in their diagnosis processes and in healthcare treatments.

Finally, using telemedicine systems, patients can reduce time and costs needed to get efficient medical care.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

The research was supported by the Apulia Region and Italian Ministry of Economic Development within the Innonetwork Call (POR 2014-2020 - FESR Funds, Axis I, Action 1.6), project HINT (Healthcare INtegration in Telemedicine) number 7ZKNZM4 - CUP: B37H17002850007. The authors also acknowledge the contributions of all the project partners to the definition of healthcare scenarios and to the development of the HINT platform.

References

[1] A. Caforio, T. Calogiuri, M. Lazoi, G. Mitrano, R. Paiano, "HINT project: a BPM teleconsultation and telemonitoring platform", in Proceedings of the 10th International Conference on Information Systems and Technologies, Lecce Italy, 1–8, 2020. doi: 10.1145/3447568.3448542.

[2] A.V. Gaddi, F. Cascini, E. Chiarolla, B. Delfrate, S. Forti, M. Marchetti, "Telemonitoraggio e telecontrollo per i cittadini con malattie cardiologiche, respiratorie e diabete", MONITOR, **47**, 21–26, 2022.

[3] J. Wiegel, B. Seppen, M. van der Leeden, M. van der Esch, R. de Vries, W. Bos, "Adherence to Telemonitoring by Electronic Patient-Reported Outcome Measures in Patients with Chronic Diseases: A Systematic Review", IJERPH, **18**(19), 2021, doi: 10.3390/ijerph181910161.

[4] D. Voran, "Telemedicine and beyond", Mo Med, 112(2), 129–135, 2015.

[5] I. Khemapech, W. Sansrimahachai, M. Toahchoodee, "Telemedicine – Meaning, Challenges and Opportunities", Siriraj Med Bull, **71**(3), 246–252, 2019, doi: 10.33192/Smj.2019.38.

[6] J. Adler, C. Yu, M. Datta, "The changing face of radiology: from local practice to global network", Medical Journal of Australia, **190**(1), 20–23, 2009, doi: 10.5694/j.1326-5377.2009.tb02256.x

[7] C.L. Snoswell, M.L. Taylor, T.A. Comans, A. C. Smith, L.C. Gray, L.J. Caffery, "Determining if Telehealth Can Reduce Health System Costs: Scoping Review", J Med Internet Res, **22**(10), 17298, 2020, doi: 10.2196/17298.

[8] E.B. Allely, "Synchronous and asynchronous telemedicine", J Med Syst, **19**(3), 207–212, 1995, doi: 10.1007/BF02257174.

[9] M. Batool, A. Jalal, K. Kim, "Telemonitoring of Daily Activity Using Accelerometer and Gyroscope in Smart Home Environments", J. Electr. Eng. Technol., **15**(6), 2801–2809, 2020, doi: 10.1007/s42835-020-00554-y.

[10] D. Dias, J. Paulo Silva Cunha, "Wearable Health Devices—Vital Sign Monitoring, Systems and Technologies", Sensors, **18**(8), 2414, 2018, doi: 10.3390/s18082414.

[11] A. Rosotti, Informatica medica. Sistemi informativi sanitari e reti di telemedicina. McGraw-Hill Education, 2021.

[12] ACR Data Science Institute, "ACR DSI and Standards Bodies Active in Artificial Intelligence Space", ACR DATA SCIENCE INSTITUTE, American College of Radiology.

[13] A.K. Gupta, A. Garg, M.S. Sandhu, "Diagnostic radiology: advances in imaging technology, Third edition", Medical Publishers, 2019.

[14] R. Nuzzi, P. Marolo, A. Nuzzi, "The Hub-and-Spoke Management of Glaucoma", Front. Neurosci, **14**, 180, 2020, doi: 10.3389/fnins.2020.00180.

[15] T. Chong et al., "The California Telepathology Service: UCLA's Experience in Deploying a Regional Digital Pathology Subspecialty Consultation Network", Journal of Pathology Informatics, **10**(1), 31, 2019, doi: 10.4103/jpi.jpi_22_19.

[16] G. Boland, J.T. Schlakman, J.H. Thrall, "Teleradiology", in PACS, K. J. Dreyer, J.H. Thrall, D.S. Hirschorn, e A. Mehta, A c. di New York: Springer-Verlag, 523–560, 2006, doi: 10.1007/0-387-31070-3_26.

[17] E.R. Ranschaert, F.H.B. Binkhuysen, "European Teleradiology now and in the future: results of an online survey", Insights Imaging, **4**(1), 93–102, 2013, doi: 10.1007/s13244-012-0210-z.

[18] B. Horn, D. Chang, J. Bendelstein, J.C. Hiatt, "Implementation of a Teleradiology System to Improve After-Hours Radiology Services in Kaiser Permanente Southern California", TPJ, **10**(1), 47–50, 2006, doi: 10.7812/TPP/05-119.

[19] R.S. Lewis, J.H. Sunshine, M. Bhargavan, "Radiology Practices' Use of External Off-Hours Teleradiology Services in 2007 and Changes Since 2003", American Journal of Roentgenology, **193**(5), 1333–1339, 2009, doi: 10.2214/AJR.09.2984.

[20] H. Eklöf, E. Radecka, P. Liss, "Teleradiology uppsala–sydney for nighttime emergencies: preliminary experience", Acta Radiol, **48**(8), 851–853, 2007, doi: 10.1080/02841850701422120.

[21] A. Kalyanpur, J. Weinberg, V. Neklesa, J.A. Brink, H.P. Forman, "Emergency radiology coverage: technical and clinical feasibility of an international teleradiology model", Emergency Radiology, **10**(3), 115–118, 2003, doi: 10.1007/s10140-003-0284-5.

[22] A. Martinon et al., "Teleradiology saves times in cases of vital emergencies: A comparative study with on-call radiology in two urban medium-sized French hospitals", European Research in Telemedicine / La Recherche Européenne en Télémédecine, **3**(4), 151–160, 2014, doi: 10.1016/j.eurtel.2014.10.001.

[23] T.C.M. Pechet, G. Girard, B. Walsh, "The value teleradiology represents for Europe: A study of lessons learned in the U.S.", European Journal of Radiology, **73**(1), 36–39, 2010, doi: 10.1016/j.ejrad.2009.10.019.

[24] C. Eakins et al., "Second Opinion Interpretations by Specialty Radiologists at a Pediatric Hospital: Rate of Disagreement and Clinical Implications", American Journal of Roentgenology, **199**(4), 916–920, 2012, doi: 10.2214/AJR.11.7662.

[25] E. Zan, D.M. Yousem, M. Carone, J.S. Lewin, "Second-Opinion Consultations in Neuroradiology", Radiology, **255**(1), 135–141, 2010, doi: 10.1148/radiol.09090831.

[26] H.K. Huang, "Teleradiology technologies and some service models", Computerized Medical Imaging and Graphics, **20**(2), 59–68, 1996, doi: 10.1016/0895-6111(96)00032-8.

[27] G.A. Naqvi, M. Daly, A. Dawood, A. Kurkuri, S. Kutty, "Smart consultation for musculoskeletal trauma: Accuracy of using smart phones for fracture diagnosis", The Surgeon, **12**(1), 32–34, 2014, doi: 10.1016/j.surge.2013.09.001.

[28] F.J. Kammerer, M. Hammon, P.M. Schlechtweg, M. Uder, S.A. Schwab, "A web based cross-platform application for teleconsultation in radiology", J Telemed Telecare, **21**(6), 355–363, 2015, doi: 10.1177/1357633X15575237.

[29] J.H. Thrall, "Teleradiology Part I. History and Clinical Applications", Radiology, **243**(3), 613–617, 2007, doi: 10.1148/radiol.2433070350.

[30] A. Bonnin, "Medical tele-imaging: a good chance for the future", Bull Acad Natl Med, 183(6), 1123–1134, 1999.

[31] K. Deldar, K. Bahaadinbeigy, "Teleconsultation and Clinical Decision Making: a Systematic Review", Acta Inform Med, **24**(4), 286, 2016, doi: 10.5455/aim.2016.24.286-292.

[32] S. Conway, "Teleradiology and telemedicine: They are not equal", Managing Quality, Safety, and Cost in Radiology, and Learning to Wear it Well, 28–29, 2016.

[33] Sanitanova, "Concept Hub & Spoke", Monitor dolore - Sanitanova..

[34] M. Skorning et al., "Teleconsultation in pre-hospital emergency medical services: Real-time telemedical support in a prospective controlled simulation study", Resuscitation, **83**(5), 626–632, 2012, doi: 10.1016/j.resuscitation.2011.10.029.

[35] C. Chandhanayingyong, B. Tangtrakulwanich, T. Kiriratnikom, "Teleconsultation for emergency orthopaedic patients using the multimedia messaging service via mobile phones", J Telemed Telecare, **13**(4), 193–196, 2007, doi: 10.1258/135763307780908049.

[36] H. Bourdon et al., "Teleconsultation in primary ophthalmic emergencies during the COVID-19 lockdown in Paris: Experience with 500 patients in March and April 2020", Journal Français d'Ophtalmologie, **43**(7), 577–585, 2020, doi: 10.1016/j.jfo.2020.05.005.

[37] G. Pare, M. Jaana, C. Sicotte, "Systematic Review of Home Telemonitoring for Chronic Diseases: The Evidence Base", Journal of the American Medical Informatics Association, **14**(3), 269–277, 2007, doi: 10.1197/jamia.M2270.

[38] S. Crossen, N. Glaser, H. Sauers-Ford, S. Chen, V. Tran, J. Marcin, "Home-based video visits for pediatric patients with poorly controlled type 1 diabetes", J Telemed Telecare, **26**(6), 349–355, 2020, doi: 10.1177/1357633X19828173.

[39] K.L. Margolis et al., "Effect of Home Blood Pressure Telemonitoring and Pharmacist Management on Blood Pressure Control: A Cluster Randomized Clinical Trial", JAMA, **310**(1), 46, 2013, doi: 10.1001/jama.2013.6549.

[40] R. Bellazzi et al., "A telemedicine support for diabetes management: the T-IDDM project", Computer Methods and Programs in Biomedicine, **69**(2), 147–161, 2002, doi: 10.1016/S0169-2607(02)00038-X.

[41] B. Morlion, C. Knoop, M. Paiva, M. Estenne, "Internet-based Home Monitoring of Pulmonary Function after Lung Transplantation", Am J Respir Crit Care Med, **165**(5), 694–697, 2002, doi: 10.1164/ajrccm.165.5.2107059

[42] C. Maiolo, E.I. Mohamed, C.M. Fiorani, A. de Lorenzo, "Home telemonitoring for patients with severe respiratory illness: the Italian experience", J Telemed Telecare, **9**(2), 67–71, 2003, doi: 10.1258/135763303321327902.

[43] N.B. Shah, E. Der, C. Ruggerio, P.A. Heidenreich, B.M. Massie, "Prevention of hospitalizations for heart failure with an interactive home monitoring program", American Heart Journal, **135**(3), 373–378, 1998, doi: 10.1016/S0002-8703(98)70310-2.

[44] G. Labarca et al., "Efficacy of continuous positive airway pressure (CPAP) in patients with obstructive sleep apnea (OSA) and resistant hypertension (RH): Systematic review and meta-analysis", Sleep Medicine Reviews, **58**, 101446, 2021, doi: 10.1016/j.smrv.2021.101446.

[45] C. Chen, J. Wang, L. Pang, Y. Wang, G. Ma, W. Liao, "Telemonitor care helps CPAP compliance in patients with obstructive sleep apnea: a systemic review and meta-analysis of randomized controlled trials", Therapeutic Advances in Chronic Disease, **11**, 2020, doi: 10.1177/2040622320901625.

[46] C. Turino et al., "Management of continuous positive airway pressure treatment compliance using telemonitoring in obstructive sleep apnoea", Eur Respir J, **49**(2), 1601128, 2017, doi: 10.1183/13993003.01128-2016.

[47] N. Fox et al., "The Impact of a Telemedicine Monitoring System on Positive Airway Pressure Adherence in Patients with Obstructive Sleep Apnea: A Randomized Controlled Trial", Sleep, **35**(4), 477–481, 2012, doi: 10.5665/sleep.1728.

[48] E. Seto, K.J. Leonard, J.A. Cafazzo, J. Barnsley, C. Masino, H.J. Ross, "Developing healthcare rule-based expert systems: Case study of a heart failure telemonitoring system", International Journal of Medical Informatics, **81**(8), 556–565, 2012, doi: 10.1016/j.ijmedinf.2012.03.001.

[49] S. Wali et al., "Expanding Telemonitoring in a Virtual World: A Case Study of the Expansion of a Heart Failure Telemonitoring Program During the COVID-19 Pandemic", J Med Internet Res, **23**(1), 2021, doi: 10.2196/26165.

[50] World Health Organization, WHO global report on falls prevention in older age. Geneva, Switzerland: World Health Organization, 2008.

[51] H. Eren e J.G. Webster, Telemedicine and electronic medicine. Boca Raton: CRC Press, Taylor & Francis Group, 2016.

[52] S.K. Boell, D. Cecez-Kecmanovic, "What is an Information System?", in 2015 48th Hawaii International Conference on System Sciences, HI, USA, 4959–4968, 2015. doi: 10.1109/HICSS.2015.587.

[53] A.R. Bakker, "HIS, RIS, and PACS", Computerized Medical Imaging and Graphics, **15**(3), 157–160, 1991, doi: 10.1016/0895-6111(91)90004-F.

[54] T. Nakamura, H. Hitsuishi, T. Harada, "Cooperation between the Hospital Information System and the File Maker Pro", in 2007 IEEE/ICME International Conference on Complex Medical Engineering, Beijing, China, 291–294, 2007. doi: 10.1109/ICCME.2007.4381741.

[55] B. Mansoori, K.K. Erhard, J.L. Sunshine, "Picture Archiving and Communication System (PACS) Implementation, Integration & Benefits in an Integrated Health System", Academic Radiology, **19**(2), 229–235, 2012, doi: 10.1016/j.acra.2011.11.009.

[56] S. Boochever, "HIS/RIS/PACS integration: getting to the gold standard", Radiology management, **26**(3), 16–24, 2004.

[57] M. Gupta, N. Singh, K. Shrivastava, P. Mishra, "Significance of digital imaging and communication in medicine in digital imaging", Digit Med, **1**(2), 63, 2015, doi: 10.4103/2226-8561.174769.

[58] M. Hosseini, B.E. Dixon, "Syntactic Interoperability and the Role of Standards", in Health Information Exchange, Elsevier, 123–136, 2016, doi: 10.1016/B978-0-12-803135-3.00008-6.

[59] I.D. McLean, J. Martensen, "Specialized Imaging", in Clinical Imaging, Elsevier, 44–78, 2014, doi: 10.1016/B978-0-323-08495-6.00002-6.

[60] M. Licurse, W. Boonn, Computers in radiology, 26, 2016.

[61] R. Bibb, D. Eggbeer, A. Paterson, Medical modelling: the application of advanced design and rapid prototyping techniques in medicine, Second edition. Amsterdam Boston Cambridge Heidelberg: Elsevier/ Woodhead Publishing, 2015.

[62] M.M. Cruz-Cunha, I.M. Miranda, P. Goncalves, Handbook of research on ICTSs and management systems for improving efficiency in healthcare and social care. Hershey PA: Medical Information Science Reference, 2013.

[63] D. Bender, K. Sartipi, "HL7 FHIR: An Agile and RESTful approach to healthcare information exchange", in Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, Porto, Portugal, 326–331, 2013, doi: 10.1109/CBMS.2013.6627810.

[64] R. Noumeir, "Active Learning of the HL7 Medical Standard", J Digit Imaging, **32**(3), 354–361, 2019, doi: 10.1007/s10278-018-0134-3.

[65] M. Mangia, "Alla scoperta di FHIR", Salute digitale, 2022.

[66] Integrating the Healthcare Enterprise, IHE Patient Care Coordination (PCC) Technical Framework, 1, 2016.

[67] Integrating the Healthcare Enterprise, IT Infrastructure (ITI) Technical Framework, 1, 2022.

[68] A. Sakellarios, "A cloud-based platform for the non-invasive management of coronary artery disease", Enterprise Information Systems, **14**(8), 1102–1123, 2020, doi: 10.1080/17517575.2020.1746975.

[69] P. Szmodics, "Evaluating the organizational knowledge based on business process management", PhD, Corvinus University of Budapest, Budapest, 2018. doi: 10.14267/phd.2018020.

[70] A. Capodieci, G. Del Fiore, L. Mainetti, "Adopting Collaborative Business Processes To Prevent The Loss Of Information In Public Administration Organisations", 2014, doi: 10.5281/ZENODO.1092317.

[71] U. Barchetti, A. Capodieci, A. L. Guido, L. Mainetti, "Information Systems for Knowledge Workers: The Kpeople Enterprise 2.0 Tool", in Web Information Systems Engineering - WISE 2012, X. S. Wang, I. Cruz, A. Delis, e G. Huang, A c. di Berlin, Heidelberg: Springer Berlin Heidelberg, 7651, 804–807, 2012, doi: 10.1007/978-3-642-35063-4_71.

[72] T. Davenport, J. Short, The new Industrial Engineering: Information Technology and Business Process Redesign. 2015.

[73] P. Harmon, Business process change: a guide for business managers and BPM and six sigma professionals, 2nd ed. Amsterdam ; Boston: Elsevier/Morgan Kaufmann Publishers, 2007.

[74] R. Macedo de Morais, S. Kazan, S. Inês Dallavalle de Pádua, A. Lucirton Costa, "An analysis of BPM lifecycles: from a literature review to a framework proposal", Business Process Management Journal, **20**(3), 412–432, 2014, doi: 10.1108/BPMJ-03-2013-0035.

[75] W.M.P. van der Aalst, "Business Process Management Demystified: A Tutorial on Models, Systems and Standards for Workflow Management", in Lectures on Concurrency and Petri Nets, 3098, Springer Berlin Heidelberg, 1–65, 2004, doi: 10.1007/978-3-540-27755-2_1.

[76] M. Netjes, H.A. Reijers, W.M.P. Van der Aalst, "Supporting the BPM life-cycle with FileNet", EMMSAD, 135–146, 2006.

[77] M. Weske, Business process management: concepts, languages, architectures, 2nd ed. Berlin, Springer, 2012.

[78] M. Dumas, M. La Rosa, J. Mendling, H.A. Reijers, Fundamentals of business process management. Berlin: Springer, 2015.

[79] J. Chae, B.H. Park, M. Jones, M. Ward, J. Nebeker, "Converting Clinical Pathways to BPM+ Standards: A Case Study in Stable Ischemic Heart Disease", in 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 453–456, 2020, doi: 10.1109/CBMS49503.2020.00092.

[80] N. Herzberg, K. Kirchner, M. Weske, "Modeling and Monitoring Variability in Hospital Treatments: A Scenario Using CMMN", in Business Process Management Workshops, F. Fournier e J. Mendling, A c. di Cham: Springer International Publishing, **202**, 3–15, 2015, doi: 10.1007/978-3-319-15895-2_1.

[81] C. Combi, B. Oliboni, A. Zardini, F. Zerbato, "A Methodological Framework for the Integrated Design of Decision-Intensive Care Pathways—an Application to the Management of COPD Patients", J Healthc Inform Res, **1**(2), 157–217, 2017, doi: 10.1007/s41666-017-0007-4.

[82] A. Bianchi, M. Mortari, C. Pintavalle, G. Pozzi, "Putting BPMN and DMN to Work: a Pediatric Surgery Case Study", in 2021 IEEE International Conference on Digital Health (ICDH), Chicago, IL, USA, 154–159, 2021, doi: 10.1109/ICDH52753.2021.00028.

[83] I. Essefi, H.B. Rahmouni, M.F. Ladeb, "Integrated privacy decision in BPMN clinical care pathways models using DMN", Procedia Computer Science, **196**, 509–516, 2022, doi: 10.1016/j.procs.2021.12.043.

[84] E. Sulis et al., "Monitoring patients with fragilities in the context of de-hospitalization services: An Ambient Assisted Living Healthcare Framework for e-Health applications", in 2019 IEEE 23rd International

Symposium on Consumer Technologies (ISCT), Ancona, Italy, 216–219, 2019, doi: 10.1109/ISCE.2019.8900989.

[85] M. Szelągowski, J. Berniak-Woźny, C. Lipiński, "BPM Support for Patient-Centred Clinical Pathways in Chronic Diseases", Sensors, **21**(21), 7383, 2021, doi: 10.3390/s21217383.

[86] J. Gomes, F. Portela, M.F. Santos, "Introduction to BPM approach in Healthcare and Case Study of End User Interaction with EHR Interface", Procedia Computer Science, 141, 519–524, 2018, doi: 10.1016/j.procs.2018.10.132.

[87] D. Ruiz-Fernández, D. Marcos-Jorquera, V. Gilart-Iglesias, V. Vives-Boix, J. Ramírez-Navarro, "Empowerment of Patients with Hypertension through BPM, IoT and Remote Sensing", Sensors, **17**(10), 2273, 2017, doi: 10.3390/s17102273.

[88] A. Caione, A.L. Guido, A. Martella, R. Paiano, A. Pandurino, "Knowledge base support for dynamic information system management", Inf Syst E-Bus Manage, **14**(3), 533–576, 2016, doi: 10.1007/s10257-015-0294-3.

[89] R. Paiano, A. Caione, "A Knowledge Base Guided Approach for Process Modeling in Complex Business Domain", in Proceedings of the 11th International Joint Conference on Software Technologies, Lisbon, Portugal, 169–176, 2016, doi: 10.5220/0005974801690176.

[90] N.D. Brunetti, S. Scalvini, G. Molinari, "Innovations in telemedicine for cardiovascular care", Expert Review of Cardiovascular Therapy, **14**(3), 267–280, 2016, doi: 10.1586/14779072.2016.1140572.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 5, 91-101 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Self-Adaptive Routing Algorithm for Real-Time Video Transmission in VANETs

Marzouk Hassan*, Abdelmajid Badri, Aicha Sahel

Hassan II University, Electrical Engineering Department, Faculty of Science and Technology - Mohammedia City, Morocco

ARTICLE INFO

Article history:
Received: 03 July, 2022
Accepted: 16 September, 2022
Online: 10 October, 2022

Keywords:
Routing protocols
Self-adaptative
Heuristic algorithm
Video streaming
VANETs
NS2

ABSTRACT

Given the strict Quality of Experience (QoE) and Quality of Service (QoS) criteria for video transmission, such as delivery ratio, transmission delay, and mean opinion Score (MOS), video streaming is one of the hardest challenges in Vehicular Ad-Hoc Networks (VANETs). Additionally, VANET attributes, including environmental impediments, fluctuating vehicle density, and highly dynamic topology, have an impact on video streaming. Creating efficient visual communications in these networks will give drivers access to business and entertainment applications, greater support, safer navigation, and better traffic management. This work particularly investigates the mobility characteristics of autos and the causes of link failure in order to precisely anticipate the dependability of connections between vehicles and build a reliable routing service protocol to satisfy various QoS application demands. Then, a link-time duration model is suggested. Link dependability is assessed and taken into consideration when creating a new self-adaptative routing protocol for video transmission. Due to the quick changes in topology, finding and maintaining the optimum end-to-end route is quite challenging. However, the heuristic Q-Learning algorithm could continually alter the routing path via interactions with the surrounding. This study suggests an algorithm for reliable self-adaptive routing (RSAR). Through changing the heuristic function and integrating the reliability parameter, RSAR works well with VANET. The Network Simulator NS-2 is used to illustrate how well RSAR performs. According to the findings, RSAR is particularly beneficial for many VANET applications since it efficiently addresses the issues brought on by changes in topology.

1. Introduction

Vehicular Ad-hoc Networks (VANETs) are the most important element in building an intelligent transportation system and intelligent city as communication technology advances. Because VANET is a development of conventional Mobile Ad-hoc Networks MANETs, it has sparked a great deal of interest from research organizations, governments, and automakers to offer services like entertainment, consulting, information, and traffic warning; these sectors already have a lot of projects that are already started [1, 2]. According to WHO figures, 1.24 million people die globally each year as a consequence of road accidents. According to World Bank data, transportation accidents cost the global economy $500 billion a year [3]. Solving traffic congestion has become a global issue since there are more vehicles in cities. Time delays, excessive fuel usage, and environmental pollution can all be brought on by traffic congestion. In-car entertainment and vehicle-to-vehicle communication have also evolved into necessities that automakers now offer to customers in tandem with the social vehicle network [4]. Vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) wireless communication are the two categories within VANETs (V2V). Many academics prefer V2V because communication is infrastructure-free, and its implementation is adaptable. VANET, on the other hand, differs from conventional MANETs in a few ways [5]. First, the network topology will frequently change in VANETs because of the fast movement of the network nodes (cars). The links are unreliable due to the short link times maintained between nodes, and vehicle distribution dramatically affects network performance. Second, car nodes are only allowed on roadways and are influenced by a variety of variables, including speed limits and traffic lights. Third, the energy constraints of VANET are no longer a significant problem because cars can generate sufficient power and

*Corresponding Author: Marzouk Hassan, marzouk.hsn@gmail.com

computational capability for themselves. The creation of a reliable routing protocol is now a crucial step in putting VANET applications into practice as an important support technology for the realization of an intelligent transportation system. Old-fashioned routing techniques based on VANETs, such as Dynamic Source Routing (DSR), and Ad-hoc On-demand Distance Vector (AODV) [6, 7], are challenging to apply because of the frequent topology changes and poor connectivity. Since they do not take into account quick topology changes, both the transmission delay and delivery ratio are decreased for these techniques. Recently, various routing algorithms on the basis of geographic location, such as GyTAR and TFOR [8, 9], have been published [10] to address this problem. These routing techniques employ traffic flow estimates, ignore changes in topology, use GPS to find the destination node, and do not take link dependability between nodes into account. Several trustworthy routing algorithms, including EG-RAODV and SLBF [11, 12], have been presented. These algorithms were created by including characteristics like packet error rate and network dependability between nodes. Although the time delay cannot be guaranteed, this method can increase the packet delivery ratio to some amount. Rapid velocity changes are the primary cause of topology changes and faulty links in VANETs. Good evaluation of connection dependability and effective routing algorithm designs can only result from a thorough understanding of vehicle motion characteristics. Assuming that cars are traveling along a set path, a number of variables, including the two vehicles' directions of travel, their distance from one another, their speeds, and their acceleration, can cause a link to be severed between two nodes. Given the low-density, high-speed environment, these aspects must be completely considered in evaluating connection reliability between nodes properly. When maintaining traffic safety and assessing link reliability, the space between vehicles is crucial. Many models of vehicle separation, including the log-normal distribution, the normal distribution model, and the exponential distribution model, have been put forth. In [13], the authors used tests to demonstrate that when the distance between vehicles follows a log-normal distribution, It may be more in line with the way traffic really moves and with safe separations. Effective evaluation of link dependability between nodes requires a thorough grasp of the moving characteristics of moving objects and the traffic flow model. Routing algorithms have been the subject of continuing research, and VANETs have used several efficient algorithms [14] and have outperformed conventional routing algorithms. Q-Learning [15] is a self-learning algorithm that may continuously interact with its surroundings to find the shortest route from a source node to a destination node. This research improves upon the Q-Learning method and suggests a new technique for adaptive routing that is both efficient and trustworthy. It may adaptively modify the Q-table while guaranteeing the dependability of each hop connection, allowing it to conform to changing network architecture like that of a VANET. This paper's major goal is to provide a trustworthy and flexible routing algorithm. The links between each hop determine how reliable a link as a whole is. By carefully examining the motion characteristics of cars, this work creates a trustworthy model of a link between nodes. Once the probability of link dependability has been determined, the Q-Learning algorithm can use it as a parameter to create the RSAR algorithm.

In the first section, the routing method is categorized and then partially explored in the second section. In the third part, we discuss how to create an accurate model for evaluating reliable links. The fourth and fifth section consecutively addresses The RSAR algorithm's fundamental idea and the simulations and analysis that were done. A summary and recommendations for further work are given in the sixth and last sections.

2. Proposed solution

2.1. Q-learning algorithm in VANETs

Standard Q-Learning is an agent-based heuristic learning technique. The three-tuple R, S, and A make up the majority of the learning process for the agent in the widely used Q-Learning algorithm. Where R denotes the instant reward for an action, A ={a1,a2,⋯,an } the activity space, and S ={S1,S2,⋯,Sn } denotes the state space. The larger the reward gained for the action, the nearer the agent is to the target. The learning process will be thoroughly explained in the following subsections once a few related definitions are presented.

2.1.1. Definitions

- Fundamental elements.

The whole VANET environment is used by the agent as its learning environment.

Learning Agent: Each node of the vehicle functions as its own independent learning agent;

State space (S): All nodes other than this agent make up the state space of a specific agent;

Activity space (A): An activity is the transmission of a beacon packet between two vehicles;

Immediate Reward (R): The instant reward that is given to an agent once they have successfully completed an activity.

- Value of a reward

The instantaneous reward, which has a range of [0,1] and is the reward received by an agent after performing an action, is defined in Definition 1. The reward value for the destination node is 1, as it can directly contact the destination node. Formula (1) specifies the beginning value R of the whole network as follows:

$$R = \begin{cases} 1 & s \in N_d \\ 0 & \text{else} \end{cases} \quad (1)$$

N_d Stands for the set of destination node D's one-hop neighbor's node; the reward value of an action is 1. The Q-value $Q(s, a)(s \in S, a \in A)$, which is in the range [0,1], stands for the potential reward value that is earned while moving from one learning stage.

Q-table definition: Every learner maintains a two-dimensional table with the Q-values of the nodes it can reach and its immediate neighbors. The Q-table is a two-dimensional table (see Table 1).

Table 1: Q-table

	D1	D2	…
N1	Q(D1,N1)	Q(D2,N1)	…
N2	Q(D1,N2)	Q(D2,N2)	…
…	…	…	…

The IDs of every potential destination node are written as D_i In the first line of the Q-table. The one-hop neighbor nodes' IDs, denoted as N_i, are listed in the first column. The value of $N_i.Q(D_1, N_1)$ represents the Q-value between the sending node and its neighbor N1 at the receiving node D_1. As can be seen in Table 1, the size of the Q-table is dependent on both the number of destination nodes and the number of neighbor nodes. One can see that it can be climbed with relative ease.

Beacon packets are periodically exchanged between nodes to update the values in the Q-table. Each node receives a share of the learning task, which causes the algorithm to swiftly find the best path and allows for prompt adaptation to network topology changes.

2.1.2. *Learning process*

The aforementioned definitions established each vehicle node as a learning agent. An individual Q-table is associated with each vehicle node, in contrast to conventional learning methods. Nodes complete their learning process by sharing beacon data and updating their Q-tables.

Along with its own speed, position, and other information, each node additionally includes the maximum Q-value of the nearby nodes to the destination node, or the greatest value in a column, in the beacon packet it transmits (as shown in Table 2). Given a VANET topology, $= \{V, E\}$, as shown in Figure 2, we assume that the state space of node A is the series of all nodes that don't include A. The collection of vehicle nodes is represented by $V = \{A, B, C, \cdots, H\}$. The set E of edges characterizes a collection of connected nodes that may exchange messages with each other with a single additional hop. As depicted in Figure 3, let's pretend that from node A, we go to node G as the endpoint. The goal now is to use a self-learning algorithm to determine the best route from node A to G, where A is the starting point.

Each vehicle node agent is given learning tasks, and as part of the learning process, the agent's Q-table and the state activity Q-values, $Q(s, a)(s \in S, a \in A)$, are primarily updated.

Table 2: Network parameter settings

Parameter	Value
Size of topology (m*m)	3511m*3009m
Video file frame size	30 fps
Image Resolution	352 * 288
CBR packet size (bytes)	512
Simulation time (s)	300
Propagation model	Two-ray ground
Transmission range (m)	250
MAC standard	IEEE 802.11 MAC (2 Mbps)

Eq. (2) contains the usual Q-Learning function:

$$Q_s(d, x) \leftarrow Q_s(d, x) + \{R + \gamma \cdot \max_{y \in N_x} Q_x(d, y)\} \quad (2)$$

R is the reward value, D represents the destination node, N_x is x^s 's neighbor node, x is its neighbor node, S is the node to go from node x to node D, the maximum Q-value between x and its neighbor is $\max_{y \in N_x} Q_x(d, y)$.

The reward a node gets for completing an action in accordance with Eq is impacted by the discount factor γ, making it a significant parameter (2). Since link stability is an essential factor, we used the between-node link reliability $r(l)$, calculated using Eq. (15) as a discount factor; thus, r = r. (l). The pace of packet transmission in VANETs is a critical metric that depends on the available bandwidth.

The definition of the bandwidth BW is as follows:

$$BW(bps) = \frac{n \times S_B \times 8}{T} \quad (3)$$

where S_B Is the packet size in bytes, T is the time interval, and packets sent and received by the node are denoted as n. With the assumption that the node's maximum bandwidth, denoted by $\max_{BW}$, is a constant Here's how you can figure out your bandwidth:

$$BF = \frac{\max_{BW} - \mathrm{BW}}{\max_{BW}}. \quad (4)$$

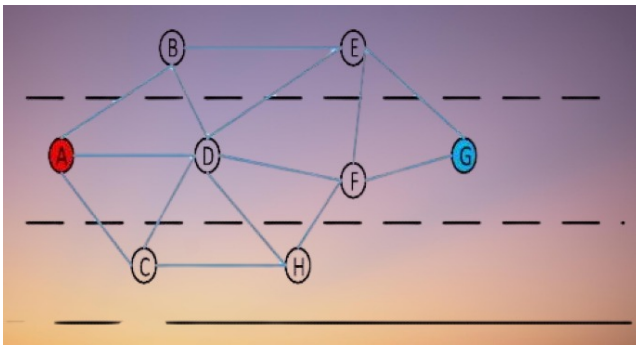


Figure 1: Learning process

Each vehicle node's learning progress is determined by the bandwidth factor. It fluctuates as the element that affects learning speed, depending on variations in effective bandwidth. A new heuristic function can be produced by changing equation (2):

$$Q_s(d, x) \leftarrow (1 - BF) \cdot Q_s(d, x) + BF \cdot \{R + \gamma \cdot \max_{y \in N_x} Q_x(d, y)\}. \quad (5)$$

According to Eq. (5), as the hops number rises, the reward value drops. The final reward value depends on the bandwidth, the dependability of the network, and the number of hops. In a dynamic network, By combining bandwidth and connection state,

the most efficient path may be determined between the source and the destination node. The destination node G's one-hop neighbors in Figure 1 are nodes F and E. According to Eq (5), $Q_E(G,G)$ and $Q_F(G,G)$ are the appropriate representations for the reward values from nodes E & F to the destination node G.

It is possible to determine that the ultimate Q-values of $Q_E(G,G)$ and $Q_F(G,G)$ are 0.7 and 0.8, respectively, after accounting for the impacts of bandwidth and link quality. A, B, C, E, F, and H are D's nearby nodes. When node D receives a beacon packet from any neighbor, it processes the packets and extracts the maximum Q-value to node G by using F, $\max_{y\in N_F} Q_F(G,Y)$. Calculate the relevant Q-value, or $Q_D(G,F)$ In accordance with Eq. (5) and update the Q-table.

Being the largest, $Q_F(G,G)$ will have a value of 1 within the beacon packet. Data packets from other nodes will go through the same processes, after which a selected column in the Q-table will be updated. As shown in Figure 2, the remaining neighbor nodes' Q-tables are changed such that $Q_D(G,F) = 0.5$ in compliance with Eq (5). The maximum Q-value between node D and its neighbors is updated in real-time by node D's Q-table as neighbor data packets are constantly received. In a similar vein, when node D transmits a beacon packet, it looks for the maximum Q-value among itself and its neighbors in a certain column of the Q-table. As soon as node A receives the beacon packet from node D, it will send the highest possible Q-value and continue the calculation to update QA. The same procedure is employed when additional nodes send in beacon packets. The outcomes depicted in Figure 4 will ultimately be attained by continuous data packet exchange [16,17].

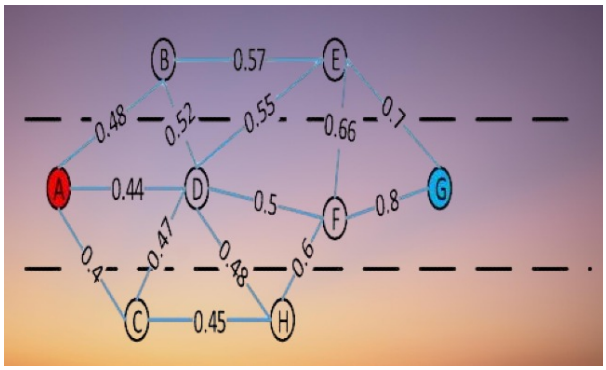


Figure 2: Q-values to the destination node G saved by each node

The best route from A to G may be easily identified, as shown in Figure 2. The best route is the one with the highest Q-value of its nodes. As seen in Figure 4, the best path is ABEG since it has the highest Q-value. The dynamic update-and-save of the Q-table allows the method to quickly, consistently, and robustly react to topology changes.

2.2. The proposed RSAR algorithm

2.2.1. Transmission process

The three actions listed below are carried out by the RSAR algorithm when it routes data packets:

Step 1. Node A checks its internal Q-table before sending a data packet to Node B to see whether Node A is the next hop for Node B. If not, it begins the route development process, which is covered in the next subsection; when this is the case, it picks the node nearest to it that has the greatest Q-value.

Step 2. At the point when the route has been constructed, the fundamental path from the starting node to the ending node has been determined, and the necessary vehicle nodes have been learned. We initiate the route maintenance operation to dynamically keep the end-to-end connection up and running, which is necessary for finding the best path across the whole network topology and resolving the segmentation problem. The next subsection details the route maintenance procedure.

Step 3. The aforementioned procedures establish the best route for the overall network topology. A vehicle node will carry out the first step if it receives or sends data packets; If not, it will proceed to step 2.

2.2.2. Route development and Maintenance Process

Before sending data, a node checks its internal Q-table to determine whether there is a next-hop node between it and the destination. The data packet is sent to the neighboring node with the highest Q-value relative to the target node. If there isn't one, route discovery is initiated. A path request timer is started at the source node, which subsequently sends out a $R_$req packet to the entire network. Within the $R_$req data packet, the source node keeps a record of the IDs of all of the nodes that the data packet traveled through while it was being routed across the network. The $R_$rep packet is saved by the destination node once it has received it for the first time from the source node. Subsequent $R_$rep packets received by the destination node are deleted. After extracting the ID information that it recorded from the $R_$rep data packet and flipping it, the node at the end of the route creates a data packet labeled $R_$ req and stores the reversed path information there. Finally, the $R_$req data packet is discarded. After waiting for an open time window, it sends the $R_$req data packet via the recorded reverse route. Upon receipt, a node that has been previously recorded will make the necessary adjustments to the data packet's next hop address, refresh its Q-table, and then broadcast the packet using a single hop. When the packet is received by other nodes that are not the destination, those nodes will just alter their Q-tables and then discard the packet. The source node will suspend the request timer as it works on updating its Q-table while waiting for $R_$ rep to be transmitted to the destination node, it also serves as the $R_$req source node. At this step, a route from the source to the destination node has been identified, and the Q-tables of the nodes on the first path are modified; their starting values are 0.

The values of the Q-tables of some of the nodes that are adjacent to the initial route path are also updated when that path is constructed. It is essential to get the process of route maintenance going as soon as possible in order to guarantee that the path will continue to be useful despite the dynamic changes in the network [18,19]. The maintenance of the Q-tables in a dynamic manner and the resolution of the issue of network segmentation are the primary goals of the process of route maintenance. Each node will, at regular intervals, broadcast

beacon packets in order to keep the Q-tables of its neighboring nodes up to date. The beacon data packet will primarily provide the node's position, speed, and maximum Q-to-Value ratio. max(Q - Value) was discussed in conjunction with the method of education. In the experiment, the beacon packet's transmission delay was set to a random value between [0.5, 1], which was done to guarantee that the update would be effective. For the sake of the Q-table, the actual time at the destination node has been supplied. When a destination node's time hasn't been changed for more than the specified length of time, the data column associated with it is deleted, and the destination node is deemed invalid. In the event of a network split due to a moving vehicle, RSAR initiates the route request timer and makes use of the store-and-forward method to ensure that an *R*-req data packet is transmitted. In the event that the sending node sometimes doesn't receive the *R*-rep data packet from the receiving node before the timeout expires, the receiving node is considered inaccessible, and the transmission is aborted. Failure to do so will result in the routing route being restored at the point of disruption.

3. Experimental tests

3.1. System Model

As stated in [20, 21], it is reasonable to assume that the highway is often a straight road and that the broadcasting distance is much greater than the width of the road, both of which are necessary for an accurate evaluation of the connection quality between nodes. The road's width may also be assumed to have negligible effects on forwarding node selection at the following hop and, therefore, may be disregarded. This is essential in order to accurately gauge the quality of inter-node connections. As shown in Figure 3, the simulated highway was created using the above methods.

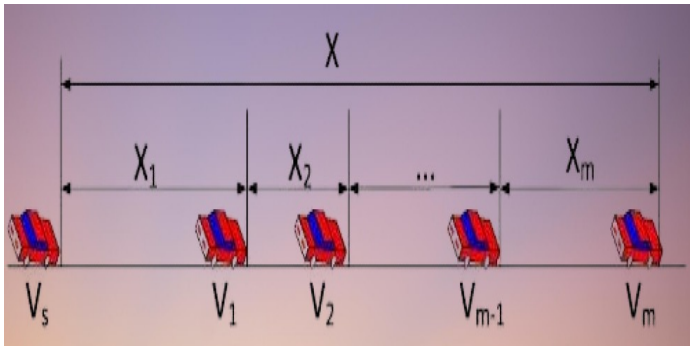


Figure 3: Road model

It is also reasonable to presume that other maneuvers, like speeding up and slowing down, switching lanes, passing other vehicles, and so on, are taking place on the roads. In addition, the distance that separates cars follows a log-normal distribution [22, 23], which can be written as $X_i \in \log N(\mu_i, \delta_i)$ A random variable with a log-normal distribution is depicted in Figure 1 as $X_i = \{X_i(m), m = 0,1,2, \cdots\}$, where m can take the values 0, 1, or 2. X_i denotes the distance between vehicles i and $i+1$ and $X_i(m)$ denotes a random variable at moment m, representing the distance between node I and its related nodes. If the node V_s in Figure 1 is considered to be the reference node, then the value X reflects the distance between V_s and any other node. Since X denotes the distance between other nodes and V_s, and $X = \sum_{i=1}^{m} X_i$X is also log-normally distributed.

Proposition 1: Let's assume that $X \in \log N(\mu, \delta)$ and that the variable $T = \sqrt{aX+b} + c$ is log-normally distributed, where $a, b\ c \in R, a, b, c \neq 0$, and $aX + b \geq 0$.

If the probability distribution function of the t is G_T Then, for each positive $t, G_T(t) = \Pr[\{T \leq t\}]$. Apparently, as T is continuous,

$$\begin{aligned} G_T(t) &= \Pr[\{T \leq t\}] \\ &= \Pr[\{\sqrt{aX+b} + c \leq t\}] \\ &= \Pr[\{aX \leq (t-c)^2 - b\}] \\ &= \begin{cases} F_X\left(\frac{(t-c)^2-b}{a}\right) & \text{when } a < 0 \\ 1 - F_X\left(\frac{(t-c)^2-b}{a}\right) & \text{when } a > 0 \end{cases} \end{aligned} \quad (6)$$

F_X is the probability distribution function of X. When a>0, it is evident that T follows a log-normal distribution. When $a < 0$, according to [13], $F(z) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{z-\mu_z}{\sigma_z\sqrt{2}}\right)$ And letting $z = ((t-c)^2 - b/a)$

$$\begin{aligned} 1 - F_X(z) &= \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\ln z - \mu(X)}{\sigma(X)\sqrt{2}}\right) \\ &= F_Y\left(\frac{a}{(t-c)^2-b}\right) \end{aligned} \quad (7)$$

Y is a log-normal random variable with parameters $-\mu(X)$ and $\sigma(X)$. Since the fact $-\text{erf}(x) = \text{erf}(-x)$ is used. Consequently, T conforms to a log-normal distribution, concluding the proof.

The direction of motion is indicated by the black arrows in Figure 2, where two cars are shown in "a" going in the same direction, whereas two vehicles are depicted in "b" moving in the opposite way.

Proposition 2: Assuming that $X \in \log N(\mu, \sigma)$, $T = aX + b$ is log-normally distributed, where $a, b, c \in R$ and a, $b, c \neq 0$. (The reasons offered in the proof of 1 may be used to simply demonstrate this)

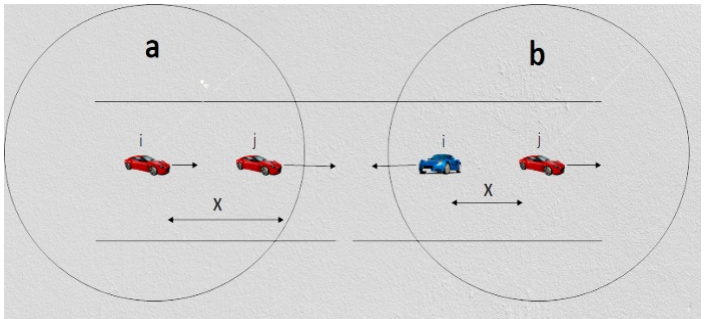


Figure 4: Two cases of link breakage

There are two basic circumstances (shown in Figure 4) in which the connection between two nodes breaks if we assume that the cars are traveling along a fixed route. Both of the vehicles depicted in Figure 4a are moving in the same direction, whereas

both of the vehicles depicted in Figure 4b are progressing in opposing directions. Think of vehicle i as the transfer point and vehicle j as the reference point in this scenario. Under normal circumstances, the longest communication durations between vehicles occur between i and j. The next thing that needs to be done is a comprehensive analysis of the link duration for these two different scenarios.

At time $t_0 = 0$, if the car i is in front of car j, then car i can communicate with car j with just one-hop. This is demonstrated in Figure 4a.

A random variable will be used to determine the initial distance between vehicle X. There is not going to be any change in the maximum communication radius of vehicle R. X fulfills these conditions at the outset:

$$0 \le X < R \tag{8}$$

On a highway, there will reportedly be instances of accelerating, decelerating, and passing other vehicles in accordance with the system model. On this stretch of road, the maximum speed limit is v_m, which means that all vehicles must travel at a speed that is either lower than or equal to v_m. Consider the fact that the vehicle is moving at a speed of $v(0)$ and that its acceleration is $a(0)$.

When time is less than zero, the acceleration is denoted by the symbol $a(t)$, and the speed is denoted by the $v(t)$. The acceleration, as measured at time t, is calculated as follows:

1 If $a(0) = 0$, for all $t \ge 0$,

$$a(t) = 0. \tag{9}$$

2 If $a(0) > 0$

$$a(t) = \begin{cases} a(0) & t \le \frac{v_m - v(0)}{a(0)} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

3 If $a(0) < 0$

$$a(t) = \begin{cases} a(0) & t \le \frac{-v(0)}{a(0)} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Based on the analysis presented above, it is reasonable to conclude that when $t_0 = 0$, assuming that the current acceleration $a(0)$ is zero, the instantaneous acceleration would be 0 as well, which is denoted by the equation $a(t) = 0$. If the speed is less than (0) the maximum speed limit, the acceleration is presumed to be $a(0)$; if the speed is more than the maximum speed limit, the acceleration is set to 0. If the speed is above the maximum speed limit, it is changed to 0. Taking into consideration the speed at the beginning, $v(0)$, the instantaneous speed, $v(t)$, can be defined by the following formula at time t:

$$v(t) = v(0) + \int_0^t a(u)du \tag{12}$$

where $u \in [0, t]$, $a(u)$ is the acceleration at time u.

The instantaneous speed may now be calculated by combining Eqs. (6-9):

1 If $a(0) = 0$, for all $t \ge 0$,

$$v(t) = v(0)$$

2 If $a(0) > 0$

$$v(t) = \begin{cases} v(0) + a(0)t & t \le \frac{v_m - v(0)}{a(0)} \\ v_m & \text{else} \end{cases} \tag{13}$$

3 If $a(0) < 0$

$$v(t) = \begin{cases} v(0) + a(0)t & t \le \frac{-v(0)}{a(0)} \\ v_m & \text{else} \end{cases} \tag{14}$$

The following formula can be used to calculate the distance traveled by any vehicle traveling at a speed of $v(x)$ during the interval of time represented by [0, t]:

$$S(t) = \int_0^t v(x)dx \tag{15}$$

Again, in accordance with the definition presented above, it is possible to determine the gap in space that exists between vehicles i and j at a particular instant in time, as shown in Figure 4.

If we assume that the initial acceleration and speed of i and j vehicles are, respectively $a_i(0), v_i(0), a_j(0)$, and vj(0), then the instantaneous acceleration and speed of vehicles i and j at time t are, respectively $a_i(t), v_i(t), a_j(t)$, and $v_j(t)$. The following equation can be used to determine the distance between the cars I and j at the time [0, t] if Eqs. 9–15 is followed to their logical conclusions:

$$S_i(t) = \int_0^t v_i(x)dx \tag{16}$$

$$S_j(t) = \int_0^t v_j(x)dx \tag{17}$$

If the distance between vehicles i and j at the beginning of the race is X, the distance $d_{i,j}$ between those two vehicles at the end of the race is:

$$d_{i,j} = \begin{cases} S_j(t) + S_i(t) + X & \text{same direction} \\ S_j(t) - S_i(t) + X & \text{opposite direction} \end{cases}. \tag{18}$$

The conclusion that the link is severed can be drawn from Equation 13, which states that when $d_{i,j} > R$, the connection is severed. The following thing that needs to be done is an analysis of the link duration in Figure 2b. This demonstrates how two automobiles meet while traveling in different directions:

$$S_j(t) + S_i(t) + X = R \tag{19}$$

Now it is possible to determine the maximum link duration t. We can calculate the maximum link duration t as follows using

the assumption that $S_j(t) + S_i(t) = \frac{1}{2}a_r t^2 + v_r t$, where $a_r = a_i + a_j$ and $v_r = v_i + v_j$.

$$t = \frac{-v_r + \sqrt{v_r^2 + 2a_r(R-X)}}{a_r}. \quad (20)$$

If two vehicles are going in the same direction, as shown in Figure 2a, determine which one is in front by overtaking, decelerating, and accelerating. Vehicle j is in front of vehicle i when $S_j(t) - S_i(t) + X > 0$; conversely, vehicle i precedes vehicle j. A symbolic function was defined as follows to efficiently convey which car is in front:

$$I(i,j) = \begin{cases} 1 & S_j(t) - S_i(t) + X > 0 \\ -1 & \text{otherwise} \end{cases} \quad (21)$$

When the connection is critically disconnected:

$$S_j(t) - S_i(t) + X = R \cdot I(i,j) \quad (22)$$

To compute the link duration at this time, two cases must be taken into account.

When $I(i,j) = 1$, vehicle j is in front of vehicle i. From Eq. (20), $S_j(t) - S_i(t) + X = R$. Similarly, to Eq. (20), because $S_j(t) - S_i(t) = \frac{1}{2}a_r t^2 + v_r t$, where $a_r = a_j - a_i$ and $v_r = v_j - v_i$The time t can be determined as:

$$t = \frac{-v_r + \sqrt{v_r^2 + 2a_r(R-X)}}{a_r}. \quad (23)$$

The position of vehicle i in relation to vehicle j is the same when $I(i,j) = -1$. $S_j(t) - S_i(t) + X = R$, and the link duration t may be calculated from Eq. (1) as follows:

$$t = \frac{-v_r - \sqrt{v_r^2 - 2a_r(R+X)}}{a_r} \quad (24)$$

To calculate the link duration between two nodes within a one-hop range: the transmitting node and the other nodes utilize equations (20), (23), and (24).

Proposition 3: If vehicle i and j's communication link fails at time t, the remaining link duration is a linear or square root function of X.

Proof: It is known that, at a constant speed v_i, $S_i(t) = \int_0^t v_i(x)dx$ is a linear function of t, i.e., $v_i = v_m$, based on the definition of S i (t) according to Eq. (15), which happens after the connection is lost at time t. S i (t) = at + b. Similarly, it may be inferred that $S_j(t) = ct + d$ is a linear function of t if v_j is a constant. Eq. (23) clearly states that $(a + c)t + b + d + X = R$ when the two cars are moving in different directions. The linear function for the link duration time is $t = \frac{R-b-d-X}{a+c}$. Once a connection is made between two co-directional vehicles i and j, the link duration is also a linear function, as shown by the equation $(c - a)t - b + d + X = R \cdot I(i,j)$. As a result, when v_i and v_j are both constants, and the link duration t is a linear function. A quadratic polynomial will represent the distance function if either $v_i(t)$ or $v_j(t)$ are not constants. Allow $v_i(t) = v_{\ i}(0) + a_i t$ without losing generality. The distance function is, by definition:

$$\begin{aligned} S_i(t) &= \int_0^t v_i(0) + a_i \\ &= v_i(0)t + \frac{1}{2}a_i t^2 \end{aligned} \quad (25)$$

Eqs. (14) and (22) demonstrate that $S_j(t) \pm S_i(t) = \frac{1}{2}a_r t^2 + v_r t$. Consequently, the answer is a square root function of X.

First Theorem Vehicles i and j's link's duration T have a log-normal distribution.

Proof According to Proposition 3, the connection duration can be written as $aX + b$ or $\sqrt{aX + b} + c$. The expression $\sqrt{aX + b} + c$ is log-normally distributed according to Proposition 1. $aX + b$ is distributed log-normally according to Proposition 2. As a result, the link's duration always follows a log-normal distribution. The theorem's proof is now complete.

The link duration of the cars i and j is log-normally distributed as $T \in \log N(\mu_t, \sigma_t)$, where the variance is σ_t and expected is μ_tBased on the link duration between the nodes. The duration of the link may then be used to assess its dependability. The likelihood of two vehicles directly communicating throughout the time period T_p. is known as link reliability. When $t = t_0$The connection dependability between nodes is, supposing that any two vehicle nodes have a communication link (l):

$$r(l) = P\{t_0 + T_p \mid t_0 \in M\} \quad (26)$$

where M is the time at which the two vehicles' connection began. The link dependability is as a result:

$$r_t(l) = \begin{cases} \int_{t_0}^{t_0+T_p} f(T)dT & T_p > 0 \\ 0 & \text{else} \end{cases} \quad (27)$$

where the probability function for the time period T is expressed by f(T).

3.2. *Simulation setup*

As seen in Figure 5, the experiment was simulated using NS-2 and a 3511m*3009m square topology using a real-life map of " Anfa District in Casablanca." The topology included intersections and straight roads, each of which had been built up with two-way lanes. To add authenticity to the simulation, each vehicle node in the network employed an intelligent driving model (IDM) that featured waiting, avoiding, overtaking, and lane switching.

The default NS-2 parameter settings are shown in Table 2. we have used a CIF (H.261) video file format with a 352 x 288. UDP was utilized for transmission following the work approach shown in Figure6.

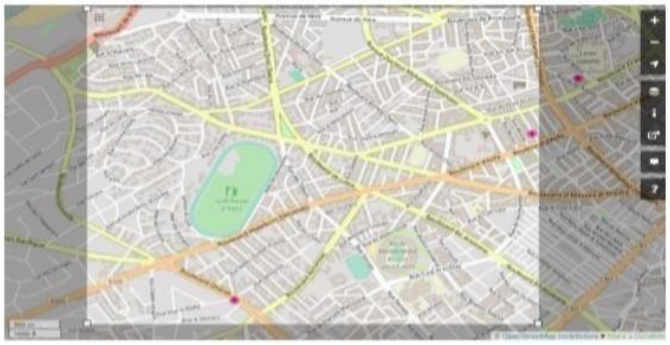


Figure 5: Anfa District Casablanca "OpenStreetMap"

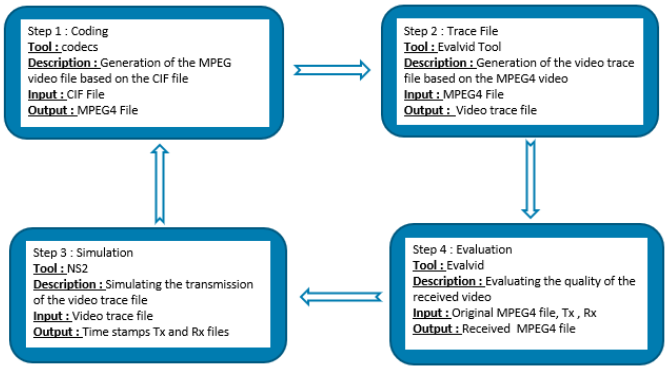


Figure 6: Work approach

In contrast with the Enhanced Distributed Channel Access (EDCA) defined in the 802.11p standard, the video transmission was carried out using an Enhanced Dynamic Cross-layer mechanism for real-time HEVC streaming. This method has proven to deliver improved performance in video quality at reception and end-to-end latency. QoS and QoE calculations have been in use for the verification of the effectiveness of the cross-layer mechanism shown in Figure7 in a separate paper.

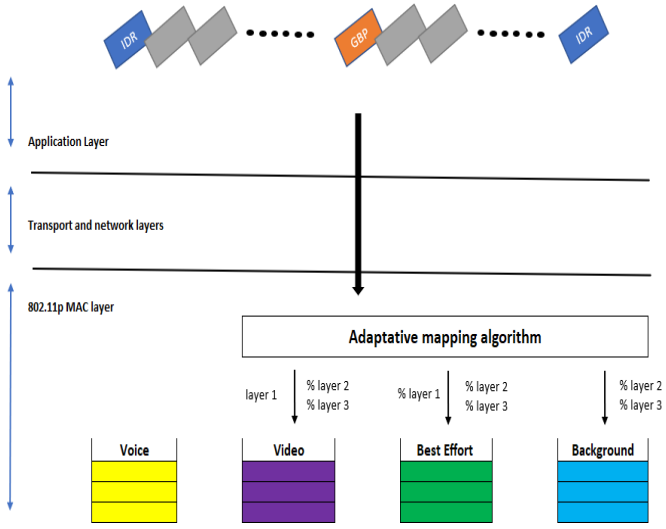


Figure 7: Illustration of the Adaptative mapping algorithm.

Each beacon packet's size was determined based on the data that was transmitted. Two scenarios that simulated the scenario in Figure 5 were put up to successfully assess the RSAR algorithm's performance. The first scenario used a network with 80 nodes and vehicle peak speeds between 30 and 90 kilometers per hour. In this case, we will focus on how the speed of the vehicle impacts the routing protocol and provide an explanation. The alternative scenario went as follows: the cars' top speed was fixed at 40 km/h, and there was anything between 60 and 120 nodes. The discussion in this instance will center on how vehicle density affects the routing protocol. The nodes were first scattered at random along several roadways and followed a predetermined path.

Each simulation was repeated 20 times with an average value of 300 s between each run. The simulation data were then assembled.

3.3. *Simulation results*

The QLAODV algorithm, the SLBF algorithm, the GPSR algorithm [11, 14, 17], and the techniques described in [22] were contrasted with the RSAR algorithm suggested in this study. All of these algorithms are examples of routing algorithms. It is feasible to assess the benefits and drawbacks of RSAR thanks to this comparison. The results illustrated in Figs. 8, 9, 10, 11, 12, 13, and 14 were acquired by comparing end-to-end delays; the packet delivery ratios hop number in various scenarios. Figures 8, 9, and 10 show the comparisons of the first test, while figures 11, 12, and 13 show the comparisons in the second test case.

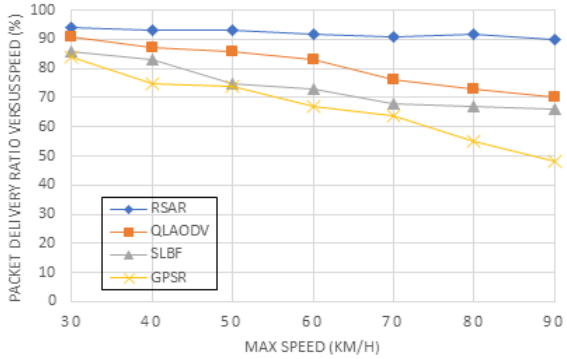


Figure 8: Packet delivery ratio on different speeds

Figure 8 shows that there is a direct association between the speed and packet delivery ratio. A node's packet delivery ratio is calculated by data packets supplied by that node by the data packets received by the destination node. Figure 8 demonstrates that the RSAR suggested in this paper was able to achieve a consistent and high packet delivery ratio of video data; as speed increased, the average ratio up surged to more than 90%. With increasing speeds, the packet delivery ratios of the other three investigated routing methods dropped dramatically. This enhancement was made possible thanks to RSAR's careful consideration of how to speed variations affect link reliability. The Q-Learning method uses this information as a learning parameter to make routing decisions by analyzing the links between nodes and determining the dependability of those links. Due to its greedy approach and lack of consideration for connection reliability, while choosing the next hop node, GPSR's packet delivery ratio decreased quickly. Until the speed reached 54 kilometers per hour, QLAODV's packet delivery ratio was over 90%, but it rapidly dropped after the speed was exceeded. QLAODV, although using the Q-Learning paradigm, must

constantly repair the route with increasing velocity to provide an always-reliable path from beginning to finish. As a result, the proportion is reduced. Due to its sensitivity to topology changes, SLBF's packet delivery ratio decreased and does not take connection stability into full account. It was more similar to the GPSR algorithm because it used a greedy approach.

Figure 9 shows the connection between speed and end-to-end time delay; In this case, the time delay is measured as the typical amount of time it takes for the receiving node to receive a valid data packet. Figure 9 shows that when the vehicle nodes' speed rose, all four techniques' time delays showed a growing trend. The RSAR algorithm suggested in this paper had a rather constant time delay that fell between SLBF and GSPR. When the maximum speed was greater than 60 km/h, the time delay of SLBF quickly grew and surpassed that of RSAR. The effects of topology modifications and the requirement for a new computation of the re-transmission method or effective forwarding area were the main causes of this trend. While topological modifications have less effect on RSAR, this method ensures that the route with the best Q-value has the most available bandwidth, the least unreliable connection, and the shortest routing distance. Because GPSR solely employs the greedy forwarding method, its time delay was the shortest. The GPSR had the lowest packet delivery ratio since it exclusively employs the greedy mechanism, dropping packets as soon as they don't arrive. Although QLAODV also employs the Q-Learning model, its increased speed requires frequent path switching in order to maintain an efficient routing path.

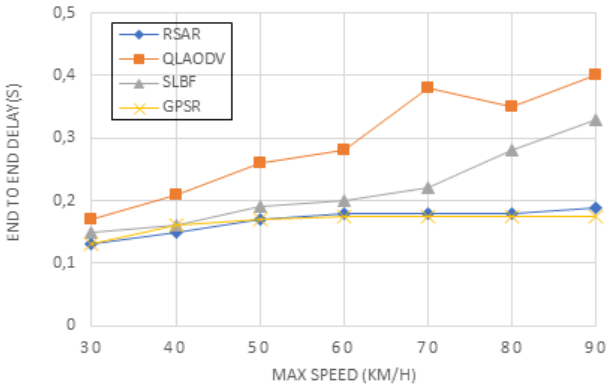


Figure 9: End-to-end delay on different speeds

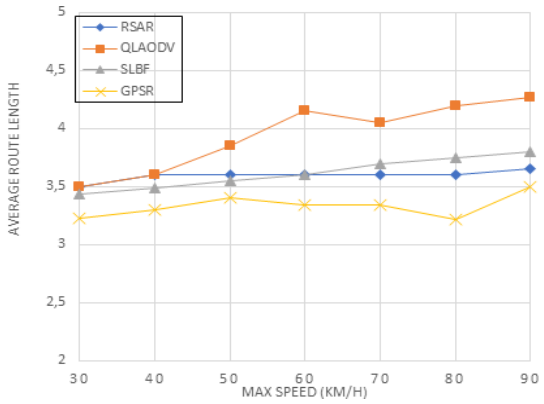


Figure 10: Routing length on different speeds

This resulted in a significant lengthening of the packet delivery time. According to Figure 9, in the context of a topology that was rapidly changing, the RSAR packet delivery ratio ranged from 0.1 to 0.2.

Figure 10 shows the relationship between speed and route length using the average number of hops required by valid data packets to reach their destination. Rather than relying on a route transformation mechanism to ensure that the whole path is maintained, this approach employs a simple forwarding decision to determine which node has the greatest Q-value; as shown in Figure 10, RSAR's route length was less than that of QLAODV. The RSAR suggested in this paper has a consistent route length, as seen in Figure 10. The maximum Q-value and store-and-forward selection mechanisms make this feasible by always using the shortest route possible. Both GPSR and SLBF use a greedy packet forwarding strategy, however when the speed increases over 60 km/h and topological changes become more pronounced, SLBF's route length increases; despite the fact that GPSR employs no trustworthy technique, the lowest number of hops was reached.

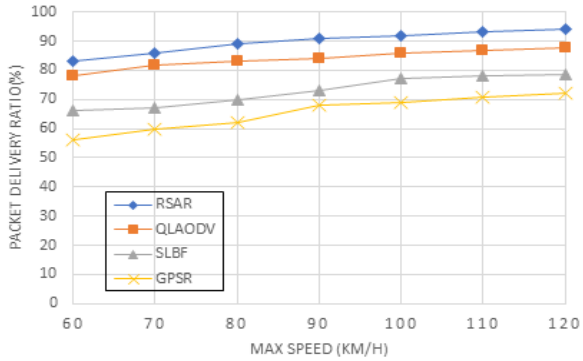


Figure 11: Packet delivery ratio by number of nodes

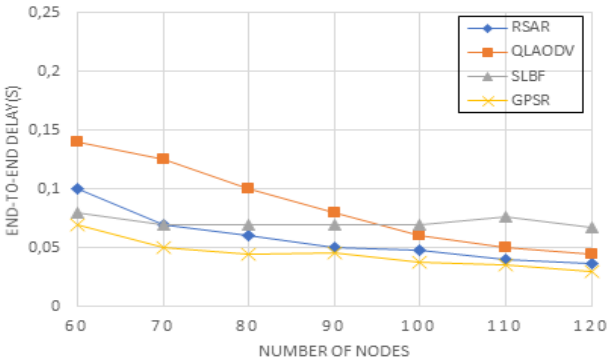


Figure 12: End to end by the number of nodes

The relationship between packet delivery ratios and the number of nodes is seen in Figure 11. Overall, the packet delivery ratio improves with increasing network size across all four approaches as the number of nodes increases. Over 90% packet delivery ratio is achieved in RSAR with 90 nodes, while the packet delivery ratio for QLAODV is 85%. RSAR fully takes into account connection stability between nodes despite the fact that both of these algorithms are based on the Q-Learning concept,

which results in a significantly higher packet delivery ratio. The store-and-forward mechanism makes RSAR's packet delivery ratio superior to that of GPSR and SLBF in the presence of a growing number of nodes. As SLBF employs the retransmission technique, it has a higher packet delivery ratio than GPSR, but because topology has a significant impact because the sending node and vehicle are frequently on different paths, it has a lower packet delivery ratio than RSAR.

Figure 12 describes the link between the end-to-end time delay and the number of nodes. The RSAR's end-to-end latency is quickly catching up to the GPSR's. This is because when nodes are added, RSAR learning nodes are also created, reducing the time delay and shortening the route to the target node. The time delay of QLADOV eventually approaches that of RSAR as the number of nodes rises, but with fewer nodes, it is significantly bigger than that of RSAR. As QLAODV spends so much time maintaining the routes, this behavior arises. Because it employs a scheduled broadcast technique that causes an increase in time delay, SLBF drops slowly with an increase in the number of nodes. As the number of nodes rises, the time delay for each of the four algorithms decreases.

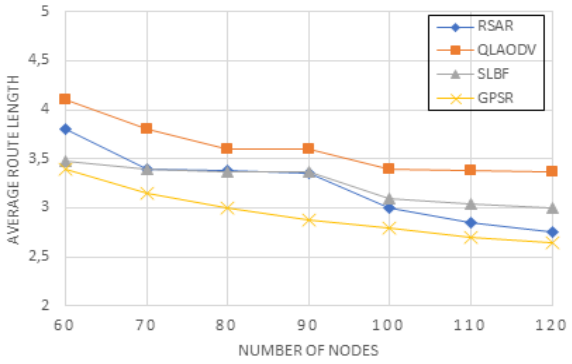


Figure 13: Routing length versus the number of nodes

Figure 13 shows the association between the number of nodes and the average route length. The average route length of all four techniques decreases as the number of nodes rises. The primary reason for this is an increase in the number of effective forwarding nodes. RSAR's route discovery is shorter than QLAODV's path discovery. As GPSR and SLBF both employ a greedy strategy, the path lengths of GPSR, SLBF, and RSAR are near to one another. RSAR chooses the next hop that is closest to the furthest node as the number of nodes increases.

3.4. Summary and discussion

As seen in the previous section, the simulation was based on the two factors that affect the most quality of video transmission those are Speed and the number of nodes.

When speed increases, we have determined that in terms of packet delivery ratio, RSAR was able to achieve a high and consistent packet delivery ratio of video data, with an average ratio of more than 90%, as speed increased. In terms of end-to-end delay, the RSAR algorithm had a rather constant time delay that fell between GSPR and SLBF. In terms of route length, RSAR's route length was less than that of QLAODV, thanks to the maximum Q-value selection and store-and-forward mechanisms.

In terms of packet delivery ratio, we have concluded that when the number of nodes increases, RSAR kept a ratio above 90%, while the packet delivery ratio for QLAODV was around 85%. In terms of end-to-end delay, The RSAR steadily approaches the GPSRs thanks to the fact that when nodes are added, RSAR's learning nodes are added as well, making the path to the time delay and destination node shorter. Each of the four approaches has a reduced average route length as the number of forwarding nodes increases.

4. Conclusion and future work

In this research, a reliable adaptive routing service method is suggested to address the issues of dramatic topological changes and unstable links brought on by quick vehicle movements in VANET. First, it was established that links had a log-normal distribution in terms of duration after studying the characteristics of vehicle motion and the causes of links' lack of reliability. Second, a link reliability calculation model was created using this information. After the link reliability between nodes was evaluated as a parameter and employed in the Q-Learning process, the RSAR method was then recommended. The effectiveness of four routing methods was then evaluated using simulated trials to measure the number of hops, packet delivery ratios, and end-to-end delays. Under varying circumstances, the findings demonstrated that RSAR achieved a greater packet delivery ratio with a shorter transmission time delay. Due to its self-learning capabilities, RSAR is well-suited to deal with the issues caused by topological changes. Local diffusion is a kind of technique used in learning, and there may be a huge number of nodes involved in choosing the route, so the routing cost in a big network environment will be quite high. As a result, the authors' future work will concentrate on the issue of routing expense.

The creation of the algorithm that was used in finishing tracks at the MAC layer while meeting low latency requirements. Due to the time limit, packages that are not received by the recipient within the time limit do not count towards the total duration of the film. This is done in order to maintain the integrity of the timeline. Because of this, their transmission is completely worthless and does nothing more than burden the network. This feature of the algorithm must be taken into account, as eliminating them at the transmitter would enhance transmission. The technique should form a relationship between the application time constraint, the round-trip time limit, and the transition time represented by the completion of the waiting list to be considered valid.

References

[1]. M.G. W.L. Junior, D. Rosário, E. Cerqueira, L.A. Villas, "A game theory approach for platoon- based driving for multimedia transmission in VANETs" , Wireless Communications and Mobile Computing, **2414658**, 1–11, 2018, doi:https://doi .org /10 .1155 /2018 /2414658

[2]. A.V.V. M. Jiau, S. Huang, J. Hwang, "Multimedia services in cloud-based vehicular networks" , IEEE Intelligent Transportation Systems Magazine, **7** (3), 62–79, 2015, doi:https://doi .org /10 .1109 /MITS .2015 .2417974

[3]. X.Z. M. Gerla, C. Wu, G. Pau, "Content distribution in VANETs", Vehicular Communications, **1**(1), 3–12, 2014, doi:https://doi.org/10.1016/j.vehcom.2013.11.001

[4]. R.S. C. Campolo, A. Molinaro, "From today's VANETs to tomorrow's planning and the bets for the day after", Vehicular Communications, **2** (3), 158–171, 2015, doi:https://doi .org /10 .1016 /j .vehcom .2015 .06 .002

[5]. R.F.S. D. Perdana, "Performance comparison of IEEE 1609.4/802.11p and 802.11e with EDCA implementation in MAC sublayer", International Conference on Information Technology and Electrical Engineering (ICITEE), 285–290, 2013, doi: 10.1109/ICITEED.2013.6676254

[6]. N. Mittal, A. Singh, "A Critical Review of Routing Protocols for VANETs" , International Conference on Innovative Computing and Communications, Springer Nature Singapore Pte Ltd, 135-142, 2019, doi: doi.org/10.1007/978-981-13-2324-9_14

[7]. F. Goudarzi, H. Asgari, and H. S. Al-Raweshidy, "Traffic-Aware VANET Routing for City Environments—A Protocol Based on Ant Colony Optimization", IEEE SYSTEMS JOURNAL, 1-11, 2018, doi: 10.1109/JSYST.2018.2806996

[8]. IEEE P1609.4, "Trial-Use Standard for Wireless Access in Vehicular Environments (WAVE) — Multi-Channel Operation ", IEICE transactions on fundamentals of electronics, communications and computer sciences, **12** (2010), 2691-2695, 2006.

[9]. S. ZAIDI, "Le streaming vidéo dans les réseaux véhiculaires ad-hoc," Doctoral thesis, UNIVERSITE MOHAMED KHIDER- BISKRA, 2018, doi : http://thesis.univ-biskra.dz/id/eprint/3870

[10]. M. Rouse, "Vehicle-to-vehicle communication (V2V communication)" , Internet of Things Agenda, 2014, doi : https://doi.org/10.48550/arXiv.2102.07306

[11]. M. Zorkany, N.A. Kader, "Vanet routing protocol for v2v implementation: A suitable solution for developing countries" , Cogent Engineering, **4**(1), 1362802, 2017, doi : https://doi.org/10.1080/23311916.2017.1362802

[12]. S. Chakrabarti, L. Technologies, A. Mishra, V. Tech, "QoS Issues in Ad Hoc Wireless Networks" QOS and resource allocation in the 3rd generation wireless networks, IEEE, 142-148, 2001, doi : 10.1109/35.900643

[13]. G. Ya, S. Olariu, "A probabilistic analysis of link duration in vehicular ad hoc networks", IEEE Intelligent Transportation Systems Magazine, **12** (4), 1227-1236, 2011, doi : 10.1109/TITS.2011.2156406

[14]. J. Toutouh, "Intelligent OLSR routing protocol optimization for VANETs", IEEE Transactions on Vehicular Technology, **61**(4), 1884-1894, 2012, doi : 10.1109/TVT.2012.2188552

[15]. Y.N. Zhu, "A new constructing approach for a weighted topology of wireless sensor networks based on local-world theory for the internet of things (IOT)", Computers & Mathematics with Applications, **64** (5), 1044-1055, 2012, doi : https://doi.org/10.1016/j.camwa.2012.03.023

[16]. YP . Liang, " A kind of novel method of service-aware computing for uncertain mobile applications", Mathematical and Computer Modelling, **57** (3–4), 344–356, 2013, https://doi.org/10.1016/j.mcm.2012.06.012

[17]. XD Song, X Wang, " Extended AODV routing method based on distributed minimum transmission (DMT) for WSN", AEU - International Journal of Electronics and Communications, **69**(1), 371–381, 2015, doi : https://doi.org/10.1016/j.aeue.2014.10.009

[18]. Z Ke, Z Ting, " A novel multicast routing method with minimum transmission for WSN of cloud computing service", Soft Computing, **19**(10): 1817-1827, 2015, doi : https://doi.org/10.1007/s00500-014-1366-x

[19]. XD Song, X Wang, " Extended AODV routing method based on distributed minimum transmission (DMT) for WSN", AEU - International Journal of Electronics and Communications, **69**(1), 371–381, 2015, doi : https://doi.org/10.1016/j.aeue.2014.10.009

[20]. T Saravanan, N.S. Nithya, " Modeling displacement and direction aware ad hoc on-demand distance vector routing standard for mobile ad hoc networks", Mobile Networks and Applications, **24**(9), 1804–1813, 2019, doi : https://doi.org/10.1007/s11036-019-01390-9

[21]. T Saravanan, N.S. Nithya," Energy aware routing protocol using hybrid ANT-BEE colony optimization algorithm for cluster based routing", 4th IEEE ınternational conference on computing communication and automation (ICCCA), 1–6, doi : 10.1109/CCAA.2018.8777662

[22]. P Sherubha, N Mohanasundaram, " An efficient network threat detection and classification method using ANP-MVPS algorithm in wireless sensor networks", International Journal of Innovative Technology and Exploring Engineering (IJITEE), **8**(11), 2278-3075, 2019, doi : https://www.ijitee.org/wp-content/uploads/papers/v8i11/F3958048619

[23]. V. Tavares, J.M.R.S. Boulogeorgos, A.A.A. Vuppalapati, " International Conference on Communication, Computing and Electronics Systems", Lecture Notes in Electrical Engineering, Springer, Singapore, 733, doi : https://doi.org/10.1007/978-981-33-4909-4_60

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 4, 92-97 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Metamaterial-Inspired Compact Single and Multiband Filters

Ampavathina Sowjanya*, Damera Vakula

ECE Department, National Institute of Technology, Warangal, 506004, India

ARTICLE INFO

Article history:
Received: 25 May, 2022
Accepted: 26 July, 2022
Online: 16 August, 2022

Keywords:
Bandpass Filter
Metamaterial
Fractional Bandwidth
Roll-off
Insertion loss

ABSTRACT

In this paper, Compact bandpass filters have been designed. A single bandpass filter was designed using novel triple concentric complementary split-ring resonators placed along the microstrip line in the ground plane. Gaps and via were placed on the microstrip line to control electromagnetic characteristics, resulting in a single bandpass filter. In turn, spiral resonators were attached to the microstrip transmission line at the gaps in the transmission line to obtain a compact dual passband filter. Stepped impedance microstrip line and T-shaped stubs were attached to the microstrip line in between spiral resonators. The structure designed resulted in a Triple bandpass filter. A fractional bandwidth of 3% was achieved at the center frequency of 3GHz. The filter had a 1.5dB insertion loss which is the minimum value in the operating frequency band. The filter resonance frequency was 1.32 GHz and 2.47GHz which have a fractional bandwidth of 7.5% and 4.85% respectively and the corresponding insertion loss was 1.3dB and 1.8dB respectively. The triple bandpass filter had a fractional bandwidth of 1.16%, 11.4%, and 1.86%, centered at 1.29 GHz, 2.27 GHz, and 3.21GHz with 1.6dB, 1.3dB, and 1.8 dB insertion loss at the respective frequencies. The proposed bandpass filters are useful for GPS, WLAN, WiMAX, and radar applications.

1. Introduction

In modern wireless communication systems, there is an urgent need for compact filters with low insertion loss, and high selectivity to transmit a specific band of frequencies. Bandpass filters allow desired frequencies and reject unwanted frequencies. The basic requirement for wireless communications is portability. For wireless devices to be portable, internal components must be as compact as possible. Metamaterial provides the opportunity to exploit its unusual properties to control the electromagnetic characteristics of microwave devices such as filters, antennas, couplers, power dividers, etc., to design compact devices with better performance characteristics. The controlling of electromagnetic parameters is done by the attainment of negative permittivity and/or permeability over a certain range of frequencies by engineering materials such as altering their dimensions or creating gaps for better performance of microwave devices. In [1], backward wave propagation with split ring resonators which are loaded along the transmission line is demonstrated. In [2], a bandpass filter is designed using a high-temperature superconducting structure based on SRRs, but the filter was large. In [3] S-shaped complementary resonant structures were used to design a bandpass filter, but insertion loss was high in the passband. In [4], a complementary split-ring resonator (CSRR) based dual-mode patch bandpass filter is proposed, but the filter was large. In [5], a hybrid bandpass filter using substrate integrated waveguide and CSRR is proposed, but it had high insertion loss. In [6], a bandpass filter using multiple SRR and CSRRs was designed, but with large size. In [7], a microstrip bandpass filter for X band applications is presented. The filter is designed with SRRs but had high insertion loss with large size. In [8], a bandpass filter using an optimized coupling matrix synthesis method is designed, which was large. In [9], a bandpass filter based on half mode substrate integrated waveguide, periodic CSRRs, and defected ground structures is proposed, but is large. In [10], a low-profile metamaterial bandpass filter loaded with a four-turn complementary spiral resonator for wireless power transfer applications is designed. In [11], a full mode substrate integrated waveguide bandpass filter using CSRR is designed, that has high insertion loss.

Multiband filters are in demand to avoid the crowding of multiple devices for different applications. To accommodate multiple applications within a device and have better performance characteristics becomes challenging for the researchers. Numerous methods have been proposed. In [12], two dual bandpass filters one using CSRRs, complementary spiral resonators (CSR), and tapered substrate integrated waveguide;

*Corresponding Author: Ampavathina Sowjanya, sowjanya6754@gmail.com

another one with CSRRs, CSR, and substrate integrated waveguide are designed. In [13], a filter with dual passbands using tapered SRRs which consist of asymmetric shape square rings with interlinking is developed. In [14], a bandpass filter with CSRR and defected ground structures using substrate-integrated waveguide technology is proposed. In [15], a dual bandpass filter based on coupling between two identical SRRs and a CSRR is presented. In [16], varactor diodes were loaded on split ring slots to achieve a unit cell for frequency tunability is designed. In [17], a bandpass filter was designed for three bands using stub loading on resonators. The resonators were further coupled with internal resistors for dual mode operation. In [18], a compact three stubs loaded open-loop resonator-based triple bandpass filter is presented. In [19], resonators with open loops for filters with three pass bands are designed. In [20], a bandpass filter for triple bands using metamaterials is proposed. Rectangular stubs with a meander line were used in the design. The filters mentioned are large [12-20].

In this paper, metamaterial-inspired LC resonant structures were implemented to design a single bandpass filter. Firstly, gaps were formed along the microstrip transmission line. Via filled with copper was introduced along the microstrip transmission line. Triple CSSRs were etched in the ground plane just beneath the microstrip transmission line. Bandpass response with better sensitivity and insertion loss was achieved using the above configuration. The dual bandpass filter was designed by adding spiral resonators connected to the gaps introduced in the microstrip line to the proposed single bandpass filter. Stepped impedance resonant microstrip line and T-shaped stubs were added between spiral resonators attached to the microstrip line to the proposed dual bandpass filter to design a triple bandpass filter that helps in improving return loss.

2. Design Methodology

The microstrip transmission line allows all frequencies to pass through it. An electric field was induced from the microstrip line to a ground plane and a magnetic field around the microstrip transmission line. A bandpass filter was designed in this paper first by stopping all frequencies using gaps along the microstrip transmission line. By introducing a via filled with copper along the microstrip transmission line between the microstrip line and ground plane through the substrate, a passband was observed with high insertion loss and low return loss. Triple complementary concentric split-ring resonators were etched in the ground plane along the microstrip transmission line to act as LC resonant structure was excited by a time-varying electric field parallel to its axis and provided a single negative permittivity medium which introduced a passband over a certain range of frequencies but with low return loss in the passband and better insertion loss; sensitivity from passband to stopband transition and vice versa was also observed. This combination of gaps, vias, and complementary structures was introduced along with the transmission line resulting in passband response with better insertion loss and sensitivity by choosing appropriate dimensions for concentric rings, microstrip lines, vias, and gaps. Spiral resonators with optimum dimensions were connected to the gaps of the microstrip transmission lines which are coupled electrically to the triple concentric split ring resonators introducing one more passband, which helps in reducing insertion loss. Thus, a compact dual bandpass filter is designed with better insertion loss and high selectivity in the first band, and better insertion loss and selectivity in the second band. The stepped impedance resonant microstrip line for the proposed dual bandpass filter generated one more passband and the T-shaped stubs attached to the microstrip line in between spiral resonators helped to improve return loss. Thus, a compact triple bandpass filter was designed.

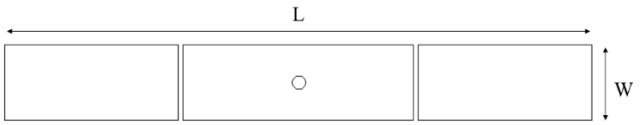


Figure 1: Top view of the single bandpass filter

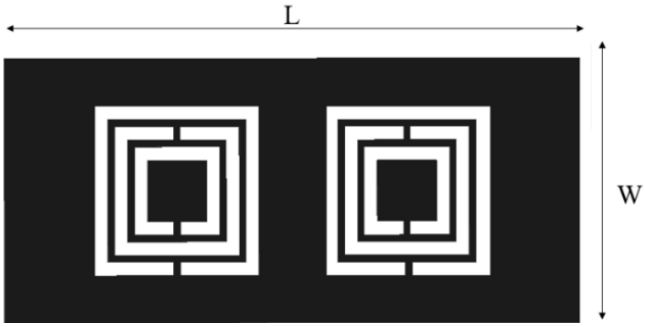


Figure 2: Bottom view of the single bandpass filter

The proposed filters were realized on RT Duroid 3010 substrate that has a height of 0.13 mm, and a dielectric constant of 10.2. Microstrip line made with 35 µm thickness copper material had a width of 2.25 mm for better impedance matching with 50Ω transmission line. Gaps along the microstrip line were 0.15 mm while via filled with copper had a diameter of 0.4 mm, and placed through the substrate from the microstrip line to the ground plane. Complementary split rings had a width of 0.4 mm and gaps in them had a width of 0.2 mm with gaps between concentric rings 0.2 mm. The outer ring had a length and width of 5 mm on each side. The top and bottom views of the simulated single bandpass filter are shown in Figure.1 and 2 respectively. The dimensions of the proposed filter were 17.5 mm×7.81 mm which was approximated by the guided wavelength λ_g at the center frequency as 0.55 λ_g ×0.25 λ_g.

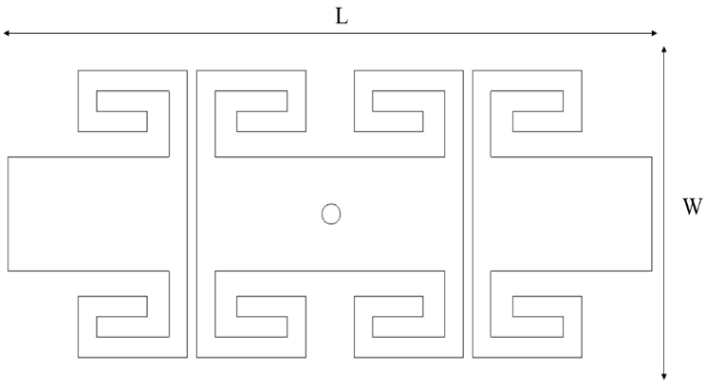


Figure 3: Top view of the dual bandpass filter.

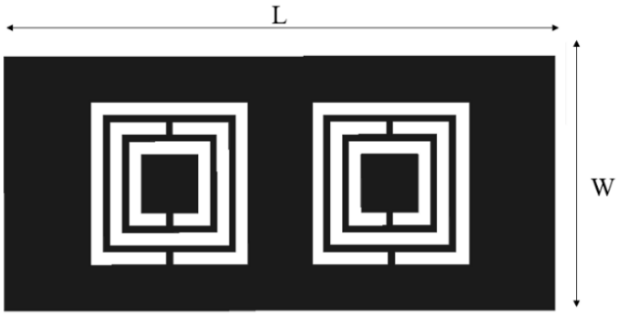


Figure 4: Bottom view of a dual bandpass filter

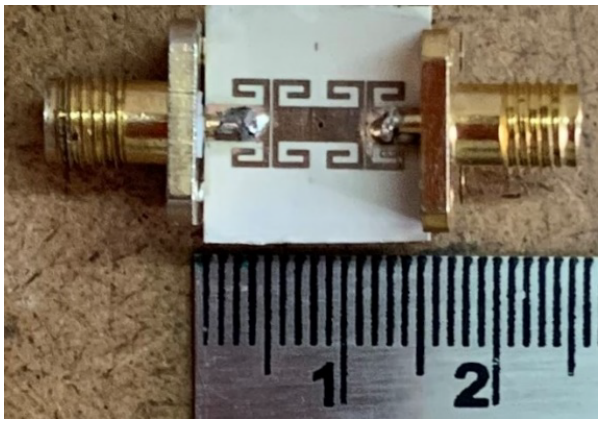


Figure 5: Fabricated top view of a dual bandpass filter

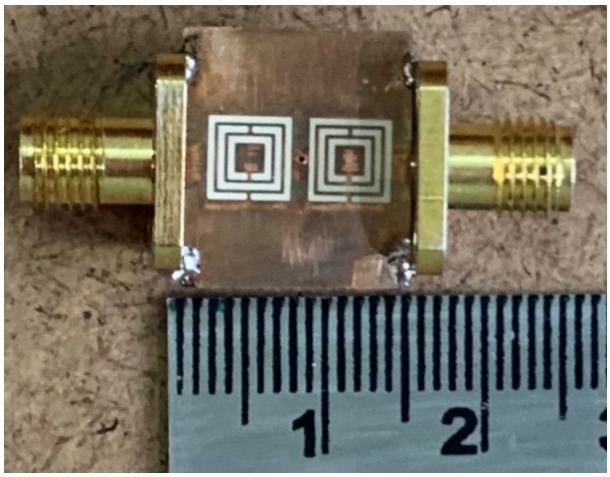


Figure 6: Fabricated bottom view of the dual bandpass filter

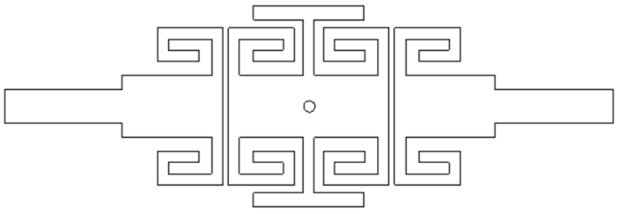


Figure 7: Top view of the triple bandpass filter

The dimensions of the filter were 14 mm × 15.81mm which is approximately 0.19 λ_g × 0.21 λ_g, where λ_g is the guided wavelength at the lower resonant frequency. The top and bottom views of the simulated dual bandpass filter are shown in Figures.3 and 4 while that of fabricated prototypes are shown in Figures.5 and 6, respectively.

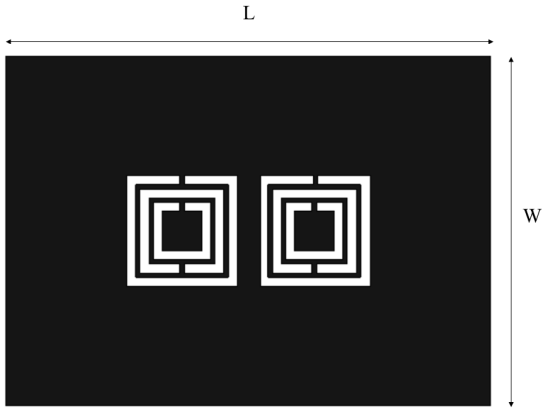


Figure 8: Bottom view of the triple bandpass filter

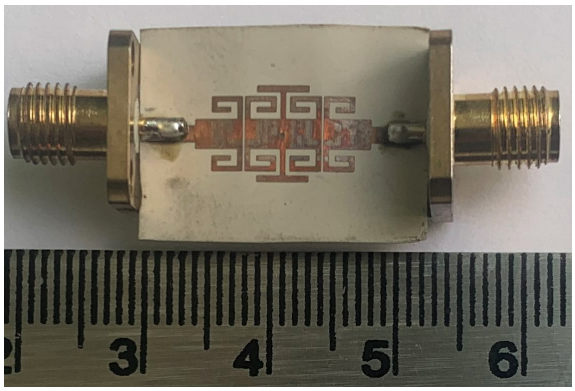


Figure 9: Fabricated top view of Triple bandpass filter

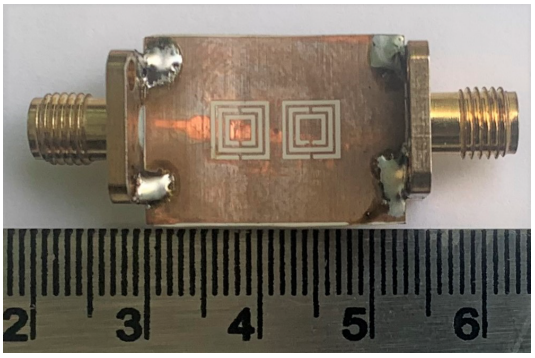


Figure 10: Fabricated bottom view of Triple bandpass filter

The dimensions of the designed filter were 22mm × 15.81mm which can also be represented as 0.3 λ_g × 0.2 λ_g, where the guided wavelength (λ_g) is calculated at the lowest resonant frequency. The structure of the triple bandpass filter is shown in Figures 7 and 8, while that of the fabricated filters is shown in Figures 9 and 10.

3. Results and Discussion

The proposed bandpass filters were designed and simulated with the help of the Ansys HFSS simulator using the finite element method (FEM). Ansys HFSS is three-dimensional electromagnetic simulation software. This software is used for designing and simulating high-frequency electronic products such as antennas, filters, etc., The most important component of FEM is mesh. Mesh can handle inaccuracies in the 3D model and produce reliable results consistently. The simulation results for the single bandpass filter are shown in Figure.11. The fractional bandwidth of the proposed filter is 3% at the 3GHz center frequency. The minimum insertion loss of the filter is 1.5 dB. The roll-off rate of the filter is 58.8 dB/GHz on the lower side and 67.8 dB/GHz on the upper side.

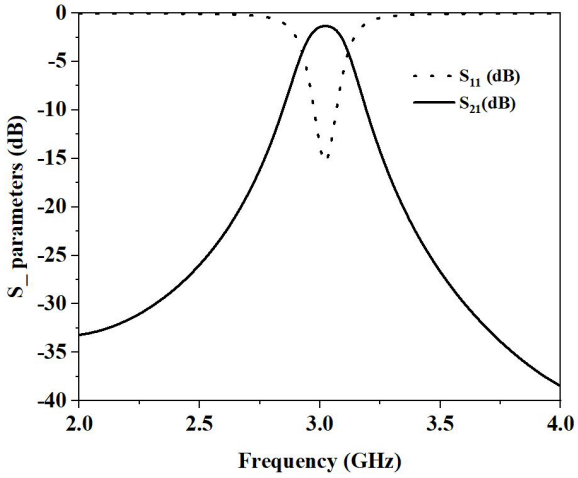


Figure 11: Simulation S-parameters of the single bandpass filter

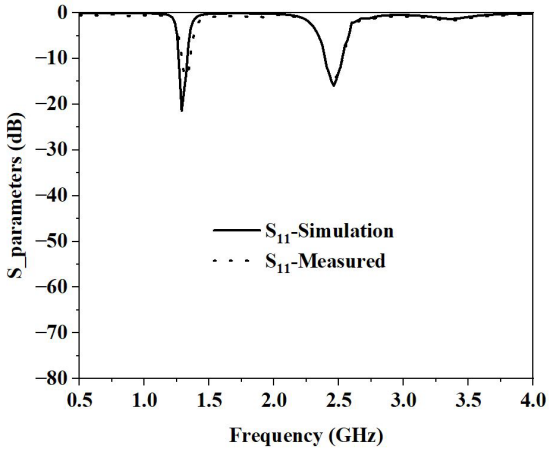


Figure 12: Simulation and measured return loss S_{11} of the dual bandpass filter

The performance of the proposed filters was validated by measuring S_parameters using vector network analyzer N5222A. The simulation and measured S_{11} and S_{21} results for the dual bandpass filter are shown in Figure.12 & Figure.13 respectively. The simulated dual bandpass filter has center frequencies at 1.3 GHz and 2.46 GHz. The fractional bandwidth was 8% for the first band, and 4.85% for the second band. The minimum insertion loss obtained was 1.15 dB and 1.7 dB for both bands respectively. The return loss was 21 dB for the first band and 17 dB for the second band. The measured dual bandpass filter had center frequencies at 1.32 GHz and 2.47 GHz. The fractional bandwidth of the proposed filter was 7.5% for the first band and 4.85% for the second band. The minimum insertion loss obtained was 1.3 dB for the first band and 1.8 dB for the second band. The return loss was 15 dB for the first band and 16 dB for the second band. The roll-off rate for the first band was 203 dB/GHz on the lower side, 64 dB/GHz on the upper side, for the second band 35 dB/GHz on the lower side, and 94.6 dB/GHz on the upper side. The electrical size of the dual bandpass filter is 0.039 λ_g^2.

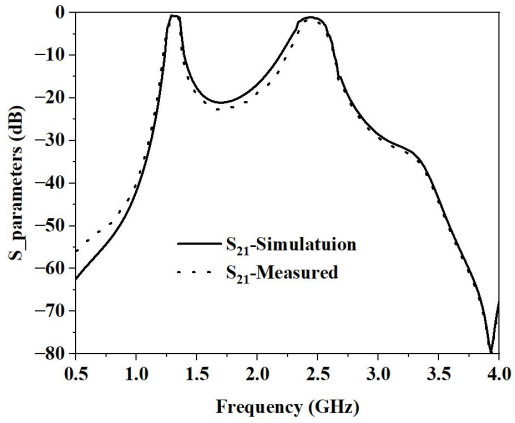


Figure 13: Simulation and measured insertion loss S_{21} of the dual bandpass filter

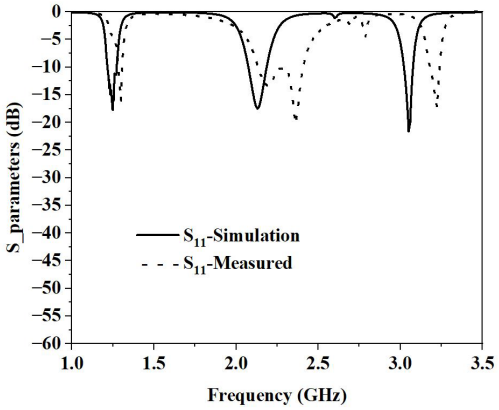


Figure 14: Simulation and measured return loss S_{11} of the triple bandpass filter.

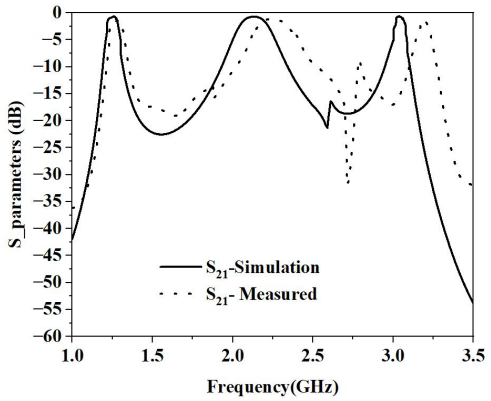


Figure 15: Simulation and measured insertion loss S_{21} of the triple bandpass filter.

The simulation results and those determined experimentally for S_{11} and S_{21} of the triple bandpass filter are shown in Figure.14 & Figure.15 respectively. The simulated triple bandpass filter has center frequencies at 1.27GHz, 2.2GHz, and 3GHz. The fractional bandwidth for the first band is 1%, for the second band is 6%, and for the third band is 1.8%. The minimum insertion loss obtained was 1.2 dB, 0.99 dB, and 1.5 dB respectively. The return loss was 18 dB,17 dB, 22 dB respectively. The measured triple bandpass filter had center frequencies at 1.29 GHz, 2.27 GHz, and 3.21 GHz. The fractional bandwidth was 1.16%, 11.4%, and 1.86% for all three bands respectively. The minimum insertion loss obtained was 1.6 dB, 1.3 dB, and 1.8 dB respectively in the passband. The return loss was 17 dB,13 dB,18 dB respectively. The roll-off rate for the first band was 262 dB/GHz on the lower side and 115 dB/GHz on the upper side, for the second band it was 56 dB/GHz on the lower side, 49 dB/GHz on the upper side, for the third band it was 47 dB/GHz on the lower side and 180 dB/GHz on the upper side. The electrical size of the triple bandpass filter is 0.06 λ_g^2.

Table 1: Single bandpass filter comparison table

References	Center frequency (GHz)	Fractional Bandwidth (%)	Insertion loss (dB)	Electrical size (λ_g^2)
4	3.41	3.5	<1	0.38
5	10	12	1.59	0.39
6	9.1	3	0.04	0.58
7	9	<1	2.24	0.92
8	2.4	12.5	0.7	0.39
Single bandpass filter	3	3	1.5	0.13

The measured dual bandpass filter operating at center frequencies of 1.32 GHz and 2.47 GHz with a fractional bandwidth of 7.5% and 4.85% respectively has a 0.039 λ_g^2 size is compact and useful for GPS and WLAN applications. The measured triple bandpass filter operates at center frequencies of 1.29 GHz,2.27 GHz, and 3.21 GHz with a fractional bandwidth of 1.16%, 11.4%, and 1.86% has 0.06 λ_g^2 is compact and useful for GPS, WLAN, WiMAX applications. The simulation and measurement results were in good agreement. The difference in measured results is due to dimensional tolerances during the etching of the copper material at the time of fabrication. The right shift in the measurement results may be due to a delta reduction in the dimensions of the design during fabrication.

From table.1., it is observed that the single bandpass filter is compact. From table 2, it is observed that the dual bandpass filter is compact. From table.3., it is observed that the triple bandpass filter is compact.

Table 2: Dual bandpass filter comparison table

References		Center frequency (GHz)	Fractional bandwidth (%)	Insertion loss (dB)	Electrical size (λ_g^2)
12	Type 1	4.32,5.52	5.76,4.98	2.79,2.92	0.22
	Type 2	3.84,4.96	6.69,5.28	1.65,3.33	0.07
13		8.89,11.04	<1, <1	2.68,2.61	3.3
14		5.57,7.84	6.8,4.1	1.8,2	0.22
15		5.02,8.92	<1, <1	6.22,5.23	0.17
Dual bandpass filter		1.32, 2.47	7.5,4.85	1.3,1.8	0.039

Table 3: Triple bandpass filter comparison table

References	Center frequency (GHz)	Fractional bandwidth (%)	Insertion loss (dB)	Electrical size (λ_g^2)
16	8.63,9.86,11.62	2.4,3.3,2.5	0.8,1.4,1.7	0.97
17	1.575,2.4,3.45	10,5,11	0.7,1.14,0.3	0.54
18	1.93,3.6,4.89	19.2, 11.6,2.86	-	0.076
19	1.96,2.6,3.9	5,11,3	1.5,0.6,1.83	0.41
20	2.1,5.7,7.3	57.1,8.7,4.1	0.8,1.6,2.4	0.073
Triple bandpass filter	1.29,2.27,3.21	1.16,11.4,1.86	1.6,1.3,1.8	0.06

4. Conclusion

A compact single bandpass filter using novel Triple concentric complementary split-ring resonator structure, gaps, and via along the microstrip line was proposed. The proposed single bandpass filter is useful for s band radar applications. The proposed filter has better insertion loss and sensitivity with a compact size. By adding spiral resonators to the proposed single bandpass filter design configuration, the dual bandpass filter was designed, fabricated, and measured resulting in a compact size, better insertion loss, and sensitivity. The proposed dual bandpass filter is useful for GPS and WLAN applications. The compact triple bandpass filter was designed using a stepped impedance microstrip line and T-shaped stubs for the dual bandpass configuration. The Triple bandpass filter is useful for GPS, WLAN, and WiMAX applications.

References

[1] R. Marques, F. Martin, and M. Sorolla, “Metamaterials with Negative Parameters: Theory, Design, and Microwave Applications”. Hoboken, NJ: Wiley, 2008.

[2] Cheng Tan, Yu Wang, Zhong Ming Yan et.al., "Superconducting filter based on split ring resonator structures", IEEE Transactions on Applied Superconductivity, **29**(4), 1-4, 2019, doi:10.1109/TASC.2019.2891017.

[3] A. K. Horestani, M. D. Sindreu, J. Naqui et.al., “S-Shaped complementary split ring Resonators and their application to compact differential bandpass filters with common-mode suppression”, IEEE Microwave and Wireless Components Letters, **24**,(3), 149-151, 2014, DOI: 10.1109/LMWC.2013.2291853.

[4] Y. Cheng, L. Zeng, W. Lu, “A compact CSRR- based dual-mode patch bandpass filter”, IEEE conference, November 2015, DOI: 10.1109/IMWS-AMP.2015.7325010.

[5] Y. Zheng, Y. Zhu, Yuandan “Compact hybrid bandpass filter using SIW and CSRRs with wide Stopband rejection”, IEEE conference, 2020, DOI: 10.1109/APMC47863.2020.9331361.

[6] S. Chandra “X-Band Metamaterial Bandpass Filter Design”, International Journal of Engineering Research & Technology (IJERT), **10** (5), 1004-1006, 2021, ISSN: 2278-0181.

[7] R. L. Defitri and A. Munir, "X-band microstrip narrowband BPF composed of the split ring resonator", Progress in Electromagnetic Research Symposium (PIERS), 3468-3471, 2016, DOI: 10.1109/PIERS.2016.7735347.

[8] A. A. Ibrahim, M. A. Abdalla, A. B. Abdelrahman, “Wireless bandpass Filters build on metamaterials” Microwaves & RF.**57**(5), 1-7, 2018, ISSN: 07452993.

[9] B. Fellah, N. Cherif, M. Abri et.al., “CSRR-DGS bandpass filter based on half mode substrate integrated waveguide for X-Band applications”, Advanced Electromagnetics, **10**(3), 39-42, 2021, doi:10.7716/aem. v10i3.1782.

[10] R. Keshavarz and N. Shariati, “Low profile metamaterial bandpass filter loaded with 4-Turn complementary spiral resonator for WPT applications”, IEEE conference, 50-53, 2020, ISBN: 9781728160443.

[11] S. Moitra, S. Nayak, R. Regar et.al., “Circular complementary split-ring resonators (CSRR) based SIW BPF”, Second International Conference on

Advanced Computational and Communication Paradigms (ICACCP), 2, 1-5, 2019, ISBN: 9781538679890.

[12] H.Y. Gao, Z. X. Tang, X. Cao, et.al., "Compact dual-band SIW filter with CSRRs and complementary spiral resonators", Microwave and Optical Technology Letters, **58**(1), 1-4, 2016, DOI: 10.1002/ mop.29482.

[13] A. Y. Rouabhi, M. Berka, A. Benadaoudi et.al., "Investigation of dual-band bandpass filter inspired by a pair of square coupled interlinked asymmetric tapered metamaterial resonator for X-band microwave applications", Indian Academy of Sciences, 2022, DOI: 10.1007/s12034-022-02693-6.

[14] G. Soundarya and N. Gunavathi, "Compact dual-band SIW bandpass filter using CSRR and DGS structure resonators", Progress in Electromagnetics Research Letters, 101, 79–87, 2021.

[15] M. Berka, H. A. Azzeddine, A. Bendaoudi et.al., "Dual-band bandpass filter based on Electromagnetic coupling of twin square metamaterial resonators (SRRs) and complementary resonator (CSRR) for wireless communications", Journal of Electronic Materials, **50**(8), 4887-4895,2021, DOI: 10.1007/s11664-021-09024-1.

[16] F. Gongora, A. E. Martynyuk, J.R.Cuevas et.al, "Independently tunable closely spaced triband frequency selective surface unit cell using the third resonant mode of split ring slots", IEEE Access, **9**,105564-105576, 2021, DOI: 10.1109/ACCESS.2021.3100325.

[17] M.U. Rahman and J.D. Park, "A compact tri-band bandpass filter using two stub-loaded dual-mode resonators", Progress in Electromagnetics Research M, **64**, 201–209, 2018, DOI: 10.2528/pierm17120404.

[18] N. Kumar and Y. K. Singh, "Compact tri-band bandpass filter using three stub-loaded open-loop resonators with wide stopband and improved bandwidth response", Electronics Letters, **50**(25), 1950–1952, 2014.

[19] M. Weng, M. H. Hsu, C. W. Lin et.al, "A simple method to design a tri-bandpass filter using open-loop uniform impedance resonators", Journal of Electromagnetic Waves and Applications **34**(1), 103-115, 2020, DOI: 10.1080/09205071.2019.1689181.

[20] A. Kumar, D. K. Choudhary, and R. K. Chaudhary, "Metamaterial tri-band bandpass filter using meander-line with rectangular-stub", Progress in Electromagnetics Research Letters, **66**, 121–126, 2017.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 4, 98-104 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Performance Adjustment Factor for Fixed Solar PV Module

Kelebaone Tsamaase[*,1], Japhet Sakala[1], Kagiso Motshidisi[2] , Edward Rakgati[3] , Ishmael Zibani[1], Edwin Matlotse[1]

[1]*Department of Electrical Engineering, Faculty of Engineering and Technology, University of Botswana, P/Bag UB0061,Gaborone, Botswana*

[2]*School of Computing and Information Systems, Botswana Accountancy College, P/Bag 00319, Gaborone, Botswana*

[3]*Energy Division, Research and Innovation, Botswana Institution for Technology Research and Innovation, P/Bag 0082, Gaborone, Botswana*

ARTICLE INFO

Article history:
Received: 10 May, 2022
Accepted: 30 July, 2022
Online: 16 August, 2022

Keywords:
Photovoltaic cell
Solar energy
Single diode model

ABSTRACT

There are different factors which contribute to the amount of output power which can be delivered by solar photovoltaic (PV) module at any time of the year. The factors include but not limited to solar irradiation, ambient temperature, relative humidity, wind velocity, position of sun in the sky, geographical position of installed solar PV module and others. Apparent position of the sun in the sky contribute to the amount of electromagnetic radiation from the sun reaching the module's surface area. With the sun further away from the module and irradiance reaching the module surface area at an angle non perpendicular to the surface leads to low output power delivered by the module. In southern hemisphere the PV module experience high output power around November/December which are summer months and low output power around June/July which are winter months. This paper develops performance adjustment factor of fixed solar PV module to adjust PV module output power such that the PV system can deliver required amount of power during winter season. The results show that the value of performance adjustment factor for fixed solar PV module or system was established and can be used to adjust performance or output power for winter periods.

1. Introduction

Performance of Solar photovoltaic (PV) module plays an important part in the overall performance of solar photovoltaic energy system or solar array in particular. To improve performance of the system various studies have confirmed that many factors need to be looked into. This include system design whereby a PV tracking system was found to yield more output power than fixed installation system. However, due to other technical considerations of the system design the fixed installation system is the most commonly preferred. Other considerations are geographical position of the system expressed in terms of latitude and longitudinal points. Weather conditions such as availability of sunlight, wind velocity, temperature, rainfall, technology applied during manufacture of modules, which lead to different types of modules such as monocrystalline, polycrystalline, thin film, and others. To research further on factors and their effects on PV module performance this paper is an extension of work originally presented in 2021 10th International Conference on Renewable Energy Research and Application (ICRERA) which was dealing with maximum power output and voltage profiles as a response to varying irradiance and ambient temperature [1]. Single diode model shown in Figure 1 [2-6] was simulated using Simulink model shown in Figure 2 to generate voltage profile, power profile and profile for maximum power of a fixed installation as it varies from January to December. The outcome of the maximum power profile showed that the maximum power output is lowest during the months of June/July which is winter season in southern hemisphere and increased gradually when summer months (November/December) approached. While the temperature and irradiance were considered when producing profiles, the output power profile followed closely the profile of solar irradiance during the same period, indicating that irradiance is a dominant

*Corresponding Author: Kelebaone Tsamaase, tsamaase@ub.ac.bw

factor in generating output power by PV module. For the northern hemisphere, the opposite takes place whereby solar irradiation profile picture was observed with the maximum monthly global irradiance recorded during May, June and July months [7]. In this paper two profiles of maximum output power are established. One profile is for the lowest values of maximum power determined in a five year period. The other profile is for the highest values of maximum power profiles established in the same five year period. The two profiles are used to come up with the performance adjustment factor for fixed PV module or PV system. The performance adjustment factor is used to adjust PV system design output to a higher value, such that the system meets the expected output throughout the year irrespective of the time of the year at which it is designed or at which data used in the design was recorded. The rest of the sections of this paper are arranged in the following manner: Section 2 deals with definitions of terms. Section 3 deals with establishment of profiles for maximum output power and voltages. Section 4 presents methodology for generating profiles. Section 5 deals with the results and discussions and section 6 gives conclusions.

2. Nomenclature

k	Boltzmann constant (1.381×10^{-23} J/K)
q	Electron charge (1.602×10-19 C)
I_{pv}	photovoltaic output current
I_{ph}	photon current, at standard test conditions (G, 1000 W/m^2, Temperature $25^oC, AM$ 1.5)
G	Solar irradiance
G_0	solar irradiance (W/m^2) at standard test conditions (G, 1000 W/m^2, Temperature $25^oC, AM$ 1.5))
(I_{SCr})	short-circuit current of a PV module at reference condition, ((STC) 1000 W/m^2, Temperature $25^oC, AM$ 1.5)
K_i	short circuit temperature coefficient at module reference conditions
I_s	reverse saturation current
I_{sh}	shunt current
V_{PV}	output voltage of PV module
T	ambient temperature (Kelvin)
n	Diode ideality constant
N_S	number of cells in series
I_D	diode current
R_S	series resistance
R_{Sh}	shunt resistance
F_p	performance adjustment factor
P_{max}	Maximum power output of PV module
$P_{\max\,(max)}$	Highest value of maximum-power output of the PV module recorded in a year
$P_{\max\,(min)}$	Minimum value of maximum-power output of the PV module recorded in a year
$P_{out(initial)}$	Output power of the PV module or PV system Design
$P_{out(final)}$	Adjusted output power of the MV module or PV systems.

3. Establishing Maximum Power Profiles for Winter and Summer Seasons

Maximum power output of the PV module for the winter and summer periods are computed by use of single diode model shown in Figure 1 [2]-[6].

The parameters in the model are current source which represents photon current (I_{ph}), diode current (I_d) passing through the diode which is connected across the source terminals, a shunt resistor (R_{sh}) which is connected across the current source terminals and represents leakages in form of leakage current (I_{sh}) experienced by the cell, and a series resistor (R_s) representing resistances due to solder bonds, metallization of cell, terminations of junction box, and emitter and base regions [2, 8, 9].

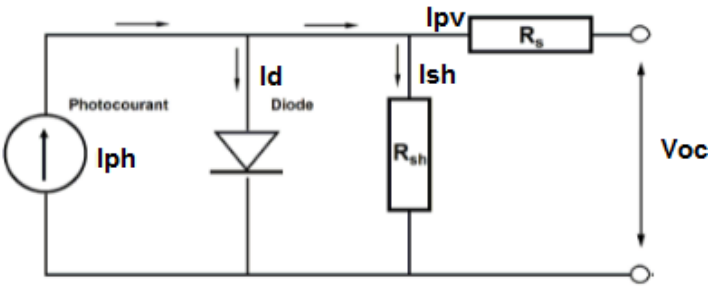


Figure 1: Single diode model representation of a PV cell

The cell model was used to represent PV module because the module is made up of electrical and mechanical interconnection of solar cells. The model in Figure 1 was simulated using Simulink and its corresponding Simulink model is shown in Figure 2 [1].

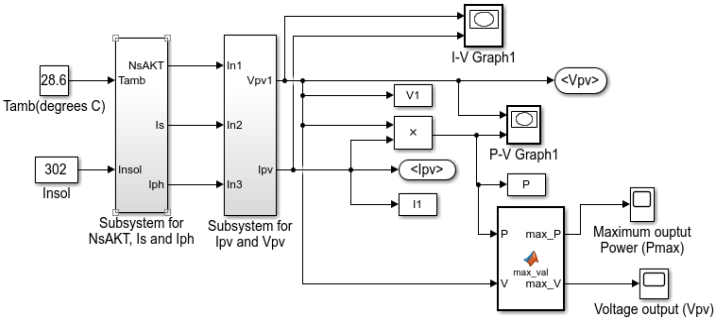


Figure 2: Simulink model for computing maximum power output

Considering expressions of different electrical parameters of the model the value of output current I_{pv} expressed in mathematical form is presented in (1) [10, 11, 12, 13, 14, 15, 16].

$$I_{pv} = I_{ph} - I_D - I_{sh} \quad (1)$$

Where I_{ph}, I_D and I_{sh} are shown in (2), (3) and (4) respectively.

$$I_{ph} = \left[I_{SCr} + K_i(T - 298)\frac{G}{G_0}\right] \quad (2)$$

$$I_D = I_S\left[e^{\frac{q(V_{PV}+IpvR_S)}{N_S nTk}} - 1\right] \quad (3)$$

$$I_{sh} = \frac{V_{PV}+R_S I_{PV}}{R_{sh}} \quad (4)$$

4. Methodology for generating maximum power output (P_{max}) profiles

The following data was used to establish P_{max} profiles. PV module characteristics, ambient temperature and irradiance from five different weather stations in Botswana [17]. Dataset for the two weather variables covers a five year period, from 2009 to 2013.

The characteristic parameters of a photovoltaic module were obtained from a module installed at a weather station called Renewable Energy Research Centre (CERC) at the University of Botswana (UB) in capital city of Republic of Botswana, Gaborone and from similar model online. The geographical positioning system location coordinates of Gaborone are 24° 39' 11.7252" S and 25° 54' 24.4512" E [18]. The parameters are detailed in Table 1.

Table 1: Photovoltaic Module Parameters

Name	**Solaire Direct Technologies**
Model	**SD ECO PLUS 150 W**
Electrical Ratings at STC: 1000 W/m^2; AM 1.5 spectrum; Temperature 25 ^oC	
Peak Power	150 W_p
ΔP_{max}	$\pm$ 2.5 W_p *with* tolerance 1%
Warranted minimum P_{max}	147.5 (tolerance 1%)
Voltage (V_{mp})	18.0 V
Current (I_{mp})	8.05 A
Open Circuit Voltage (V_{OC})	22.6 V
Short Circuit Current (I_{SC})	8.72 A
Maximum System Voltage	1000 V

Ambient temperature and solar irradiance were obtained from the following weather stations. Sir Seretse Khama Airport (SSKI) Weather Station. The weather station is located in the Capital City Gaborone, in the south-eastern part of the country with geographical positioning System (GPS) coordinates of 24° 39' 11.7252" S and 25° 54' 24.4512" E. Kasane weather station, located in the town of Kasane at the northern part of Botswana. Its GPS location coordinates are 17° 48' 10.944" S and 25° 8' 56.0364" E. Tshane weather station, located in the town of Tshane in the southern part of Botswana. Its GPS location coordinates are 24° 1' 7.032" S; 21° 51' 33.084" E. Maun weather station, located in the town of Maun at the north-western part of Botswana. Its GPS location coordinates are 19° 59' 23.6004" S; 23° 25' 24.888" E. Shakawe weather station, located at the village of Shakawe with the following coordinates 18° 21' 46.728" S 21° 50' 51" E [18].

For each weather station, ambient temperature and solar irradiance for a five year period were considered. The temperature was considered because besides irradiance it is one of the factors affecting PV module electrical power output [19, 20]. Data provided from the weather stations was for daily recordings and using that data the monthly average values of temperature and irradiance were calculated. The monthly average values were used as input variables into the Simulink model to compute monthly values of P_{max} for each month. Figure 3 shows corresponding graphical representation of power output against the voltage obtained from the model. From each graph $P_{\max}$ obtained or indicated were recorded [1]. Similar procedure was followed to generate monthly P_{max} for the years 2009 to 2013.

To plot profiles, the lowest values of maximum power ($P_{\max\,(\min)}$) from each year were plotted. Same procedure was followed to plot a profile for highest values of maximum power ($P_{\max\,(\max)}$)from each year. The difference between two profiles (ΔP_{max}) was determined and used to formulate a factor referred to as performance adjustment factor. This factor gives a relationship between P_{max} in summer season and P_{max} in winter season in southern hemisphere.

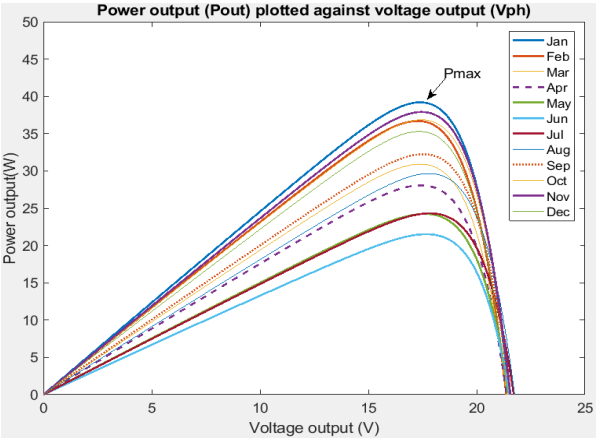


Figure 3: Photovoltaic module power output plotted against voltage for different months of the year

5. Results and Discussion

5.1. Sir Seretse Khama Airport (SSKI) Weather Station

The irradiance and temperature data recorded at the station was used in the simulation model to get values of power output and voltage for the years 2009, 2010, 2011, 2012 and 2013. The Figures 4 and 5 show how output power and voltage values respectively were determined and recorded by Simulink model, with Tables 2, 3, 4, 5 and 6, of corresponding years respectively showing values obtained from Simulink model. From the tables the maximum and minimum values of output power recorded each year were obtained and recorded in Table 7. Also shown in Table 7 are corresponding values of voltages and months of the year at which such minimal and maximum values were recorded.

To establish five year-period profiles for highest maximum value of power ($P_{\max\,(\max)}$) and lowest maximum value of power ($P_{\max\,(\min)}$) as shown in Figure 6, data from Table 7 was used. Similarly voltage profile was also generated.

Table 2: 2009 Maximum power and voltage output determined with Simulink Model

Month	Average Temperature	GHI	P_{max} (W)	V_{pv} (V)
Jan	24.2	264.0	34.6	17.6

Feb	24.9	259.0	33.8	17.5
Mar	20.8	234.0	30.8	17.7
Apr	19.4	235.0	31.1	17.8
May	18.1	188.0	24.6	17.6
Jun	14.0	150.0	19.5	17.7
Jul	10.0	187.0	25.2	18.2
Aug	14.7	226.0	30.4	18.1
Sep	19.8	254.0	33.8	17.8
Oct	22.9	252.0	33.1	17.6
Nov	22.4	271.0	35.8	17.7
Dec	25.2	316.0	41.7	17.7

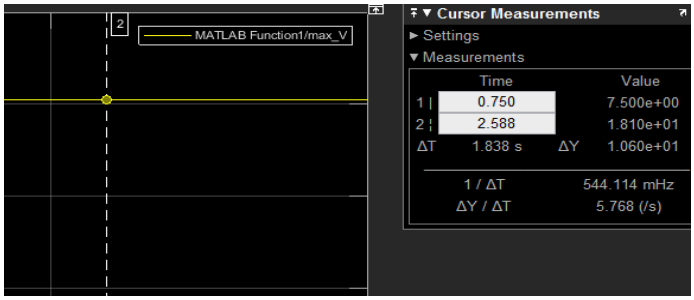


Figure 4: Simulink plotting output voltage

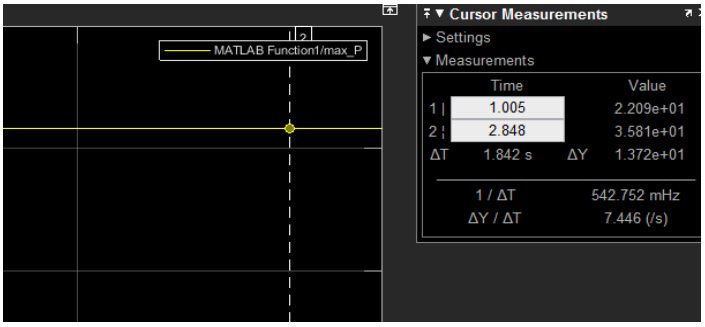


Figure 5: Simulink plotting maximum power

Table 3: 2010 Maximum power and voltage output determined with Simulink Model

Month	Average Temperature	GHI	P_{max} (W)	V_{pv} (V)
Jan	25.3	261.0	34.0	17.5
Feb	25.4	279.0	36.5	17.5
Mar	24.0	235.0	30.5	17.5
Apr	20.6	163.0	20.8	17.3
May	16.6	173.0	22.6	17.7
Jun	12.1	181.0	24.2	18.0
Jul	13.9	182.0	24.1	17.9
Aug	15.6	229.0	30.7	18.0
Sep	21.5	275.0	36.5	17.8
Oct	25.5	304.0	40.0	17.6
Nov	25.4	266.0	34.7	17.5
Dec	24.6	290.0	38.2	17.6

Table 4: 2011 Maximum power and voltage output determined with Simulink Model

Month	Average Temperature	GHI	P_{max} (W)	V_{pv} (V)
Jan	24.0	256.0	33.5	17.5
Feb	25.2	287.0	37.6	17.6
Mar	25.9	257.0	33.3	17.4
Apr	17.8	176.0	22.9	17.6
May	12.0	191.0	25.6	18.1
Jun	13.8	184.0	24.4	17.9
Jul	10.4	187.0	25.2	18.2
Aug	12.7	224.0	30.4	18.2
Sep	20.7	275.0	36.6	17.8
Oct	22.2	287.0	38.1	17.8
Nov	24.0	287.0	37.8	17.7
Dec	27.0	240.0	30.8	17.3

Table 5: 2012 Maximum power and voltage output determined with Simulink Model

Month	Average Temperature	GHI	P_{max} (W)	V_{pv} (V)
Jan	25.3	320.0	42.3	17.7
Feb	26.6	297.0	38.8	17.5
Mar	23.9	266.0	34.9	17.6
Apr	19.8	230.0	30.3	17.7
May	17.1	206.0	27.2	17.8
Jun	14.0	180.0	23.8	17.9
Jul	14.0	195.0	26.0	18.0
Aug	17.0	225.0	30.0	17.9
Sep	21.0	267.0	35.5	17.8
Oct	21.9	260.0	34.3	17.7
Nov	25.9	316.0	41.6	17.6
Dec	23.3	290.0	38.4	17.7

Table 6: 2013 Maximum power and voltage output determined with Simulink Model

Month	Average Temperature	GHI	P_{max} (W)	V_{pv} (V)
Jan	26.9	293.0	38.2	17.5
Feb	26.8	302.0	39.5	17.5
Mar	24.5	275.0	36.1	17.6
Apr	21.7	220.0	28.7	17.6
May	16.6	203.0	26.9	17.8
Jun	14.5	182.0	24.1	17.9
Jul	16.0	192.0	25.4	17.8
Aug	17.5	223.0	29.6	17.9
Sep	23.9	255.0	33.3	17.5
Oct	24.0	292.0	38.5	17.7
Nov	25.7	312.0	41.1	17.6
Dec	25.5	247.0	32.0	17.4

Table 7: Yearly highest ($P_{max\ (max)}$) and lowest ($P_{max\ (min)}$) values of maximum power and average voltage output for SSKI

Year	$P_{\max(max)}$ (W)	Mth	$P_{\max(min)}$ (W)	Mth	Average V_{pv} (V)
2009	41.7	Dec	19.5	Jun	17.8
2010	40.0	Oct	20.8	Apr	17.7
2011	38.1	Oct	22.9	Apr	17.8
2012	42.3	Jan	23.8	Jun	17.7
2013	41.1	Nov	24.1	Jun	17.7

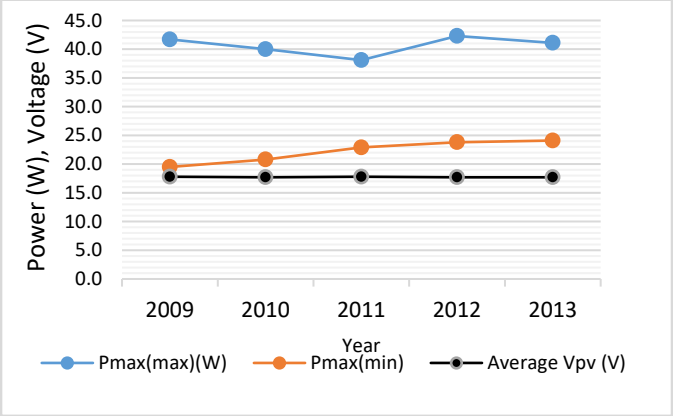


Figure 6: Profiles of highest and lowest values of maximum power and voltages for SSKI Weather station

5.2. Kasane, Maun, Shakawe and Tshane Weather Stations

To develop profiles for Kasane, Maun, Shakawe and Tshane weather stations the approached described in SSKI weather station at section 5.1 was followed. Simulated values of power output and voltages arising from irradiance and temperature values for different months and years are shown in Table 8. Corresponding power and voltage profiles of the same weather stations are shown in Figures 7 to 10 respectively.

Table 8: Highest and lowest values of maximum power and average voltage output

Place	Yr	Highest value of Maximum power, $P_{\max(max)}$ (W)	Mth	Lowest value of Maximum power, $P_{\max(min)}$ (W)	Mth	Average Output voltage, V_{pv} (V)
Tshane	2009	42.5	Dec	22.3	Jun	17.7
	2010	43.6	Dec	25.3	Jun	17.7
	2011	41.6	Nov	24.5	Jun	17.7
	2012	40.1	Nov	24.4	Jun	17.7
	2013	42.7	Nov	24.3	Jun	17.7
Maun	2009	37.5	Nov	25.4	Jun	17.6
	2010	39.2	Oct	25.6	Jul	17.5
	2011	40.6	Oct	27.8	Jun	17.6
	2012	38.5	Jan	27.0	Jun	17.6
	2013	39.9	Feb/Oct	27.2	Jun	17.6
Kasane	2009	36.9	Jan	27.2	Jun	17.6
	2010	38.9	Sept	25.0	Jul	17.6
	2011	38.9	Sept	28.8	Jun	17.6
	2012	38.1	Sept	28.4	Jun	17.6
	2013	38.9	Oct	28.8	Jun	17.6
SSKI	2009	41.7	Dec	19.5	Jun	17.8
	2010	40.0	Oct	20.8	Apr	17.7
	2011	38.1	Oct	22.9	Apr	17.8
	2012	42.3	Jan	23.8	Jun	17.7
	2013	41.1	Nov	24.1	Jun	17.7
Shakawe	2009	36.6	Nov	27.0	Jun	17.6
	2010	38.8	Oct	26.1	Jul	17.6
	2011	40.9	Oct	29.3	Jun	17.7
	2012	38.8	Sept	28.4	Jun	17.7
	2013	39.5	Oct	28.6	Jun	17.6

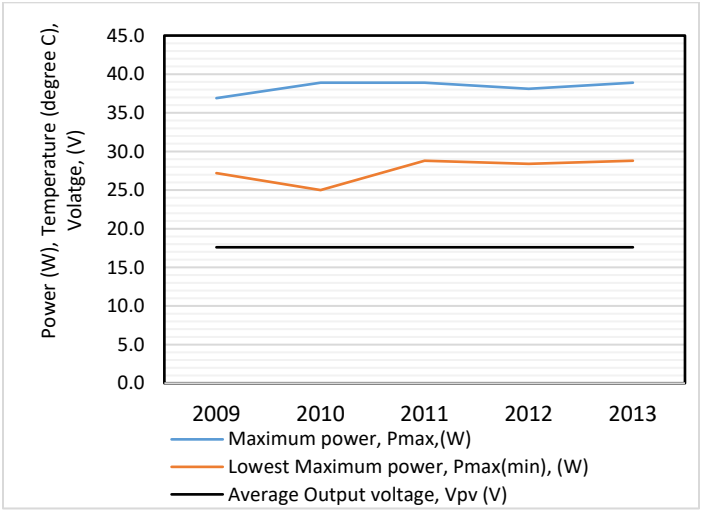


Figure 7: Profiles of highest and lowest values of maximum power and voltages for Kasane Weather station

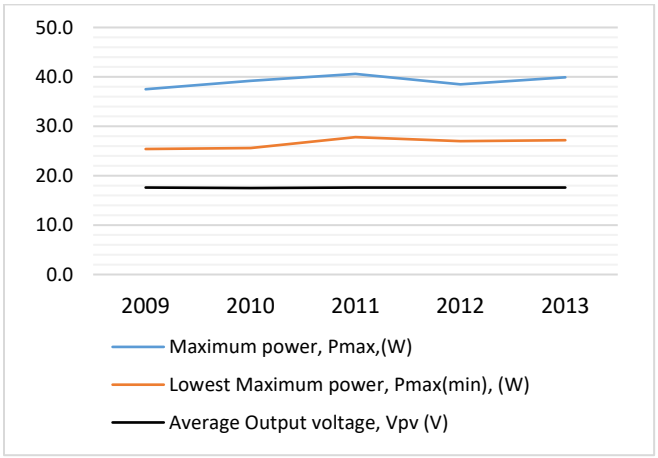


Figure 8: Profiles of highest and lowest values of maximum power and voltages for Maun Weather station

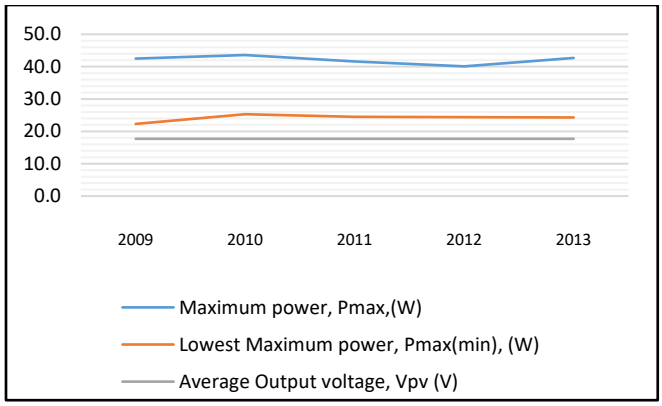


Figure 9: Profiles of highest and lowest values of maximum power and voltages for Tshane Weather station

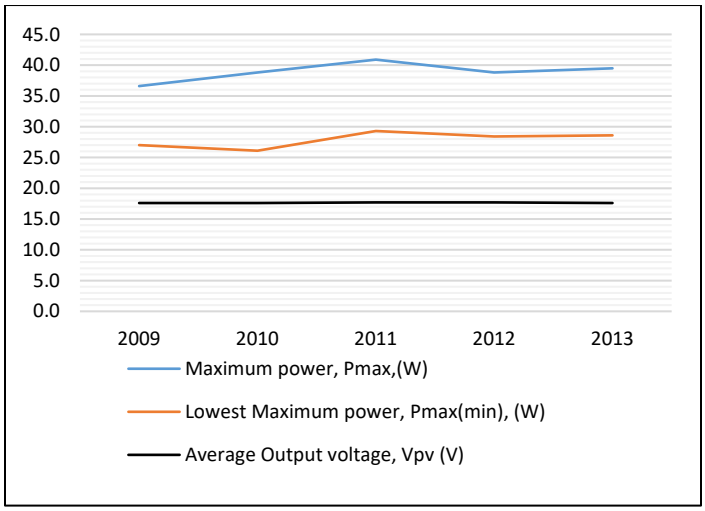


Figure 10: Profiles of highest and lowest values of maximum power and voltages for Shakawe Weather station

5.3. Five year national average profiles

Data from five weather stations shown in table 8 was used to come up with yearly averages shown in table 9, where also the yearly averages were computed to come up with a five year average value for $P_{\max(\max)}$ and $P_{\max(\min)}$ respectively. Table 9 was used to generate figure 11 where average values of highest value of maximum power $P_{\max(\max)}$, lowest value of maximum power $P_{\max(\min)}$ and voltage for each year were plotted to come up with national average profiles. Profiles in figure 11 were refined further as shown in figure 12 by plotting average values stretching on a five year long period.

Table 9: Yearly averages for maximum power and voltage output

Year	Average Highest Maximum power, $P_{\max(\max)}$ (W)	Average Lowest Maximum power, $P_{\max(\min)}$ (W)	Average Output voltage, V_{pv} (V)
2009	39.04	24.28	17.7
2010	40.10	24.56	17.6
2011	40.02	26.66	17.7
2012	39.56	26.40	17.7
2013	40.42	26.6	17.6
Average	39.8	25.7	17.7

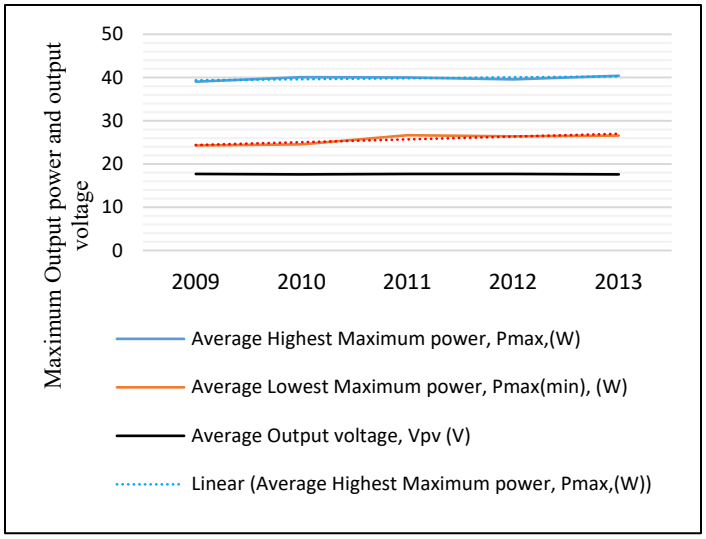


Figure 11: Five year period national averages of maximum power and voltage output

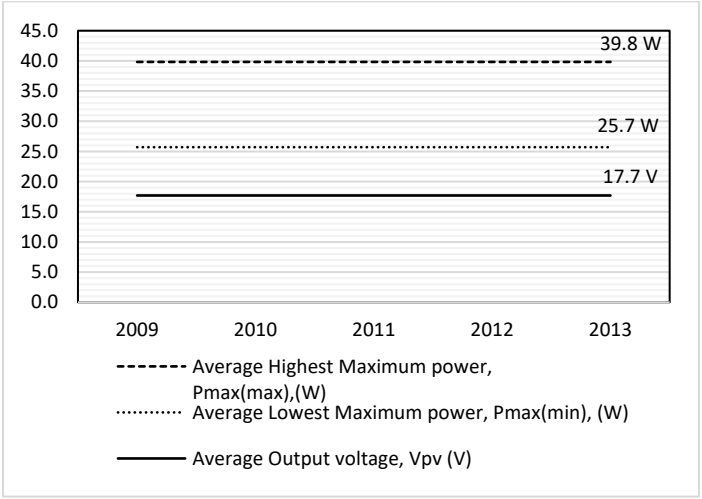


Figure 12: Refined profiles for five year period national averages of maximum power and voltage

5.4. Determining the value of performance adjustment factor, F_p

The relationship between two final average values of maximum power as shown in table 9 and figure 12 is represented by a factor herein referred to as performance adjustment factor, F_p. The purpose of this factor is to make adjustment on PV systems design output power which would have been determined or calculated during PV system design stage. The importance of adjustment is to ensure that the PV system or the PV module gives a required amount of power output throughout the year regardless

of the season or month of the year at which such system was designed. F_p is obtained as shown in (5) to (8)

$$F_p = \frac{P_{\max\,(\min)}}{P_{\max\,(\max)}} = \frac{25.7}{39.8} = 0.645728643 \quad (5)$$

$$F_p \approx 0.6457 \quad (6)$$

The final value of the solar PV system or module output power obtained after adjustment is designated, $P_{\text{out (final)}}$, and is presented as in (3)

$$P_{\text{out}(final)} = \frac{P_{out(initial)}}{F_p} = \frac{1}{F_p}(P_{out(initial)}) \quad (7)$$

where:

$$0.6457 \leq F_p \leq 1 \quad (8)$$

$P_{\text{out(initial)}}$ is the output power, before adjustment, of the fixed solar PV system or module calculated using site specific parameters (temperature and irradiation) at any time of the year.

5.5. Conditions for F_p

According to (7) and (8) if $P_{out(initial)}$ is calculated using solar irradiance and temperature values of summer season, example October/November, then to get $P_{out(final)}$ the value of F_p at this season is maximum with a value of 1.0. For the same PV system or module, to adjust its calculated output for the purpose of winter performance the $P_{out(initial)}$ should be divided by 0.6457. The $P_{out(final)}$ obtained after adjustment can now be used to resize the system to ensure that it will deliver such adjusted power in winter.

If the $P_{out(initial)}$ is calculated using solar irradiance and temperature values of winter season, the value of F_p will remain maximum throughout, that is value of 1. In this case $P_{out(initial)}$ is taken as $P_{out(final)}$. This means that the system designed or sized using winter weather data (irradiance and temperature) is capable of delivering required output power throughout the year including during summer season and there is no need for any performance adjustment

6. Conclusion

According to the results obtained the module maximum output power in summer (around November/ December) is higher than the maximum power in winter (June / July). To ensure that PV system or module produce required amount of power throughout the year a performance adjustment factor was derived to be used in calculations for determining power output in winter. Further work will involve experimental setup to refine performance adjustment factor with other variables considered.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] K. Tsamaase, J. Sakala, E. Rakgati, I. Zibani, K. Motshidisi, "Solar PV Module Voltage Output and Maximum Power Yearly profile using Simulink-based Model," in 2021 10th International Conference on Renewable Energy Research and Application(ICRERA), 26-29, 2021, doi: 10.1109/ICRERA52334.2021.9598794

[2] C.B.Honsberg, S.G.Bowden, "Photovoltaics Educationte," www.pveducation.org, 2019.

[3] H. Ibrahim, H. Anani, "Variations of PV module parameters with irradiance and temperature," in 9th International Conference on Sustainability in Energy and Buildings, 5-7, 2017, doi:10.1016/j.egypro.2017.09.617

[4] T. M Silverman, M. G. Deceglie, I. Subedi, N. J. Podraza, I. M. Slauch, V. E. Ferry, I. M Slauch, "Reducing Operating Temperature in Photovoltaic Modules," Photovoltaics, **8**(2), 532-540, 2018, doi.org/10.1109/JPHOTOV.2017.2779842

[5] H. Singh, D. Kaur, P. S. Cheema, "Optimal design of Photovoltaic Power System for a residential load," in International Conference on Inventive Systems and Control (ICISC), 19-20, 2017, doi: 10.1109/ICISC.2017.8068595

[6] Y. Wang, Y. Liu, C. Wang, Z. Li, X. Sheng, H. Lee, N. Chang, H. Yang, "Storage-Less and Converter-Less Photovoltaic Energy Harvesting With Maximum Power Point Tracking for Internet of Things," Computer-aided design of integrated circuits and systems, **35**(2), 173 - 186, 2016, doi:10.1109/TCAD.2015.2446937

[7] T. Abu Dabbousa, I. Al-Reqeb, S.Mansour, M. Ammirrul Atiqi Mohd Zainuri, "The Effect of Tilting a PV Array by Monthly or Seasonal Optimal Tilt Angles on Energy Yield of a Solar PV System," in 2021 International Conference on Electric Power Engineering (ICEPE), 23-24, 2021, doi:10.1109/ICEPE-P51568.2021.9423475

[8] P.Hao, Y. Zhang, "An Improved Method for Parameter Identification and Performance Estimation of PV Modules From Manufacturer Datasheet Based On Temperature-Dependent Single-Diode Model," Photovoltaics, **11**(6), 1446 - 1457, 2021, doi:10.1109/JPHOTOV.2021.3114592

[9] P. Natarajan, and R. Muthu, "Performance Improvement of PV Module at Higher Temperature Operation," Engineering science and technology, 2(5), 2012, doi: ISSN 2250-3498

[10] S. T. Kim, S. Bae, Y. C. Kang, J. W. Park, " Energy Management Based on the Photovoltaic HPCS With an Energy Storage Device," Industrial electronics, **62**(7), 4608 - 4617, 2015, doi: 10.1109/TIE.2014.2370941

[11] V. Tamrakar, S.C. Gupta, Y. Sawle, "Single-diode PV cell modeling and study of characteristics of single and two-diode equivalent circuit," Electrical and electronics, **4**(3), 13-24, 2015, dio: 10.14810/elelij.2015.4302,

[12] H. Park, Y. Kim, H. Kim, "PV Cell Model by Single-diode Electrical Equivalent Circuit" Electrical engineering and technology, **11**(5), 1323-1331, **11**(5): doi:10.5370/JEET.2016.11.5.1323,

[13] T. Ahmed, T. Gonçalves, M. Tlemcani, "Single diode model parameters analysis of photovoltaic cell," in 2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA), 20-23, 2016, doi: 10.1109/ICRERA.2016.7884368,

[14] B. C. Babu, S. Gurjar, "A Novel Simplified Two-Diode Model of Photovoltaic (PV) Module," **4**(4), 1156-1161, 2014, doi: 10.1109/JPHOTOV.2014.2316371

[15] N. Islam Sarkar, "Effect of various model parameters on solar photovoltaic cell simulation: a SPICE analysis," Sarkar Renewables, **3**(13), 2016, doi: 10.1186/s40807-016-0035-3

[16] H. K. Mehta, H. Warke, K. Kukadiya, A. K. Panchal, "Accurate Expressions for Single-Diode-Model Solar Cell Parameterization," Photovoltaics, **9**(3), 803-810, 2019, doi: 10.1109/JPHOTOV.2019.2896264

[17] Department of Meteorological Services, Plot No. 54216, Corner of Metsimotlhaba/Maaloso Roads, meteo@gov.bw. Botswana. 2021.

[18] Longitude and latitude finder. 2021. www.latlong.net

[19] V. Herbort, R. von Schwerin, B. Compton, L. Brecht, H. te Heesen, "Insolation dependent solar module performance evaluation from PV monitoring data," in 38th IEEE Photovoltaic Specialists Conference, 3-8, 2012, doi:10.1109/PVSC.2012.6317834

[20] M. Islam, M. Mehedi Hasan Shawon, S. Akter, A. Chowdhury, S. Islam Khan, M. Rahman, "Performance Investigation of Poly Si and Mono Si PV Modules: A Comparative Study," in International Conference on Energy and Power Engineering (ICEPE), 14-16, 2019. doi: 10.1109/CEPE.2019.8726598

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 1, 101-107 (2023)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Conception and Simulation of an Electronic Nose Prototype for Olfactory Acquisition

Mostapha Harmouzi *, Aziz Amari, Lhoussaine Masmoudi

Conception and Systems Laboratory, Faculty of Sciences, Mohammed V University in Rabat, Morocco

ARTICLE INFO

Article history:
Received: 01 September, 2022
Accepted: 08 January, 2023
Online: 07 February, 2023

Keywords:
e-nose
Sensing chamber
Sensor array
Gas compartment

ABSTRACT

The "Electronic Nose" approach, which is exclusive to gas measurement systems, uses gas sensors as odor detectors. Design faults exist in the existing electronic nose (e-nose) chamber, such as its large volume, difficult construction, etc. In order to obtain measurements in a satisfactory state, we want to create a gas chamber that can provide favorable conditions for the sensor array, taking into account the ideal gas flow morphology and detector placement. To describe and identify the design capable of offering the best performance for a genuine idea, the e-nose chamber was created using ParaVIEW simulation and FreeCAD conception. According to the results, the spherical sensing container with connections from both pipes in a tangential arc style gives the highest performance in terms of turbulence reduction, in that case, we are printing this chamber and put it in a gas flow prototype to see the performance of the quality measurement of the sensors inside it, and the result shows that these sensors have good acquisition responses by testing the homogeneity distribution inside the chamber.

1. Introduction

The human olfactory system, which consists of the region from the olfactory epithelium to the olfactory brain, is a sensory system for the detection of odor molecules. Olfaction is the term for the sensory function of smell that allows humans to detect and identify a wide range of odorants. Odorants enter the mucus-covered olfactory epithelium in the nasal cavity, bind to the olfactory receptors present in the cilia of the olfactory sensory neurons, and then send odorant information to the brain [1].

Odorants are volatile, hydrophobic compounds that have molecular weights of less than 300 Daltons. The largest known odorant to date is labdane that has a molecular weight of 296 [2].

The sense of smell is a very sensitive organ that can detect even the smallest amounts of substances. It is estimated that only 2 % of the volatile compounds available in a single sniff will reach the olfactory receptors, and as few as 40 molecules of some mercaptans are sufficient to perceive an odor [3].

Instead than being caused by a single chemical, most odors are created by combining hundreds of different odorants. Perceptual fusion results from the tendency of individual components to harmonize or blend together in mixes. The maximum number of components that humans can detect in a combination of odors is three to four [3].

A system that addresses the demand for increasing production or quality control against adulterations is crucial in the fields of food and agricultural product quality control [4]. The most crucial component of the overall e-nose system is the sensor array, which responds differently to different gas molecules [5]. The detecting system of an e-nose is composed of a sensing element, chamber, and sensor array. In order to maximize interaction between the sensor array and volatile substances, the sensing chamber can be a dynamic closed-loop area. The sensing chamber may also be a static area that prevents gas leaks during the detecting stage, which might have a negative impact on the interpretation of the data [6].

One of the biggest issues with an e-detection nose's stage efficiency is poor sensor response to volatile chemicals. One of the most frequent issues related to subpar sensor response signal performance is also thought to be the longer time it takes for sensors to attain steady state.

The size and shape of the detecting chamber, together with the placement of the intake and outflow gases, are all factors that might significantly affect the performance of the sensor's response. The gas flow and concentration in the detecting chamber are impacted by these variables.

* Corresponding Author: Mostapha Harmouzi, Email: mostapha7.harmouzi@gmail.com

2. Technology for an e-nose system

An e-nose system is a device with a sensor array that is intended to distinguish between and identify complex scents. Three primary components make up e-nose instruments: a sample handling system, a detection system or sensor array, and data processing models, including categorization and prediction models. The sensor array consists of general-purpose sensors, often metal oxides, that have undergone various chemical treatments. When the volatile molecules are exposed to the sensor array, a particular smell print (or fingerprint) is instantaneously produced from them. To build the database and train a new pattern recognition system that is used to categorize and identify novel scents, patterns or fingerprints from known odors are employed [7].

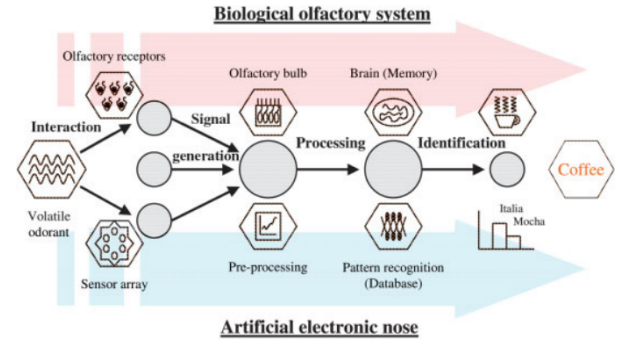


Figure 1: Basic diagram showing the analogy between biological and artificial noses [8].

Electronic nose is a quick and powerful approach that does not require any specific sample preparation and can determine a product's whole volatile profile. Devices called e-noses are composed of several sensors and may mimic the sense of smell.

As shown by the similarity between biological and artificial noses in Fig.1, e-nose is designed as a match-model for the natural nose and includes the many phases between a volatile odorant and its detection, namely: interaction, signal generation, processing, and identification. A sensor array, electrical interface circuitry, and a pattern-recognition device that serves as a signal processing system are all included in the system. However, a more straightforward model based on a collection of sensors, signal amplification through pre-processing [9], and a pattern recognition system aids in better understanding and illustrating how the nose works [8]. Pattern recognition methods include all statistical and neural techniques for the classification and recognition of odors.

Machine learning is used in the electronic nose. The core concept is that the measurement system learns from training measurements to achieve the desired goal rather than the scientist or application engineer adapting the measurement system to a specific measurement activity [6].

2.1. Measurement conditions stability

The electronic nose's seeming simplicity has made processing samples and delivering sample gas easier. Furthermore, despite the fact that many electronic nose scientists have academic backgrounds in mathematics or electronics, they lack knowledge of chemical-analytical concerns. The same is true for the measurement systems. Data from a basic electronic nose measuring system, such as one that merely consists of a few gas sensors in a straightforward chamber, will be noisy and have a high degree of measurement uncertainty [6].

Therefore, the only means of achieving stability are controlled sampling methods and the transfer of volatiles to the measurement apparatus. The electronic nose must also be as technologically advanced as is feasible to block sources of noise and outside impacts.

2.2. Detection measurement

The three main parts of an electronic nose are a detecting system sample, a processing system, and a calculation system. The sample handling technique enables the generation of a sample's headspace (volatile chemical smells [6]). The detecting component of the electronic nose is then injected with this headspace. The sample handling method is essential for ensuring constant operational conditions [10].

Our aim study is based on the detection system sample, because firstly before entering to the processing and computing phases we should make a good measurement from the sensor array in e-nose system by using the procedure below [4] (Figure 2):

- To suppress any strange gases in the cell, first of all to do is injecting the gas vector (oxygen, nitrogen, or air) into the chamber, which is controlled by opening the upper electrical valve (the strange gases are considered as noise in the measurement).
- When to confirm that the chamber is cleaned, then flowing the gas vector into the sample to through the chamber sensor array by opening the downer electro-valve and closing the first one.
- Recording the data from any acquisition electronic board and plotted to see the measurement of the sensor array.

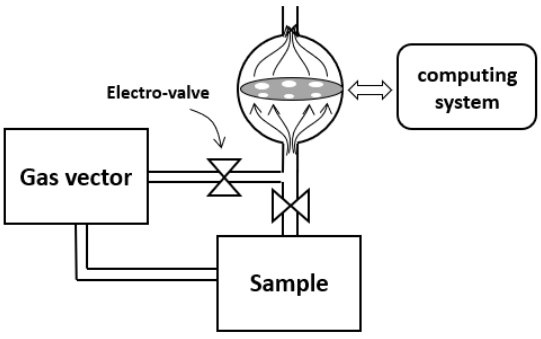


Figure 2: Procedure sensing of an e-nose system [4].

3. Prototype system design

In order to make sure that gas sensors offer reliable readings, we look at the process design of the sensing chamber and the notion of hydrodynamic and fluid flow in this section [11]. We will assess the form, taking into account the sizes depend on the number of sensors employed, and temperature and humidity impact.

3.1. Specification

Two fluid systems types are, static and dynamic systems. According to their names, static systems are those in which the fluid is at rest and dynamic systems are those in which the fluid is in motion. [12].

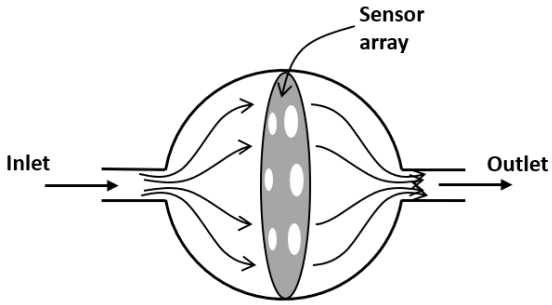


Figure 3: Sensor array inside the gas chamber of e-nose [4].

A dynamic system maintains gas concentrations that are utilized for processing. As seen in Figure 3, the chamber's design should allow for a more effective dispersion of the gas introduced within. To do this, a variety of sensors with the ability to simulate fragrance must be used [13].

3.2. Chamber design

With a single inlet and outflow, the chamber was meant to resemble a spherical chamber (Figure 3). This shape may be compared to another in which the detecting system performed poorly because the stationary zone was big, for instance because the rectangular shape's borders trapped the gas (Figure 4). As a result, the chamber's sensitivity to compost emissions was poor, and its signal-to-noise ratio was less stable [5].

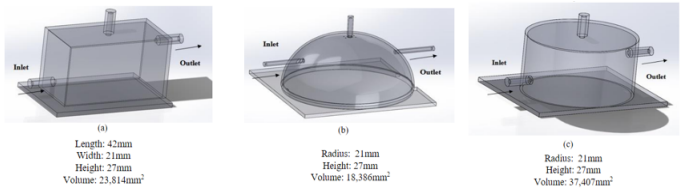


Figure 4: Shapes and volumes of the (a) rectangular, (b) hemisphere and (c) cylindrical sensing chambers [5].

In [5], the author concluded from a comparison of various chambers that the hemisphere chamber, shown in Figure 4, which they used to assess the sensor's time response and stagnant region, had the best performance. For each sensor chamber and the alcohol content, the average of three readings was obtained in 100 seconds [5]. When there was just clean air in the sensing chambers at first, the alcohol content remained consistent for all of them. The three sensing chambers show a rapid rise in alcohol concentration level with passing time, however the rectangle chamber responds more slowly than the hemisphere and cylindrical chambers when they inject the alcohol gas detected by the gas sensor (MQ-3) inside the sensing chambers [5].

However, the author study [5] encouraged us to specify the e-nose geometrically as a spherical shape, which we base on the temporal response and additional measuring benefits, as illustrated in Figure 5.

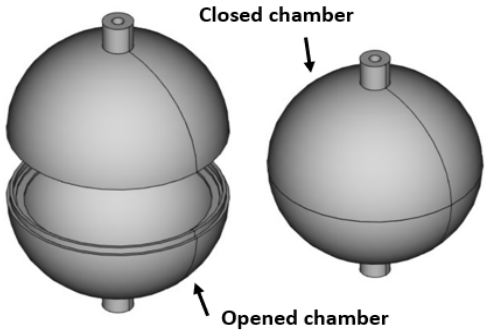


Figure 5: Spherical chamber opened and closed[4].

3.3. Sensor array arrangement

Odors of interest in practical applications are typically complicated mixes rather than pure gases. A fruit's aroma, for instance, is a sophisticated fusion of dozens of distinct odors. Finding sensors for each individual component of the gas mixture is very hard due to its complexity [14].

In the literature we found that, six to seven gas sensors are employed in the nasal sensor array [15]-[16]-[17], some researchers supporting them to eleven or more [18]-[19]. In our case, six sensors used for the measurement, which are useful to describing the odor of the sample. To be more practical, addition of temperature and humidity sensors is advised. (DHT11), which are important in the measurement and signal processing [4].

Our sensor array should be built on a PCB circle shaped that fits our spherical chamber. The hexagonal dispersion geometry makes a symmetrical form of these sensors as seen in Figure 6.

Temperature has an impact on other aspects of thermal comfort and indoor air quality [20]. The temperature in the occupied space ought to stay constant. On the other side, high humidity can promote the development and proliferation of molds and bacteria that can spread via the air. Controlling quality is a difficulty in this situation. On the other hand, relative humidity has a big role in thermal comfort. The optimum range for the relative humidity of indoor air, as per international norms and regulations, is between 30 and 60% [21].

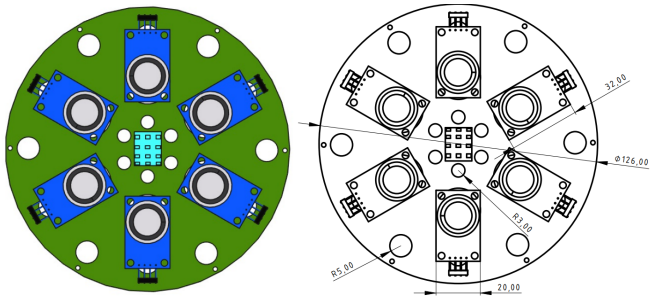


Figure 6: Sensor array network distribution in hexagonal form, as well as temperature and humidity sensors.

3.4. Flow rate

The time needed to reach a uniform condition and the chamber's capacity were two of the many variables used to evaluate the sensor's response signal performance. As shown in Figure 7, the design of a spherical chamber performed well since the stationary zone was minimal and the chamber was less intricately constructed.

The continuity equation, a fundamental principle of fluid mechanics [11], asserts that the volume flow is constant and that the quantity of any incompressible amount of fluid entering and departing are equal, The measurement item and the chamber volume determine the flow rate. Figure 7 depicts the geometry of the cell.

The calculation of velocity and flow rate, provide us to know the volume of chamber designed by applied some formulas.

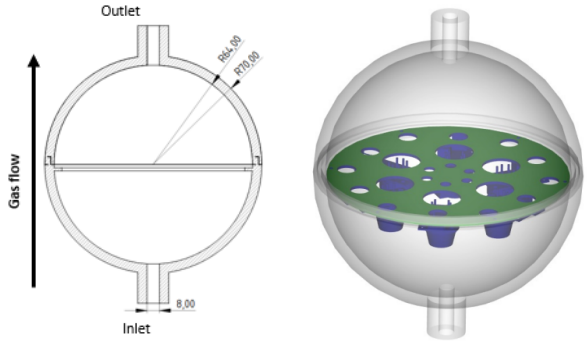


Figure 7: Dimension of the e-nose shape (mm unit).

$$V = \frac{4}{3} * \pi * r^3 \quad (1)$$

$$S = r*\pi*2 \quad (2)$$

$$D = S*v \quad (3)$$

where:

V = Volume of the chamber
S = Section of the tube
D = Debit (flow rate)
v = velocity
r = Radius

For the velocity, our object is to make a symmetrical concentration of the gas inside the sensing chamber in a good time, fixed about 180 seconds [16], and the inlet and outlet pipe section is 50.27 mm2 by (2), the volume of the sensing chamber using (1) is 1098 cm3, and velocity equal to 3.98 cm/s. With these parameters we can calculate the optimum flow rate 2 cm3/s using (3).

3.5. Architectural prototype system for sensing chamber test

The prototype's main objective is to control the quality of agricultural products by using smell (gas sensor matrix) through the first step to ensure the reliability of the information as much as possible. Performed by a procedure, elaborated in the next section, well determined at the fluidic level, to acquire the necessary data for the processing system.

In general, the experiment takes place in a well-defined time (according to the response of the gas sensor), at first the odorant is putted in the sample door, and we wait for any moment that the smell is concentrated before the flow. Then any smell is released inside the gas cell with the air by the action of opening the solenoid valve at the top V1 plus the start of the pump until the stabilization of the sensors as shown in Figure 8. The next step is to open the solenoid valve at the bottom V2 and close the solenoid valve at the top V1, so that the smell of the sample flows to the gas cell, which causes a measurement exchange at the sensor level.



Figure 8: Prototype system conception.

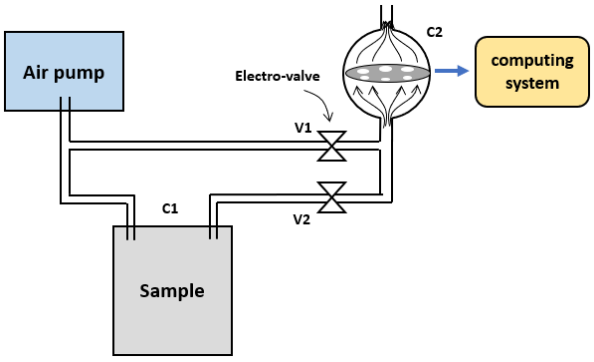


Figure 9: Synoptic diagram.

4. Results and discussions

4.1. Chamber simulation

Working on FreeCAD software using the workbench CfdOF with OpenFOAM tool to simulate the sensing chamber by setting some parameters as, ambient temperature, boundary conditions, iterations, the temperature, density and flow rate. And importing the mesh result to ParaVIEW software. According to Figure 9, we visualize the streamline movement.

In fact, energy is wasted at the level of gas movement (turbulence) [11], resulting in difficult cleaning and an imbalance in the gas concentration in the cell linked to the measuring sensors. This is just one final thought on the simulation. The cause of this effect is the connection between tube and the spherical form is 90-degree.

As a result, we have created a new modification that uses a tangential arc to eliminate the corner on both sides (Figure 11). We maintain the same settings and setup as in Figure 10, the distribution of the airflow in Figure 11 is generally uniform and devoid of turbulence. In comparison to Figure 10, both corners, which are challenging to clean, are replaced by a tangential arc, which is simple to clean. Additionally, the air pressure inside the sensor array is raised, which is good for the way the sensors and the gas under measurement interact.

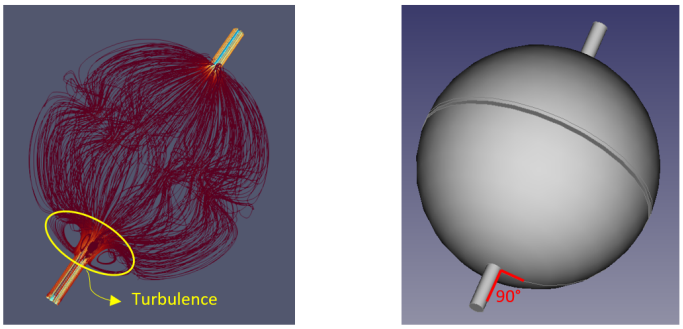


Figure 10: outcomes of fluid characteristic simulation (including the sensor array inside the chamber).

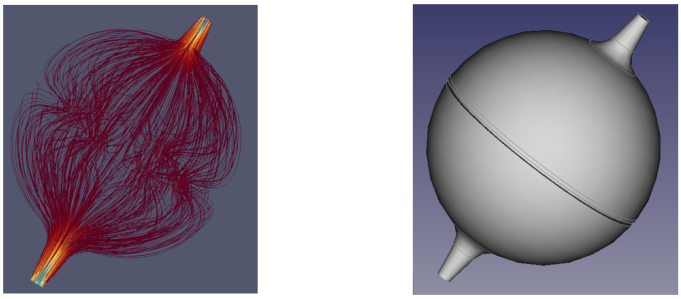


Figure 11: Simulation outcomes with the same flow rate and changed shape as in Figure 9.

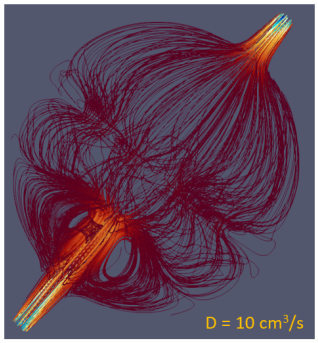


Figure 12: Simulation outcomes at 10cm^3/s flow rate.

Figure 12, which keeps the same setup as Figure 11 but changes the flow rate parameter to 10cm3/s, displays the simulation results. As depicted in Figure 10, we may draw the conclusion that the system is constrained in terms of specific parameters, such as flow rate, and that it cannot be utilized for sensing measurements since the response will be non-confidential when employed (there will be a lot of turbulence). The values used for Figure 11 therefore provide the highest geometrical performance.

4.2. Experience description

According to a huge number of tests carried out at the validation of built system dedicated to the quality control of agricultural products. This part results from tests done in order to come out with a significant percentage of system validation regarding data acquisition.

Sample: to ensure that the system works well (Figure 8) and with the lack of some sensor on the market, we worked with the most available sensors such as the MQ2 and MQ5 sensors, which are used for smoke detection and flammable gauges (for example butane), in this case our sample will be a lighter that can be contained in the cell, with a mechanical action so as to pull out the gas towards the outside.

Technical part: The sensor array consists of Six MOS chemiresistive type sensors (GS1, GS2, GS3, GS4, GS5 and GS6) capable of detecting various odors. To test the homogeneity of the chamber, the sensor array integrated with an Arduino UNO microcontroller having an on-chip analog to digital converter was used for data acquisition and plot it in the screen, two electro-valves are used for starting or stopping the gas flow coming from the pump.

A chemiresistive sensor or gas detector, principle of functionality in it [22], look like a resistor vary with different gases.

4.3. Testing the sensors responses inside the chamber

We aim to test the response of these sensors (MQ2 and MQ5), based on the above procedure of the experiment, as shown in the Figure 13.

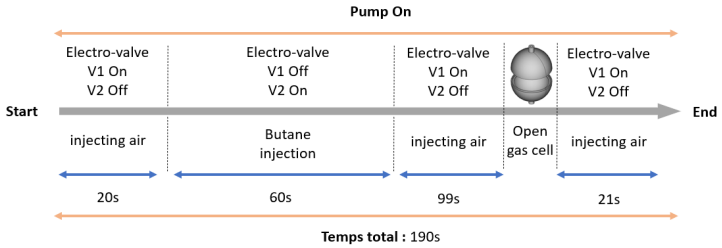


Figure 13: Procedure of the experiment.

The first 20 seconds are considered an initialization step for the sensors by airflow. Then we observe that there is a significant growth detected. This deduces that a new gas was captured. Finally, we return to the initial state performed another time by the injection of air, which will create a well on a signal decrease (the

gas cell fills with air) slowly. With this much-delayed decrease, we gave it a helping hand at time 178 s by manually opening the sensor chamber to release the butane gas in a faster way.

The sensor response is illustrated in the figure below:

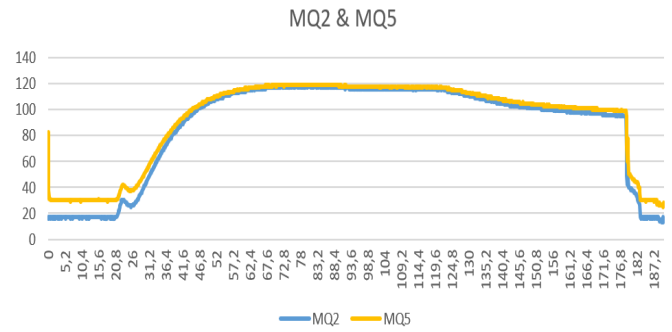


Figure 14: response of the butane gas flow from the sample cell to the measuring cell (C1 to C2 see Figure 9) observed by the sensors MQ2 and MQ5.

We can see the reaction of these sensors with butane gas, which we have an increasing after 20 seconds of butane flowing as shown in Figure 13, the amplitude gone from 20 as minimal to 120 analog values, and decreasing when we inject the air flow, with these responses we can see that the two sensors inside the chamber are functional, for reason to make sure that the system is working.

4.4. Homogeneity test in the gas chamber

To evaluate the homogeneity of the gas sensor cell, the sensors must have almost the same answer, for this, the idea is that we will follow it is to work with only one sensor of type MQ2 at a time.

Since the electronic board supports 6 gas sensor locations, this involves 6 tests for each position, because only one measurement sensor was used. The following figure shows the procedure of the experiment:

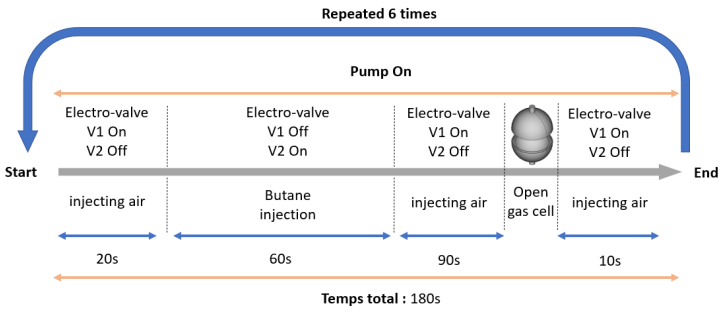


Figure 15: Homogeneity experiment procedure.

We respect as much as possible all the details of the experiment with the same condition as possible, the result is shown in the following figure.

By respecting the same conditions at each turn of the experiment, namely the concentration of gas in the sample cell in C1, the opening and closing of the solenoid valves and the practical operation of the MQ2 sensor.

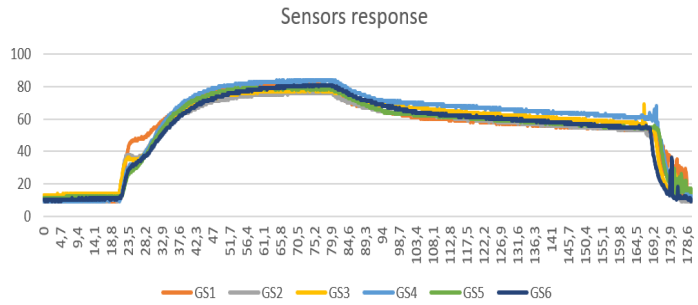


Figure 16: Homogeneity test response on the 6 sensor positions in C2 cell.

From these measurements it is found that they are almost homogeneous with some small deference either at the beginning of the transitional regime and the end of the measurement, the first is due to the fact that the concentration of gas mixture (air + butane) are not the same when it touches the sensor membrane (the sensitive part of the sensor against the gauze) for the 6 different positions on the electronic board, for the second is generated by the opening of the chamber causes a vibration of the sensor giving an observable disturbance at the measurement level.

But in general, these answers present a quality of measurement in the permanent diet that is most important in the data processing phase.

5. Conclusion and future work

For several uses in machine learning, including quality control, the e-nose system was an essential tool. The sensor chamber is crucial in the detection phase of any electronic nasal application. This investigation's analytical, which mainly focused on geometrical restrictions of e-nose design, discovered that, the spherical shape shows the highest geometrical performance for the debit due to several factors including volume and shape.

On the one hand, this practical work is an image that must respond to the simulation of the gas cell on the homogeneity of data capture according to the sensors. In the same way, we found a compatibility relationship between them, since the 6 sensors positions have an important reliability of the information, which gives the next step towards the processing and the pattern recognition phases.

References

[1] H. J. Ko and T. H. Park, "Bioelectronic nose and its application to smell visualization," J. Biol. Eng., **10**(1), 17, Dec. 2016, doi: 10.1186/s13036-016-0041-4.

[2] G. Ohloff., "Chemistry of odor stimuli.Experientia," **42**, 271-279., 1986.

[3] S. S. Schiffman and T. C. Pearce, "Introduction to Olfaction: Perception, Anatomy, Physiology, and Molecular Biology," in Handbook of Machine Olfaction, T. C. Pearce, S. S. Schiffman, H. T. Nagle, and J. W. Gardner, Eds. Weinheim, FRG: Wiley-VCH Verlag GmbH & Co. KGaA, 2002, 1–31. doi: 10.1002/3527601597.ch1.

[4] M. Harmouzi, A. Amari, and L. Masmoudi, "The conception of gas chamber for electronic nose," in 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET),

Meknes, Morocco, Mar. 2022, 1–4. doi: 10.1109/IRASET52964.2022.9737802.
[5] N. S. Samiyan and M. M. Addi, "Characterization of Sensing Chamber Design for E-Nose Applications," J. Telecommun. Electron. Comput. Eng. JTEC, **9**, 3–9, Art. no. 3–9, Dec. 2017.
[6] P. Boeker, "On 'Electronic Nose' methodology," Sens. Actuators B Chem., vol. 204, 2–17, Dec. 2014, doi: 10.1016/j.snb.2014.07.087.
[7] S. Kiani, S. Minaei, and M. Ghasemi-Varnamkhasti, "Instrumental approaches and innovative systems for saffron quality assessment," J. Food Eng., **216**, 1–10, Jan. 2018, doi: 10.1016/j.jfoodeng.2017.06.022.
[8] E. L. Hines, P. Boilot, J. W. Gardner, and M. A. Gongora, "Pattern Analysis for Electronic Noses," in Handbook of Machine Olfaction, T. C. Pearce, S. S. Schiffman, H. T. Nagle, and J. W. Gardner, Eds. Weinheim, FRG: Wiley-VCH Verlag GmbH & Co. KGaA, 2002, 133–160. doi: 10.1002/3527601597.ch6.
[9] R. Gutierrez-Osuna, H. T. Nagle, B. Kermani, and S. S. Schiffman, "Signal Conditioning and Preprocessing," in Handbook of Machine Olfaction, T. C. Pearce, S. S. Schiffman, H. T. Nagle, and J. W. Gardner, Eds. Weinheim, FRG: Wiley-VCH Verlag GmbH & Co. KGaA, 2002, 105–132. doi: 10.1002/3527601597.ch5.
[10] A. Koocheki and E. Milani, "Saffron adulteration," in Saffron, Elsevier, 2020, 321–334. doi: 10.1016/B978-0-12-818638-1.00020-4.
[11] J. Roussel, "Cours de mécanique des fluides – femto-physique.fr," 68.
[12] "Understanding Pressure Measurement." https://www.meddeviceonline.com/doc/understanding-pressure-measurement-0001 (accessed Oct. 19, 2021).
[13] S.-M. Jafari, M. Z. Tsimidou, H. Rajabi, and A. Kyriakoudi, "Bioactive ingredients of saffron: extraction, analysis, applications," in Saffron, Elsevier, 2020, 261–290. doi: 10.1016/B978-0-12-818638-1.00016-2.
[14] K.-T. Tang, S.-W. Chiu, C.-H. Pan, H.-Y. Hsieh, Y.-S. Liang, and S.-C. Liu, "Development of a Portable Electronic Nose System for the Detection and Classification of Fruity Odors," Sensors, **10**(10), 9179–9193, Oct. 2010, doi: 10.3390/s101009179.
[15] S. Kiani, S. Minaei, and M. Ghasemi-Varnamkhasti, "Integration of computer vision and electronic nose as non-destructive systems for saffron adulteration detection," Comput. Electron. Agric., **141**, 46–53, Sep. 2017, doi: 10.1016/j.compag.2017.06.018.
[16] M. Ezhilan, N. Nesakumar, K. J. Babu, C. S. Srinandan, and J. B. B. Rayappan, "Freshness Assessment of Broccoli using Electronic Nose," Measurement, **145**, 735–743, Oct. 2019, doi: 10.1016/j.measurement.2019.06.005.
[17] K. Heidarbeigi, S. S. Mohtasebi, A. Foroughirad, M. Ghasemi-Varnamkhasti, S. Rafiee, and K. Rezaei, "Detection of Adulteration in Saffron Samples Using Electronic Nose," Int. J. Food Prop., **18**(7), 1391–1401, Jul. 2015, doi: 10.1080/10942912.2014.915850.
[18] P. Le Maout., "Polyaniline nanocomposites based sensor array for breath ammonia analysis. Portable e-nose approach to non-invasive diagnosis of chronic kidney disease," Sens. Actuators B Chem., **274**, 616–626, Nov. 2018, doi: 10.1016/j.snb.2018.07.178.
[19] L. Kumari, "Various techniques useful for determination of adulterants in valuable saffron: A review", 3 March 2021, doi: https://doi.org/10.1016/j.tifs.2021.02.061
[20] T. Hübert, "Chapter 12 - Electronic Noses for the Quality Control of Spices," 115–124, 2016, doi: https://doi.org/10.1016/B978-0-12-800243-8.00012-3.
[21] M. S. Hallé, M. L. Dufresne, and M. L. Lamarche, " MODÉLISATION DE LA QUALITÉ DE L'AIR DANS UNE UNITÉ DE BRONCHOSCOPIE : INFLUENCE DES STRATÉGIES DE VENTILATION," 132.
[22] M. A. Franco, P. P. Conti, R. S. Andre, and D. S. Correa, "A review on chemiresistive ZnO gas sensors," Sens. Actuators Rep., **4**, 100100, Nov. 2022, doi: 10.1016/j.snr.2022.100100.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 4, 105-113 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

A Comparison of Cyber Security Reports for 2020 of Central European Countries

Kamil Halouzka*, Ladislav Burita, Aneta Coufalikova, Pavel Kozak, Petr Františ

Department of Informatics and Cyber Operations, University of Defence, Brno, 66210, Czech Republic

ARTICLE INFO

Article history:
Received: 11 May, 2022
Accepted: 23 July, 2022
Online: 16 August, 2022

Keywords:
Cyber security
Cyber threat
Cyber-attack
Annual reports
Strategic documents
Phishing email defense

ABSTRACT

The aim of the article is to analyze the annual reports on cyber security of Central European countries, i.e. the Czech Republic, Slovakia, Poland, Germany, and Austria. The article focuses on the development of the state of cyber security, actors of threats in cyberspace, cyber threats, and the most common types of attacks. The article evaluates the objectives of cyber-attacks from the point of view of state institutions, organizations, and state and private companies, and they have listed the follow-up measures here. The method used is a critical verbal evaluation with comments and comparative analysis to find the strengths and weaknesses of the evaluated cyber security strategies and learn from them. The experiment of the cyber defense against phishing attacks is mentioned as an example of the cyber defense of individuals. The rules in Microsoft Outlook were used by filtering incoming email messages. The result is promising by stopping 88 % of phishing emails. The discussion and conclusion state that COVID-19 played a big role in the cyber security situation in countries to the analyzed documents.

1. Introduction

This review paper is an extension of the work originally presented in the 2021 Communication and Information Technologies Conference Proceedings [1].

The goal of the article is to analyze the situation in the field of cyber security, using the latest strategic documents of Central European countries, reports of Computer Emergency Response Teams (CERT), and The National Security Agencies (NSA).

The result of the analysis is somewhat pessimistic because in cyberspace we are the object of an ever-increasing number of cyber-attacks and technologically we lag behind the attackers. On the other hand, the paper mentioned a positive example of an effective defense against phishing email attacks.

Cyberspace is not limited by geographical boundaries, and its actual state can be characterized by continuing enlargement of cyber threats and cyber-attacks, whereas any of them are very serious and the form of cyber warfare is approaching.

Private companies and public organizations have to face permanently against cyber-attacks and deal with their consequences. International cooperation in cyber security is very important, countries participate by exchanging information about cyber threats and attacks and organizing joint exercises.

Cyberspace was recognized by NATO as a new domain of warfare in 2016; comparable to other domains: land, air, sea, and outer space. Appropriate policies, action plans, committees, and agencies have been adopted to ensure Member States' cyber security. Cyber headquarters and operational centers with relevant troops have been established and intensive preparations are underway against cyber threats and cyber-attacks.

The article is interesting in existing threats in cyberspace with regard to their danger and with a focus on cyber-attacks, associated with real military activities (aggression), especially from Russia.

The core of the article is the analysis of new strategic documents in cyber security in selected countries in order to look for their identical and different areas and obtain the necessary lessons learned for the development of better endurance in cyber security and preparation of suitable sources for education.

The analyzed documents often deal with phishing attacks and defense against them, because these attacks are usually at the beginning of larger cyber-attacks.

Phishing is a form of attack with the help of social engineering techniques, in which an attacker pretends to be a trusted authority in order to obtain sensitive data from the victim. The attacker thus often tries to gain the trust of the victim, who then actually communicates the necessary information or data voluntarily.

*Corresponding Author: Kamil Halouzka, kamil.halouzka@unob.cz

An example of a possible effective defense of individuals against phishing attacks is mentioned in an experiment about the application of Microsoft Outlook email client functions.

2. The Literature Review

The section presents selected comparative studies of national and international documents in cyber security and analyzes the ability of cyber defense. The analyzed studies are prepared in the form of a verbal and tabular description of the comparison, but suitable analytical procedures and models with the form of graphical outputs are also used. The aim is to identify the strengths and weaknesses of the analyzed documents and learn to improve their own approaches.

An extensive study [2] prepared for the Italian Parliament compares NATO member countries' approaches to cyber defense and finds that some countries have more proactive approaches (US, UK, France), while others have a more defensive approach (Germany, Spain). Although there are differences in the approaches to cyber defense, it must be seen that all nations are affected by NATO's unified regulatory and doctrinal framework, so that despite existing differences, national elements of cyber security can be integrated with the Alliance's command structure. The 2019 NATO Summit declared that due to the geopolitical activities of China and Russia, preparation for cyber defense had become a top priority. From the technical point of view, the NATO Communications and Information Agency (NCIA) provide capabilities necessary to the Alliance's structures in terms of cyber defense. The NCIA administrates some of the allied networks with the NATO Cyber Security Centre (NCSC) and the NATO Computer Incident Response Capability (NCIRC). Finally, outside the NATO military command structure, the Cooperative Cyber Defence Centre of Excellence (CCDCOE) in Estonia, created in 2008, prepares reports and other documents in the field of cyber defense and, since 2010, hosts regularly cyber security exercises.

A paper [2] further analyzes the procedures and results in the cyber defense of individual NATO nations. The US approach to cyber defense is qualitatively and quantitatively different from that of most European countries. The National Security Strategy from 2017 underlines the cyber domain as one of the main future battlegrounds, and the 2018 Strategy warns against adversarial capabilities damaging American armed forces, economy, and society in cyberspace. The US Department of Defense established a Cyber Command (USCYBERCOM) in 2009, within the Strategic Command, whose commander is at the same time the Director of the National Security Agency (NSA), to ensure seamless cooperation between cyber and intelligence operations. The USCYBERCOM strategy is focused to:

1. Achieve and sustain capabilities, by anticipating technological changes and exploiting them faster and more effectively than the adversaries.
2. Create cyberspace capabilities to support operations in other warfare domains.
3. Ensure information superiority to achieve strategic impact.
4. Operationalize the cyber battlespace for agile maneuvers.
5. Expand and deepen partnerships with agencies, the private sector, and academia.

The conclusion stated that cyber defense at the NATO level is not limited to the creation of command structures and the employment of dedicated personnel, but also involves broader partnerships. The necessity of equipping NATO with cutting-edge technology led in 2014 to the formation of specific cooperation with industries operating in the cyber sector. NATO-EU cooperation was already listing the cyber dimension among priority areas of collaboration in 2016.

Analysis of the Polish National Cyber Security Strategy (NCSS) in comparison to EU strategy and regulations and to other NCSS of the countries (US, UK, France, Lithuania, and Estonia) is the content of the paper [3]. The definition of cyberspace is not enough clear and often is in the documents missing. An example is a definition by the US Department of Defense "A global domain and the information environment including networks, information technology infrastructure, data sources, telecommunications, the Internet, computers, and embedded systems", cited in [4]. The EU security strategy in cyberspace, issued in 2013, clarified goals, responsibility roles, and tasks as achieving cyber resilience, preparing an EU Cyber Defence Policy and sources, reducing cybercrime, and developing needed technologies. In 2017, the EU published a cyber security document including initiatives in resilience to cyber-attacks and cyber security capacity, effective criminal law, and complex stability in international cooperation. The 2019 Cyber Security Act has provided a consolidated cyber security certification framework. The sanctions system Cyber Diplomacy Toolbox, allows the EU to impose targeted restrictive measures to prevent and respond to cyber-attacks. The European Parliament adopted 2021 a decision on the EU's Cyber Security Strategy for the next digital Decade to make developed tools and services secure from the start of development, resilient to cyber threats, and able to quickly react if vulnerabilities are discovered. Poland has agreed in 2017 to its NCSS for 2017-2022 which defines cyber security as "The resilience of information and communication systems, at a given level of confidence, to any activity that compromises the availability, integrity, authenticity, or confidentiality of data, or the related services oriented by or accessible via these information networks and systems". In 2018, Poland adopted the National Cyber Security System Act establishing coordination measures of state policy in the area of cyber security, and in 2019, in accordance to the European Union Agency for Network and Information Security (ENISA) lifecycle approach, Poland adopted its improved NCSS for 2019-2024. The final evaluation of the Polish NCSS contains recommendations for its further development, based on comparisons with the approaches of other countries.

Complex comparative analyses [5] of national cyber security strategies (NCSS) combine areas of industry, economy, technology, and defense. The study characterizes the NCSSs of countries US, UK, Japan, and EU, and describes cyber security agendas for the revision of NCSS in South Korea, by applying topic modeling. Topic modeling involves statistical techniques to identify hidden structures from a set of documents. The result is 15 agendas in the areas of Infra Stability, Protection and Response Capability, Industry and Technology, and International Cooperation. The NCSS of the US emphasized improving incident response capabilities, especially cybercrime law enforcement and investigation capabilities, and establishing cyber security governance. Similar to the US, the UK prioritized protection and

response capability. On the other hand, the NCSS of Japan establishes a relatively high proportion to the cyber security industry and technology sector. Finally, the NCSS of the EU picks up international cooperation.

The article [6] is interesting in terms of currently ongoing Russian aggression in Ukraine, it includes a comparative analysis of cyber security systems in Russia and Armenia. In the introduction, the close cooperation between both countries in the economic, political, and military fields is appreciated. Two levels (legal and practical) are used to analyze cyber strategies, policies, and institutions. Key theoretical concepts in information security, information warfare, etc. are described. The cyber security definition is mentioned as "A set of technical and non-technical (policies, security arrangements, actions, guidelines, risk management) measures allowing to provide social, ethnic and cultural evolutionary modernization of the critical cyber infrastructure, as well as protection of vital interests of human, society, and state." The experiences from the military operation of Russia in August 2008 in South Ossetia and Georgia changed the Russian Defense Ministry's intention to create informational troops, whose functions include all aspects of information warfare, from psychological operations and propaganda to security of computer networks and cyber-attacks on the enemy's information systems. Cyberspace in Armenia is rather liberal. The principle is "allowing everything that is not prohibited" when prohibited are direct and clear criminal acts. Similar to the Russian approach, the Armenian side uses a wider concept of information security without specifying the concept of cyber security.

3. Analysis of selected cyber threats

The aim of the section is not to list every possible cyber threat. A brief introduction into cyber threats was presented in [1]. The initial point of an intrusion into a system by attackers is a very often successful phishing campaign. Therefore, the aim of the section is to analyze phishing from several points of view, especially phishing associated with real military actions.

Phishing campaigns usually have the following aims:

1. They try to lure sensitive data from end-users and misuse obtained data. It can be personal data, corporate data, or governmental data.
2. The goal is to infect a computer for later abuse by adding malicious attachments to e-mails or creating links leading to fake malicious websites, thus computers may become part of a botnet, be infected with ransomware, contain a key logger, etc.

In addition to classical phishing attacks that target as many victims as possible, there are also other types of phishing [7]:

- Spear phishing attempts are targeted toward specific individuals or groups of individuals. They may include the recipient's name, position, or company. IT administrators can be great targets because of the level of access they usually have within the organization.
- Whaling targets high-level employees, like executives or directors. They typically have access to the most valuable information in a company, making them appealing targets for attackers.
- Clone phishing is typically targeted at a small group of people. Attackers copy a legitimate email that has previously been sent by a trusted organization but replace links to redirect the victim to a malicious site.

Phishing attacks usually take place over email but attacks using other mediums have also been observed. Smishing is the text message (SMS) version of phishing attacks. Vishing is phishing that is executed via telephone (Voice).

Phishers usually exploit three types of events. Firstly, phishing campaigns misuse long-term affairs such as humanitarian aid to countries affected by perpetual fighting, or an education for African children. Secondly, regularly occurring events are commonly exploited by phishers, e.g., holidays, summits (EU, NATO), or elections. Thirdly, phishers take advantage of current events, e.g., outbreaks of fighting, rapid changes in the financial market. A common characteristic of phishing is the feeling which is intended to evoke in people. This effect encourages users to react, and as a consequence, it increases the effectiveness of phishing. The feelings they usually evoke can be: sympathy, fear, joy from prize, urgency, stress, and patriotism.

Patriotic or nationalist hackers see themselves as irregular soldiers, or conscripts fighting a war for their country, a form of cyber militia. Their attacks are motivated by strong feelings of patriotism and nationalism, reflected in the language and rhetoric used. The actions of the patriotic hacker may result in serious damage to targeted systems [8].

In the case of intensified Russian military activities there are 2 possible vectors of cyber-attacks:

1. From Russia - attacking information systems (IS) of the enemy and enemy sympathizers.
2. Against Russia - attacking IS of Russia and Russian sympathizers.
3. From all around the world - attacking IS of adversaries and their sympathizers, depending on which side the attackers are inclined towards.

Phishing campaigns related to current events in the context of Russian activities may use the terms humanitarian aid, solidarity, support for the fight, signing petitions, and providing accommodation to refugees. The impact of such phishing campaigns tends to be personal or sensitive data leakage, payment card details leakage, subsequent misuse of the leaked data, and payments to the attackers' accounts instead of accounts for humanitarian purposes, computers infected with malicious codes, and their subsequent abuse for adversary activities.

Phishing can also contain fake actual news. It aims to manipulate people across the board and significantly influence their behavior. This method is often used in state information operations to weaken an adversary or, on the contrary, to strengthen the confidence of its own population in the government aggressive activities. Such phishing messages include fake news of invasion, troop movements, shortages of goods in shops, fuel shortages, shutdown of gas supplies by Russia, power cuts due to gas shortages, etc. On the other hand, fakes news about military successes, humane treatment of the enemy, liberation of the population from oppression and other justifications for fighting are

then used to influence the confidence of the population in the governmental aggressive decisions.

In a widespread phishing campaign against an adversary, the consequences can be immeasurable - population panic, buying frenzies, and stockpiling of resources. With an unexpectedly height public reactions, we can expect repercussions in all sectors. Particularly for the communication and information area, we can expect disruptions or even unavailability of services or entire information systems, e.g., disruption of electronic banking, unavailability of telephone lines, unavailability of web information portals, and unavailability of bank card payments. Additionally, if the critical infrastructure is affected, the functioning of the state and human lives may be endangered.

Another frequent impact of successful phishing can be a ransomware infection or a data leakage. Both encrypted data as well as data leakage are common subject of a ransom. It is not recommended to pay the ransom due to uncertainty as to whether the attackers keep their promises. In case of encrypted data there is a risk the attackers will not send decrypting keys after payment or the keys will not work properly. In case of the data leakage there is a risk the attackers will sell stolen data to a third party after payment anyway. A new approach to ransomware has emerged. Anybody can buy ransomware as a service for a fee or a shared ransom. Effects of ransomware on information systems range from denial of service, and data loss, to loss of reputation, and bankruptcy. The consequences of an infection in the critical infrastructure are immense, as it was mentioned above.

A new and additional set of phishing related to Russian military activities and the current situation in Russia and Ukraine is emerging. This new set has specific keywords in conjunction with the words Russia or Ukraine, e.g., solidarity, refugees, war victims, aid, fundraising, petition, cohesion, but also shortages in connection with the words gas, wheat, energy, building materials, etc. Such keywords can be effectively used in individual defense against phishing attacks which is described in section 5.

There are two main approaches to reduce a rate of success phishing. From a technical point of view we can increase detection capabilities. From a non-technical point of view we can reduce a phishing risk by spreading security awareness.

Firstly, technical measures work reliably but they need to be updated and adapted regularly. It is very difficult to respond adequately to rapidly changing links in phishing emails. Links in emails are changed by attackers faster than it would be possible to manually respond to them in real time. Manually adding malicious links into blacklists is an inefficient human-consuming and time-consuming method. A more appropriate method is to automatically and regularly download updated blacklists from selected trusted sources that deal with this issue.

Blocking selected file types (.dll, .exe, .js, .msi, .reg...) is a very effective method. However, if we block files types inappropriately, users will not receive their attachments needed for their work. For example, .pdf files may contain links to malicious sites, but it is not possible to block all .pdf attachments in bulk. The solution is a technology that runs selected suspicious attachments in a sandbox. The technology automatically evaluates the behavior of the system after the attachment is launched and if everything works normally, the email is sent to the user's mailbox. If malicious behavior is found, the user is informed about the situation and the email is sent with a modified attachment, for example an attachment is converted from a .pdf file to an image (.png, .jpeg...). This prevents the user from clicking on the malicious link or executing the malicious code, but the information from the attachment is still delivered to the user.

Another technical measure is blocking emails based on sets of keywords. The sets must be chosen with caution, taking into consideration the possibility of a high number of false positives. In long-term tuning, this method achieves a very good level of reliability. The practical application of this technical measure is discussed in section 5.

Secondly, there are many methods of security awareness spreading. For example, users can attend specialized seminars and courses to learn how to recognize phishing. However, these trainings are usually only once a year due to financial and time constraints. In addition, this approach is not proving to be as effective as expected. In practice, short training sessions followed by testing employees with mock phishing is more effective method. Metrics such as the number of tricked users, the most clicked links, types of attachments launched, and, of course, leaked passwords are analyzed and final reports are published within the organization. Modified mock phishing emails are then repeated at random time intervals and repeat clickers are then invited to additional trainings. According to [9], it is advisable to increase the difficulty of mock phishing up to 3 levels:

- Tier 01 – generic type of mass phishing attacks. Emails include misspellings, poor graphic design, and well-known scams.
- Tier 02 – more professional or more personalized phishing attacks. Emails contain victim's name and are work related.
- Tier 03 – targeted attack and customized for selected high-risk groups.

The best results for reducing the success rate of phishing attacks are achieved through a combination of technical measures and educated users. It depends on the availability of resources, how many security technical tools are used and to what extent users are trained and tested.

4. Cyber Security Status Report

The aim of the next section is to analyze the cyber security status reports of the Czech Republic, Slovakia, Austria, Germany, and Poland for the year 2020. Cyber security status reports are issued by these countries in the second half of the year. For this reason, the reports for 2020 were prepared. The reports for 2021 were not available at the time of writing. The aim is to assess the problems that each country has in the area of cyber security, to evaluate the frequency of cyber-attacks or incidents, and to evaluate the most common types of attacks.

4.1. Czech Republic

A report on the state of cyber security in the Czech Republic has been published by the National Cyber and Information Security Agency (NÚKIB) [10]. The year 2020 in the Czech Republic was characterized by an increase in the number of cyber-

attacks against Czech institutions, organizations, and companies in all sectors. In 2020, the NÚKIB recorded a more than double increase in incidents compared to 2019. The most common types of attacks in the Czech Republic in 2020 were spam (59%), phishing (16%), and scanning (12%). Respondents ranked ransomware (19%), DoS / DDoS attacks (19%), spear-phishing emails (14%) and attempts to exploit vulnerabilities (13%) as the most serious attacks. Cybercrime has long been among the most serious threats to the country's cyber security. In 2020, cybercrime emerged in the form of ransomware attacks, which hit the Czech healthcare sector to a large extent.

The most serious threats to the Czech Republic's cyber security include state-sponsored actors in cyberspace and cybercrime. A new development is Ransomware as a Service (RaaS), which is a service provided by ransomware developers to other hackers, usually for a share of the ransom, and they do not care about the actual penetration of organizations' systems.

In terms of personnel and financial security, a large number of organizations in the country have been facing a lack of experts and insufficient budgets in the field of cyber security. This situation was more evident in the government sector than in private companies. Almost none of the interviewed organizations had all cyber security positions filled. More than half of the organizations cited inadequate salary conditions as the main factor.

In terms of training, the NÚKIB placed a strong emphasis on the training of state administration employees and trained more than 22,000 state administration employees, employees of the Army of the Czech Republic, and medical and prevention personnel from the education sector in e-learning courses during 2020.

In the report on the state of cyber security in the Czech Republic, much attention is paid to the security of 5G networks. In 2020, the Prague 5G Security Conference was organized jointly with the Office of the Government and the Ministry of Foreign Affairs, the main topic was the risks associated with building 5G infrastructure. The main outcome was the presentation and the launch of the Prague 5G Security Repository, a virtual library designed to share legislative, strategic and other tools that states adopted in the past year in the area of 5G network security.

4.2. Slovak Republic

A report on the state of cyber security in Slovakia has been published by the National Security Authority of Slovakia [11]. The report is divided into seven main parts. The focus is on the description of actors in cyberspace, namely inexperienced attackers (script-kiddies), cybercriminals motivated by financial gain, state-sponsored groups, hacktivist groups interested in obtaining sensitive and classified state information, and cyberterrorists, whose role is mainly to hit civilian and military targets with cyber-attacks.

Furthermore, the report focuses on the categories of cyber security incidents (monitored by SK-CERT) and threats. The most detected and reported incidents were in the "Unwanted Content" category and the most solved incidents were in the "Attempted Intrusion" category. The most frequent threats in Slovakia were phishing campaigns (Microsoft tech support scam, abuse of Slovak Post), malicious code distribution (mainly ransomware – the attack on Slovak TV Senzi, which refused to negotiate with the attackers and filed a criminal complaint), data leaks (leak of 130,000 personal data of patients tested on COVID-19) and vulnerability exploitation.

The Slovak cyber security status report also goes into detail on cyber threats targeting specific organizations, such as healthcare, public administration, banking, electronic communications, and digital infrastructure, and energy. The report pays great attention to, among other things, issues related to national and European legislation, the preparation of the National Cyber Security Strategy, national and international activities and cooperation, cyber security audits, and cyber defense exercises (Table-Top exercise BlueOLEx 2020 and Cyber Coalition 2020). Among other important security actions that were implemented in Slovakia was the establishment of the Competence and Certification Centre for Cyber Security. The aim of this center is to assist the National Security Authority in fulfilling its professional tasks in the field of cyber security, protection of classified information and cryptographic protection, and trust services in the public interest.

4.3. Austria

The Cyber Security Status Report was prepared by the Cyber Security Steering Group (CSS) in accordance with the Austrian Cyber Security Strategy (ÖSCS) [12]. As in the above-mentioned countries, the number of malware attacks on computer systems and networks in Austria increased over the past year. A large part of Austria's annual report was devoted to the impact of the SARS-CoV-2 pandemic on cyber security. At the beginning of the pandemic, many companies were forced to change to a home office for which they were unfortunately not prepared. Companies were often forced to reduce their own cyber security to allow their employees to work away from their offices.

There has been a sharp increase in the number of fraudulent sites, seemingly related to Covid-19, designed to phish or spread malware. The authors of these scams demanded from victims to pay them USD 4 000 in bitcoins. If they didn't pay, they were told their families would be infected with the coronavirus. Other frauds involved changing delivery times for shipments due to the pandemic. Opening the link and/or file contained in the message caused the installation of malware (including AZORuIt, Emotet, Nanocore RAT and Trick-Bot) on the target computer. There were major problems in Austria with data leaks from corporate computer networks, where attackers demanded a ransom for its return. Several waves of DDoS attacks occurred during the reporting period, mainly against banks, the financial sector and Internet Service Providers (ISPs). The aim of these attacks was not only to deny services but also to blackmail their victims.

A large part of the report is devoted to cyber security cooperation between Austria and European Union, United Nations, NATO, and other important committees and forums. Equal attention is paid to clarifying national cyber security actors such as Cyber Security Centre, Cybercrime Competence Centre, CIS and Cyber Security Centre, Austrian Armed Forces Security Agency, and many others.

The Austrian Cyber Security Status Report is the only one that does not provide any specific numbers of cyber security breaches, as all values were given only as percentages.

4.4. Poland

A report on the state of cyber security in Poland has been published by CERT (Computer Emergency Response Team) Poland [13]. In the period under review, 60.7% more cyber security attacks were registered than in the previous year. The most common type of incident was phishing, which represented 73% of all cyber-attacks. In March 2020, Poland released a list of dangerous websites (List of warnings), which had a significant impact on the number of phishing attacks recorded. These phishing attacks were targeted at obtaining e-banking authentication details, payment card details, email account access details, and social media accounts.

Cybercriminals used, for example, Facebook messages with sensationalist headlines, fake SMS messages, and WhatsApp messages for this purpose. There also was an increase in unwanted messages (spam) on mobile platforms (especially Android). CERT Polska focused on analyzing IP addresses localized in Poland that were used for Distributed Reflective Denial of Service (DRDoS) attacks. For that purpose, a list of poorly configured services that were the most frequently used for DRDoS attacks was published.

As in previous years, disinformation campaigns related to attacks on information portals and accounts of Polish politicians were recorded in Poland. Criminals used the accounts to publish fake information aimed at, for example, reducing the trust of public officials or spreading negative information about the US military in Poland.

The report on the state of cyber security in Poland also contains a large number of practical examples (in text and graphic form) of the most common form of malicious software distribution, fake job offers on Facebook, fake parcel post services, fake bills for the advertisement and others. The extensive chapter is completed with recommendations on how to avoid infection.

4.5. Germany

In Germany, the Federal Office for Information Security (BSI) monitors IT security threats [14]. According to the report on the state of cyber security in Germany, the trend of attackers using malware to launch mass cyber-attacks continued during the period. The malware was the most commonly used attack and was responsible for cyber-attacks on individuals, private companies, government offices, and other institutions. The malware was also used to launch targeted attacks against pre-selected victims.

A large amount of personal data was also leaked, which unfortunately also included data on patients and their clinical records (this data was unfortunately freely accessible online). The reporting period also showed the emergence of a number of vulnerabilities in software products that attackers were able to exploit to spread malware, attack or steal data. Some of these vulnerabilities were assessed as critical.

There were targeted attacks on financially powerful victims such as car manufacturers and their suppliers, attacks on airports and airlines, and on less known high-income companies. Attackers also used increasingly the "human factor" as a starting point for attacks that use social engineering to gain an entry point for further attacks. Also in Germany, the COVID-19 pandemic was often used for cyber-attacks.

One example was the large-scale waves of spam offering fake advice about the coronavirus. These emails urged company employees to post personal or company information on copies of official websites. Cybercriminals designed these sites similar to (government) websites.

One of the biggest threats mentioned in the annual report was Emotet. Emotet was used to create a cascade of other malware attacks that can culminate in targeted ransomware attacks on selected, usually wealthy victims. Critical BlueKeep and DejaBlue vulnerabilities in Windows Remote Desktop Protocol were also published. This vulnerability allowed attackers to execute malicious code, including malware, on unpatched systems.

4.6. Brief comparison of cyber security reports

As noted, each of these states issues an annual report on the state of cyber security. Each state publishes information in this report at its own discretion, and there is generally no rule specifying what the report must contain. The following table (Tab. 1) shows a basic comparison of the cyber security status reports of these states, with information from the United States and the United Kingdom reports added for comparison.

While for the UK the Annual Review 2020 report [15] issued by The National Cyber Security Centre was used for comparison, in the US two reports were used, namely the NSA cybersecurity year in review for 2020 [16] and the FISMA FY 2020 Annual Report containing The State of Federal Cybersecurity [17]. In the case of the United States, both reports are very general and contain little specific information compared to other reports.

Tab. 1 shows who is responsible for the policy/strategy in the countries listed, when the first National Cyber Security Strategy (NCSS) was issued, and when the last update was made or the period of validity. The table also shows important aspects of each cyber security status report, when the focus was on new IT security measures in organizations and the status of security-focused budgets, which were heavily affected by the COVID-19 pandemic and the resulting increase in security and budgetary measures. Cyber security Status Report lists the most frequent incidents during the reporting period, where the most popular type of incident was phishing. The second most reported incident type was ransomware. There was also mentioned a large increase in malware on mobile platforms in 2020. An important part of cyber preparedness is also participation in international cyber exercises, which have also been affected by the COVID-19 pandemic. Even though the Locked Shields 2020 and Cyber Europe 2020 international exercises were canceled in 2020 due to the pandemic, individual countries have organized several cyber security exercises that are important for training defense against cyber-attacks. The table also shows that national security authorities support cyber security education in the form of cooperation with schools (e.g. cyber security information cards for schools), business (e.g. various levels of consulting for business customers), and voluntary sector (e.g. guidance on cyber security).

Table 1: Basic comparison of the cyber security status reports

	USA	GBR	Germany	Polland	Austria	Slovakia	Czech Republic
Policy/strategy responsibility	CISA	NCSC	BMI	MDA	BMEIA	NBÚ	NÚKIB
First NCSS	2003	2011	2011	2009	2009	2008	2012
Actual NCSS	2018	2022 - 2030	2021	2019 - 2024	2021	2021 - 2025	2021 - 2025
New IT security measures in companies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Security budget in companies	Not mentioned	Increase	Not mentioned	Increase	Increase	Increase	Increase
The most frequent incident reported during the reporting period	Spear-phishing	Phishing	It's not specified. The report generally mentions malware and phishing.	Phishing	Phishing	Phishing	Spam
The second most frequent incident reported during the reporting period	Not mentioned	Ransomware		Ransomware	Not mentioned	Ransomware	Phishing
Participation in cyber exercises	Not mentioned	Not mentioned	Yes / Crossed Swords 2020, Common Roof	Yes / KSC-EXE 2020, Capture The Flag, Hack-A-Sat	Yes / Crossed Swords 2020, Common Roof	Yes / BlueOLEx, Cyber Coalition	Yes / Cyber Coalition
Cyber security education in companies	Yes / improving	Yes / improving	Yes / improving	Yes / improving	Yes / improving	Yes / improving	Yes / improving
Cyber security incidents in 2020	30819	723	419 (mentioned only critical infrastructure)	10420	Not mentioned	3793	1267

The latest comparison is on the number of cyber incidents in 2020. It is difficult to compare the number of reported cyber incidents from the cyber security status reports of these states because each state reported these cyber incidents in the form of the most frequent, most serious, reported or resolved cyber incidents. For this reason, a clear comparison of the number of cyber incidents is not possible. Austria, for example, did not mention the number of cyber incidents at all in its cyber security report.

5. Cyber Defense of the Individual

This part of the article focuses on the cyber defense of individuals against phishing attacks. The result of research on the phishing attacks in the previous two years has collected the set of almost 400 phishing emails that were sent to the email inbox of one of the authors. The emails were the subject of analysis, the necessary knowledge was obtained, and that was used for the cyber defense of individuals. The functions of the mail client MS Outlook were used for individual defense.

One of the options for individual cyber defense against phishing emails is the possibility to block all email addresses detected from the content of phishing email messages. Email addresses were obtained from phishing emails using the intelligent function entities extraction of the Tovek [18] software, using the Tovek Agent module, see Fig. 1. The first column in Fig. 1 is the file name (phishing email) and the second column includes email addresses.

A total of 365 emails were analyzed, from which 511 email addresses were extracted, of which only 17 were reused; i.e., 3%. It is obvious that blocking that volume of addresses with such low efficiency would not be effective in terms of protection against phishing attacks, see Fig. 2; normalized email addresses in the Excel table, and ordered.

Jméno souboru	Entita email
001.docx	fredarosa1000@gmail.com
002.docx	interpet.sales@gmail.com
003.docx	p003williams@gmail.com
004cz.docx	
005cz.docx	
006cz.docx	
007.docx	admin@hiflyingllc.com
008.docx	info@info.com
009.docx	exxonmobiloilemployment@hotmail.com, exxonmobiloilgas-usa@engineer.com
010.docx	monicag.narula11@gmail.com
011.docx	alexdaniel0982@gmail.com
012.docx	abdul.hamza03@gmail.com, markmarketing922@yahoo.com
013.docx	lawrrence1020@gmail.com
014.docx	mrpeteralexandra3033@outlook.com, mrpeteralexandra3@gmail.com

Figure 1: Entity extraction

361.docx	futuresgrate@gmail.com
105cz.docx	gabinete@secult.rr.gov.br
108sk.docx	gabinete@secult.rr.gov.br
144cz.docx	gabinete@secult.rr.gov.br
075.docx	gambare@kitagami.gr.jp
055.docx	gatestephen@yandex.com
179.docx	gatul8836@gmail.com
036.docx	genbellalogan928@gmail.com
348.docx	genhermi@gmail.com
094.docx	georginabaker13@gmail.com
226.docx	gianggn_mttq@bacgiang.gov.vn
140.docx	glbifcu241@gmail.com
254.docx	gloj0027231@gmail.com
170.docx	glorapeelaez@gmail.com
314.docx	godwinnpeter@gmail.com
061sk.docx	gontar.mc@os.chernigiv-rada.gov.ua
068cz.docx	gontar.mc@os.chernigiv-rada.gov.ua

Figure 2: Extracted email addresses

Another possibility of individual cyber defense against phishing emails is the use of rules in the MS Outlook client to filter incoming emails. We have chosen rules for email detection based on the occurrence of keywords in the subject and content of the email message. The keywords were taken from the results of research on phishing emails, published in the article [19]. Keywords that characterize the particular email segment:

- BUSINESS (investment, project, contract, business, intention, export, service, product, partner, relationship, cooperation, employment, recruit, benefit, inquiry);
- FUND (deposit, fraud, scam, credit, compensation, inheritance, award, sum, price, prize, claim, winning);
- TRANSFER (transfer, property, gold, diamonds, box, packet, shipment, airport, bank, payment);
- CHARITY (charity, donate, Christ, God, sister, brother, promise, illness, disease, hospital, cancer, widow);
- OTHERS (offer, loan, communicate, message, response, contact, friendship, package, undelivered).

The rule is easily set up using the wizard in MS Outlook see Fig. 3. An overview of the rules is shown in Fig. 4.

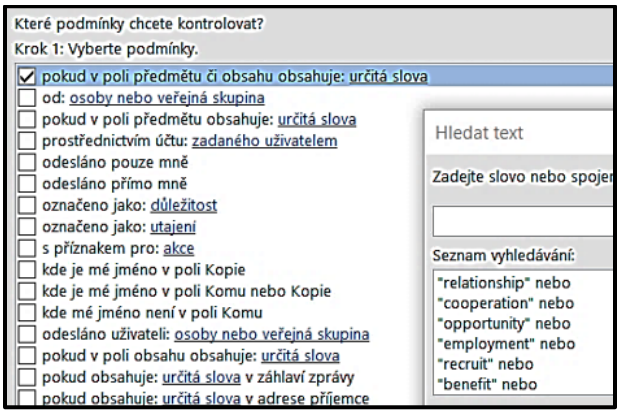


Figure 3: Setting email message filtering rules

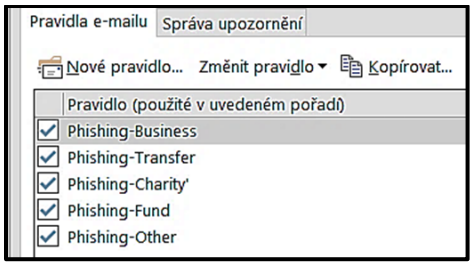


Figure 4: Overview of email message filtering rules

During the debugging of how the rules work successfully, conditions in the rules were continuously updated, on the basis of which emails were not excluded from phishing (manuscript, conference, journal, editor, publication, System NEWS, Smart Cities, Computerworld, Reuters, webcast, Deloitte, identity, Sophos). These are keywords associated with activities associated with publishing scientific articles and offering professional events. The phishing messages are stored in a dedicated folder; based on the set rules. The testing phase was realized for two weeks and after that followed the four weeks experiment.

The results of the experiment were evaluated, see Tab. 2. There were recognized two types of mistakes:

1. Uncaptured (undetected) phishing email (false negative).
2. Phishing erroneously detected, means the correct email that was included in the Phishing folder (false positive).

All seven false-negative messages were phishing, of which 2 in Czech. Content included in groups OTHERS-3, FOND-2, BUSINESS-1, TRANSFER-1, and CHARITY-0.

Table 2: Statistics of the experiment

Phish-email / Week	1	2	3	4	Σ
Total number	27	36	31	26	120
-of which undetected	1	2	2	2	7
Erroneously detected	1	3	1	2	7
Total detection errors	2	5	3	4	14

On the contrary, none of the false-positive messages was phish; they contained corporate information about the event or information about some facts. They contained some of the keywords for inclusion in the phishing folder and then included necessary changes in phish detection rules.

It is also worth mentioning that the number of phishing emails in the experiment increased, compared to the previous ones, by almost 50%. The result of the experiment is promising, correctly captured phishing emails account for 88%.

The described method of defense against phishing attacks can be recommended because it can be individually customized and is easily adjustable. It should be noted that phishing emails are an individual matter. Several years of experimentation have confirmed that the content and extent of phishing attacks against the same person vary a little.

6. Discussion and Conclusion

Cyber security is unfortunately a much-used term these days, and because of the current political situation (Russia-Ukraine conflict), it is not expected that there will be any return to the good old days. The information published in this article is not yet influenced by this conflict, but the main role in it is affected by the Covid-19 pandemic. This pandemic has greatly affected the behavior not only of attackers in the Internet world but also of ordinary users who have started to take the problem of cyber security at least a little bit seriously.

It is interesting to see how the countries mentioned in this article declare their cyber security challenges, and it is important to mention that a detailed comparison of the annual reports was almost impossible in terms of frequency of cyber incidents, structure, and content, as each national report had a different way of assessment. Comparable information was aggregated into Table 1, which shows information regarding the number of cyber incidents, where the United States clearly dominated, as well as the most frequent incidents encountered by those states, and information regarding support for training and cyber exercises.

Individual reports described each type of attack, with one of the most common being ransomware attacks, which caused

hundreds of millions of euros in damage to various state and non-state institutions during the pandemic. Cyber-attacks have most often targeted critical infrastructure, the public sector, the financial sector, industry, healthcare, and, unfortunately, education.

As already mentioned, COVID-19 played a big role in cyber security in 2020, clearly showing how adaptable cybercriminals are. The most common methods of cyber-attacks that were recorded during COVID-19, according to the annual reports of the mentioned countries, were the following attacks:

- Fake news or links to fraudulent sites with disinformation such as there is a miracle cure for COVID-19.
- Fake messages or phone calls pretending to be from companies like Microsoft or Google Drive. Their goal was to extort a password from the user under the premise of offering help or threatening to cancel the account.
- Messages about non-existent packages being delivered.
- Fake appeals to donate money.
- Emails that appear to be from a medical organization.

Subsequently, the biggest threats according to the annual reports include:

- Ransomware - demanding a ransom to recover encrypted data.
- Threats associated with personal data - data breaches/leaks;
- Malware - malicious programs.
- Disinformation - spreading misleading or false information.
- Harmless threats - human error and system misconfiguration.
- Availability and integrity threats - attacks that prevent system users from accessing their information.
- Threats related to electronic mail - e-mail attacks.
- Supply chain threats - attacks (e.g. on service providers) to gain access to customer data.

The main contribution is analysis of cyber security documents, their comparison and evaluation in individual areas. There is a benefit for learning and developing a perspective on security requirements that follow from the findings of the analysis. An experiment was carried out on the subject of phishing, which, based on the use of MS Outlook rules, was able to significantly reduce phishing emails. This approach is original.

Acknowledgment

The article represents the results of the research in cyber security of the project [20] at the University of Defence in Brno, Faculty of Military Technology, Department of Informatics and Cyber Operations.

References

[1] K. Halouzka, L. Burita, P. Kozak, "Overview of Cyber Threats in Central European Countries," in 2021 Communication and Information Technologies Conference Proceedings, Liptovsky Mikulas, Armed Forces Academy of General Milan Rastislav Stefanik in Liptovsky Mikulas, 81–86, 2021.

[2] A. Marrone, E. Sabatino, "La difesa cibernetica nei Paesi NATO: modelli a confront, Cyber Defence in NATO Countries: Comparing Models," Rome, Senate, 2020, [Online]. Available: http://www.parlamento.it/documenti/repository/affariinternazionali/osservatorio/approfondimenti/PI0164.pdf.

[3] A. Jacuch, "Comparative analysis of cyber security strategies, European union strategy and policies. Polish and selected countries strategies", Online Journal Modelling the New Europe, **37**, 102–120, 2021.

[4] US Congressional Research Service, Defence Primer: Cyberspace Operations, https://sgp.fas.org/crs/natsec/IF10537.pdf.

[5] M. Song, D.H. Kim, S. Bae, S.-J. Kim, "Comparative Analysis of National Cyber Security Strategies using Topic Modelling," in International Journal of Advanced Computer Science and Applications, **12**, 62–69, 2021, doi:10.14569/IJACSA.2021.0121209.

[6] R. Elamiryan, R. Bolgov, "Comparative analysis of cyber security systems in Russia and Armenia: Legal and political frameworks," in Digital Transformation and Global Society, **858**, 195–209, 2018, doi:10.1007/978-3-030-02843-5_16.

[7] J. Fulmer, "Complete Guide to Phishing Attacks: What Are the Different Types and Defenses?," 2022, [Online]. Available: https://www.esecurityplanet.com/threats/phishing-attacks/.

[8] M. Dahan, "Hacking for the homeland: patriotic hackers versus hacktivists," in ICIW 2013 Proceedings of the 8th International Conference on Information Warfare and Security, 51–57, 2013.

[9] SANS, "Phishing Strategic Planning Document," 2022.

[10] NÚKIB, "Zpráva o stavu kybernetické bezpečnosti za rok 2020," 2021.

[11] NBÚ, SK CERT, "Správa o kybernetické bezpečnosti v Slovenskej Republice za rok 2020", 2021.

[12] Cybersecurity Report 2020, Austria, Vienna, 2021.

[13] NASK PIB/CERT Polska, "Security landscape of the Polish Internet, Annual report from the actions of CERT Polska 2020," 2021.

[14] Federal Office for Information Security, "The State of IT Security in Germany in 2020," 2021.

[15] National Cyber Security Centre, Annual Review 2020, UK, 2021.

[16] NSA United States of America, NSA cybersecurity year in review for 2020, 2021.

[17] FISMA FY 2020 Annual Report to Congres, The State of Federal Cybersecurity, Federal Information Security Modernization Act of 2014, 2021.

[18] Tovek, "The text analytical software TOVEK," 2022, [Online]. Available: https://www.tovek.cz.

[19] L. Burita, P. Matoulek, K. Halouzka, P. Kozak, "Analysis of phishing emails," in AIMS Electronics and Electrical Engineering, **5**(1), 93–116, 2021.

[20] DZRO FVT 2_KYBERSILY, Research project Cyber forces and resources, University of Defence, Faculty of Military Technology, Brno, Czech Republic, 2021.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 3, 106-115 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Secured Multi-Layer Blockchain Framework for IoT Aggregate Verification

Ming Fong Sie, Jingze Wu, Seth Austin Harding, Chien-Lung Lin, San-Tai Wang, Shih-wei Liao*

Department of Computer Science and Information Engineering, National Taiwan University, Taipei, 106, Taiwan

ARTICLE INFO

Article history:
Received: 27 March, 2022
Accepted: 25 May, 2022
Online: 04 June, 2022

Keywords:
IoT
Blockchain
ID-based Signature
Aggregate Verification
Digital Provenance
Supply Chain Management

ABSTRACT

Technologies designed for digital provenance, especially the Internet of Things (IoT) and blockchain, may allow for security, transparency, and traceability in the global supply chain. However, upstream nodes in the supply chain that work for large-scale production suppliers are not considered. In addition, most IoT blockchain systems adopt an ID-based signature scheme that may affect the efficiency of IoT devices. We propose using aggregate verification to improve the security and efficiency of ID-based verification, reduce network traffic on the blockchain, and transfer computing overhead to aggregator nodes. This paper implements a multi-layer blockchain for Agriculture 4.0 supply chain management that has higher efficiency, effectiveness, and security in comparison to conventional blockchains. We design a Multi-Layer Aggregate Verification (MLAV) solution to improve supply chain management with IoT Blockchain for Agriculture 4.0 through the following methods. First, we use a multi-layer IoT blockchain system to reduce Ethereum gas fee. Second, we design an ID-based Aggregate Verification scheme, thereby eliminating the certificate management cost in the traditional Public Key Infrastructure (PKI) and reducing bandwidth and computation time requirements. Third, we implement a three-layer blockchain infrastructure. In Layer 1, IoT devices sense and upload data to the system's database; in Layer 2, smart contracts execute aggregate ID-based signature verification from IoT devices and upload the transactions to the private blockchain; in Layer 3, a batch converts the layer 2 data and uploads its Merkle root to Ethereum, thereby reducing the required gas fee.

1 Introduction

This paper is an extension of a work originally presented in BRAINS 2021 [1] that uses a multi-layer architecture designed to facilitate smallholders in joining an agricultural blockchain infrastructure. We use aggregate verification to solve the efficiency bottleneck of ID-based signature verification. We also lay out the framework for distributed supply chain management for access control with smart contracts to allow the smallholder to gain access to loans.

With the ushering in of Agriculture 4.0, in recent years there has been widespread adoption of technologies such as the Internet of Things, big data, artificial intelligence, cloud computing, and remote sensing [2]. Agriculture 4.0, also known as digital farming or smart farming, has been brought about by combining telematics and data management with known precision agriculture concepts. These changes have improved the accuracy and practicality of farming operations [3].

However, regarding data management, food safety, and quality monitoring of the agricultural supply chain, Agriculture 4.0 still has significant shortcomings. Particularly in the COVID-19 pandemic, it is essential to prevent cross-contamination and food pathogen outbreaks. The agricultural supply chain should have data transparency, and there should exist a high level of traceability from source to consumer. The World Government Summit in 2018 pointed out that Agriculture 4.0 will need to focus on both the demand side and the value chain (supply) side of the food equation to use technology to meet the real needs of the consumer and to re-engineer the value chain [4]. The tamper-proof property and transparency of blockchain technology allow it to meet these requirements effectively. IoT Blockchain is highly suited for ensuring traceability and consistency of information generated by IoT devices.

Large companies have already begun to adopt distributed supply chain management systems with blockchain technology. In 2017, Walmart established the Walmart Food Safety Center in Beijing and invested US$25 million to use IBM's blockchain solutions to build a global food safety system [5] already tracking 1,500 items on the supply chain blockchain in 2021. Proof of Concept (PoC) and blockchain pilot projects have been established in the United

*Corresponding Author: Shih-wei Liao, Email: liao@csie.ntu.edu.tw

States and China for two products: mango slices and fresh-cut pork products [6]. The cost of supply chain management systems is very high, and there are restrictions on their scalability; therefore, IoT blockchain systems are only suited for large companies. We believe that the benefits brought about by IoT Blockchain should serve not only high-level large-scale food and agricultural suppliers but also grassroots farmers and food producers in providing customized platforms and advice. Blockchain is becoming ever-more democratized and is decentralized in nature; it is therefore highly suited to providing equal opportunity for traditionally disadvantaged entities.

Therefore, the contributions of this paper can be outlined as follows. An Agricultural Supply Chain Finance operational procedure is created with smart contracts that define the rules and functions of three types of nodes: farmer, distribution channel, and financial institution. A corresponding IoT blockchain system is then designed and implemented to record valuable data on both production activity and order history of farmers. There are two types of IoT devices in this system: farming sensors and mobile phones. The farming sensor is a custom-designed piece of hardware that is used in an agricultural setting; for mobile phones, a custom-designed Android app is developed that connects to our IoT blockchain system. Batch verification is leveraged by our IoT Blockchain system to increase ID-based signature verification.

The remainder of this paper is structured as follows: In Section II, we summarize related works. In Section III, we discuss the benefits of using our multi-layer blockchain. Section IV defines and details the aggregate verification algorithm we use. In Section V we describe the blockchain management framework and system architecture. In Section VI, we propose the implementation and evaluation. In Section VII, we have our conclusion.

2 Related Works

2.1 Distributed Ledger Technology

Distributed ledger technology (DLT) is a shared transaction ledger technology that can store, distribute, and exchange certifications publicly or privately between peer entities. In the context of the blockchain, a distributed ledger records transactions between participants and nodes. This data can be duplicated and synchronized across decentralized peer-to-peer networks with consensus algorithms. DLT may then facilitate the flow of information between nodes and help to resolve inefficiencies relating to information asymmetry[7] [8].

Blockchain technology is a specific type of DLT that was developed in 2008 for the implementation of the cryptocurrency Bitcoin [9]. In the Bitcoin network, by leveraging the Proof of Work (PoW) consensus mechanism, blocks are added to a linearly growing, chronologically ordered blockchain. Each block contains the timestamp, transaction data, and hash value of the previous block.

2.2 Ethereum Layer 2 and Smart Contract

Our platform uses the Ethereum blockchain, an open-source, public blockchain structured around a decentralized Ethereum Virtual Machine (EVM) that may process smart contracts [10]. Data agreements are reached via the Proof of Work consensus algorithm. Ethereum 2.0 is expected to switch to the Proof of Stake (PoS) consensus algorithm, which may significantly reduce computational resources wasted during mining and prevent attacks from application-specific integrated circuits (ASIC) [11]. In addition, Ethereum layer-2 technologies such as Arbitrum could collect transactions off-chain and batch it on-chain, scaling up transaction speed, improving privacy, and holding EVM compatibility by using Arbitrum Virtual Machine (AVM), all while still benefiting from layer-1 security [12].

The concept of smart contracts was first proposed in [13]. Its original idea was centered around a ”computerized transaction protocol that executes the terms of a contract.” Developers began integrating this innovative concept into the blockchain environment. The idea of a smart contract has become ”executable code that runs on top of the blockchain to facilitate, execute, and enforce an agreement between untrusted parties without the involvement of a trusted third party” [14].

2.3 IoT Blockchain in Agriculture 4.0

IoT Blockchain is a new technology that integrates the Internet of Things with blockchain technology. It features the following advantages [15] in Agriculture 4.0:

- Transparency for participating companies [16]: New levels of transparency and visibility are essential for improving product traceability and ensuring product authenticity and legality [17, 18].
- Food safety and quality monitoring [19]: A real-time food tracking system built on blockchain technology provides an information platform that enables all supply chain members to access all information, thereby providing openness, neutrality, and reliability for the food supply chain[20].
- Promoting the digitization and disintermediation of the supply chain: Blockchain reduces verification and transaction costs by eliminating intermediaries [21]. By replacing trade financing (banks acting as financial intermediaries) with a blockchain platform, processing time may be reduced from between 7 and 10 days to between 1 and 4 hours [22].
- Improving data security of information sharing: Centralized databases may be prone to data loss or have data that is difficult to retrieve [23]. All data in the blockchain is immutable because the order of transactions is stored in chronological blocks and broadcast to all nodes [24]. The stored data is tamper-proof because updating and deleting of transactions is determined by the consensus mechanism[25, 26].
- Agricultural Finance: Blockchain technology can implement fast, real-time payments for agricultural financial services that increase cash flow and working capital while reducing transaction costs and risks [27].

Ethereum is the first Turing-complete blockchain framework that enables smart contract integration [10]; given enough time and memory as well as the necessary instructions, smart contracts can solve any computational problem no matter the complexity. There

are numerous advantages of smart contracts on the blockchain over traditional contracts.

- The content of the contract is open-source, transparent, and tamper-proof. The immutable smart contract code is guaranteed to execute, which reduces the occurrence of fraud.
- Higher efficiency: using a programming language, there are almost no misunderstandings or disputes, and consensus may easily be reached.
- No third-party arbitration is required. The system automatically executes according to the smart contract, thereby reducing time and verification costs.

2.4 IoT Blockchain

Massive amounts of data are generated from IoT devices and stored in the cloud. It is also an emerging security concern. Blockchain-based IoT systems might solve this security problem in distributed infrastructure [28]. The user could integrate with PKI to access data from IoT devices and interact with blockchain miner nodes to keep track of every transaction on chain [29]. In [30], the author introduce multi-layered network architecture by defining the IoT device layer, router layer, cloud compute layer, offers an authentication framework, and reducing IoT network burden also improves transaction throughput and security.

2.5 Supply Chain Finance

This research focuses on the agricultural supply chain, which is highly competitive. Large-scale downstream enterprises hold an advantageous position. They place heavy requirements on upstream suppliers for purchases, prices, and payment conditions, causing massive pressure on upstream suppliers. Most of these upstream suppliers are small-scale enterprises or farmers; it is difficult for suppliers to obtain loans from financial institutions, resulting in limited funding for upstream entities. A new set of modes has been developed to solve this problem: Agricultural Supply Chain Finance. Agricultural Supply Chain Finance is a proposed framework for financial support for upstream suppliers that promotes the establishment of long-term strategic, synergistic relationships between upstream suppliers and major distributors and improves the competitiveness of the entire agricultural supply chain.

The lack of mutual trust in the agricultural supply chain stems from one primary source: agricultural production suppliers rarely keep financial records. Since most agricultural operations are tax-exempt, farmers neither pay taxes nor report taxes. Therefore, most farms in Taiwan do not have records of either production or sales. As a result, when applying for loans from financial institutions, it is difficult for these agricultural production suppliers to provide sufficient financial data; coupling this with the lack of sufficient existing credit information, financial institutions are often unwilling to lend to these suppliers.

2.6 Aggregate Verification

Aggregate verification is a computationally efficient method for verifying a large number of digital signatures quickly. It is more computationally efficient than individual verification of each signature. With large quantities of data rapidly generated by a variable number of IoT devices, we adopt an aggregate verification method that may be used even in cases of high traffic in the network.

Among the most important considerations in implementing blockchain technology is the measurement of time between uploading data and attaining immutability (confirmation time). For systems with a large number of nodes frequently generating new data, aggregate verification may be implemented to efficiently verify many signatures simultaneously in one action [31]. One such example of this form of implementation may be in IoT devices which require periodic firmware and software updates. In this case, aggregate verification may be implemented for future-proofing the devices by evolving to internet threats, fixing functionality by releasing firmware updates [32]. Upon release of updated firmware, a group of distributed IoT devices may have an update securely and remotely installed automatically rather than being individually, manually installed by the owner of each device.

In the case of immediate verification of items uploaded to the blockchain, some blockchains may choose to require fewer block confirmations for signature verification, thereby compromising security [33]. Therefore, in addition to the aggregate verification case described above, a new aggregate verification application has been proposed in [31] in which a system may need to verify a smaller number of signatures with high efficiency and minimized confirmation time. For example, an IP camera surveillance system may have cameras with a framerate of 15 FPS, with each device generating approximately 1.25 million images in one day. Digital signatures may be used to quickly verify large amounts of pictures and uploaded to the blockchain in large clusters.

3 Benefits of Multi-Layer Blockchain

This section describes the advantages of using our multi-layer blockchain on the system level and on the framework level.

3.1 Multi-Layer Blockchain System

With the application of blockchain technology in the supply chain, corporations and large upstream manufacturers simplify collecting information and production-related data for all transactions in the supply chain (e.g. IBM's global food safety system), assisting institutions in strengthening risk management and control. However, the disadvantaged small upstream producers do not benefit from blockchain technology. Low-income smallholders who operate without a working contract are among the highest upstream entities in the agricultural supply chain; they and the information related to their production activities are essentially nonexistent or anonymous on the supply chain. As participants in the supply chain, they have limited or no access to the information in the blockchain system.

Following our previous work [1], we create a platform to facilitate these small upstream entities in joining the blockchain system; the overview of our platform is shown in Fig. 1. We propose the Certificateless verification process on a client node based on aggregate verification, identity-based cryptography, and zero-knowledge cryptography, which resolves the critical escrow problem and secures the

proof. The Key Generation Center (KGC) constructs only the partial private key of the IoT device. The IoT device generates the entire private key by choosing a piece of secret information and combining it with the partial private key constructed by the KGC. The system parameters published by KGC compute the corresponding public key of the IoT device, and the IoT's secret information is chosen by itself. In addition to our multi-layer platform, we also develop an Android app. The outermost layer of our platform is the Database layer which allows app users to record and read production-related data. The Database layer also provides storage for IoT-generated data for large manufacturers. In order to ensure the tamper-proof property and connectivity of all product information, the Smallholder Node and the Aggregator Node upload information to the Quorum chain in the form of a hash. As Quorum is a permissioned blockchain and lacks information transparency, the data is then uploaded to Ethereum for public access [34]. In order to reduce the gas fee and increase the blockchain recording rate in Ethereum, we convert the data for each product unit into Merkle tree format and only upload the Merkle root on Ethereum. The detailed technical process description may be found in Section V.

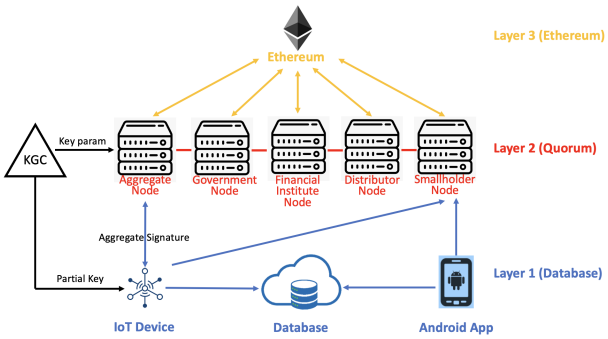


Figure 1: Multi-layer Platform: Blockchain View

3.2 *Supply Chain Management Framework Based On Smart Contracts*

When operating blockchain-based supply chains, the blockchain system must define a set of rules to determine the roles, responsibilities and read-write permissions of participants, which involves maintaining a distributed, authenticated, and synchronized ledger of transactions [35, 36]. Traceable transactions with roles could ensure the stability of the supply chain system [36].

Our supply chain framework is designed to benefit smallholders equally; a blockchain system with smart contracts clearly defines the different rules, responsibilities, and relations of the three types of nodes: Upstream Producer, Financial Institution, and Distribution Channel. The detailed operation process description may be found in "Operation Procedure" in Section V.

With the multi-layer blockchain framework, small upstream entities may upload to the blockchain and access their production activity records, contracts, and transaction records while also solving the largest hurdle: financing. In the past, due to factors such as low fixed assets, low equivalent collateral, and no access to personal financial information, it has proven difficult for smallholders to gain the trust of financial institutions or to obtain any form of funding channel. Through the application of IoT Blockchain, the upstream producer's production activity data and historical order information may be provided to third-party entities for evaluation. This ultimately provides a means by which they may gain trust and become eligible for loans from financial institutions.

Blockchain is increasingly being viewed as a viable solution for alleviating the complexity of global supply chains [20, 37]. The solutions we provide may also be applied to international trade and global supply chains. Addressing the eight technical and non-technical challenges summarized by the OECD [38], our implementation framework and blockchain system may realize the goal of both Inclusion of Informal Actors and Governance.

4 Verification Process

This section presents the details of our verification process on a client node (farmer node or aggregator node) based on aggregate verification, identity-based cryptography, and zero-knowledge cryptography. The aggregator node scheme includes system setup and registration phases, data uploading procedure, aggregate verification, and performance simulation.

4.1 *Aggregator Node*

Smallholders only need to upload a few pictures and messages to the chain; there is no real-time requirement, and they do not need to consider its writing efficiency or computational cost. Smallholders may register as a client node and perform write/read operations to the blockchain through our Android app.

For large suppliers (such as central kitchen) nodes, one node may have hundreds of IoT devices operating simultaneously, with huge amounts and multiple types of complex data simultaneously uploading to the chain; special consideration must be made for throughput, security, and the tamper-proof property.

Therefore, we design an aggregator node; IoT devices are not directly connected to the blockchain but send messages and signatures to an aggregator node which is a relay device that holds powerful computing resources for the network. After executing aggregate verification, the aggregator node then uploads the data to the chain. The advantages of using an aggregator node include:

- Reducing computational cost: IoT devices have limited computing resources, and many devices that use batteries have power restrictions. Therefore, we place energy-consuming operations (uploading procedure) on the aggregator node after centralizing the signature and message to ensure durability in IoT device operations.
- Meet real-time requirements: The message may be quickly uploaded to the chain after aggregate verification, avoiding the possibility of human tampering.

4.2 *System Setup and Registration*

When an IoT device or a client node (aggregator or farmer) first enters the system, the System Setup and Registration processes are executed. According to the aforementioned multi-layer architecture,

data uploading to the chain takes place on Layer 2: Quorum. A Quorum chain is a consortium blockchain; its implementation includes an identity-based signature scheme (IBS). A short public identifier such as an IoT device's MAC address may be used as a verification key. The following system parameters are defined in a tuple: (q, P, G1, G2, e) [39]. The notations and parameters throughout this paper are listed in Table 1.

Table 1: Notation and Parameters

Notation	Description
RID	Real Identity of the device
q	Prime order
P	Generator of G1
G1	Additive group of q
G2	Multiplicative group with same order as G1
e	G1 x G1 → G2, a bilinear mapping
h()	One-way hash function like SHA3-256
H1()	MapToPoint hash function $H1 : \{0,1\}^* \rightarrow G_1$
⊕	Exclusive or
DM	IoT device
MT	Model type
Chk	Checksum generated by SHA3-256
BT_seed	BitTorrent seed of the data

Next, the IBS selects two randomly generated numbers c_1 and c_2 as a pair of secret master keys and generates their corresponding public keys PK_{MAC1} and PK_{MAC2} as follows:

$$\begin{aligned} &c1, c2 \in Z_q^* \\ &PK_{MAC1} = c1 \times P \\ &PK_{MAC2} = c2 \times P \end{aligned} \tag{1}$$

The following hash functions are formed [40]:

$$\begin{aligned} &h : \{0,1\}^* \rightarrow \{0,1\}^f \\ &H1 : \{0,1\}^* \rightarrow G_1 \end{aligned} \tag{2}$$

The system publishes <q, P, G1, G2, e, h, H1, PK_{MAC1}, PK_{MAC2}>.

Upon entry of an IoT device or a client node, our platform must also execute a key generation procedure. The Hardware Security Module (HSM) stores system parameters c_1, c_2 and generates the pseudo-identity and its corresponding public and private keys for the client node by executing the following three procedures.

- Identity Authentication: Each device must first input <RID, fp, or pw> where RID refers to the real ID of the farmer, fp refers to fingerprint, and pw refers to password. This step is necessary for activating the device's HSM.
- Sub-Identity Calculation: HSM selects $t \in Z_q^*$ and computes the first sub-identity $SID1 = tP \in G1$ as well as the second sub-identity $SID2 = RID \oplus H1(t \times c1 \times PK_{MAC2})$.
- Key Calculation: To finish the procedure, HSM calculates the corresponding private key $PrK1 = c1 \times SID1$, $PrK2 = c2 \times H1(SID1 \oplus SID2)$. After Identity Authentication, Sub-Identity Calculation, and Key Calculation procedures have been executed, HSM outputs <ID = (SID1, SID2), PrK = (PrK1, PrK2)>.

After Identity Authentication, Sub-Identity Calculation, and Key Calculation procedures have been executed, HSM outputs <ID = (SID1, SID2), PrK = (PrK1, PrK2)>.

4.3 Data Uploading Procedure

IoT device D uses off-chain programs to download device information DM from BT_seed, and computes DM's checksum. IoT device D uploads new data to an aggregator node according to the following procedure.

- D prepares device DM and retrieves model type MT
- DM collects data, recording TimeStamp and checksum Chk generated by h(DM).
- D stores data into a peer-to-peer file sharing system such as BitTorrent or IPFS while also acquiring the BT_seed of DM.
- D provides a signature as follows:

$$\begin{aligned} v &\leftarrow h(ID \parallel MT \parallel TimeStamp) \\ w &\leftarrow h(ID \parallel Chk \parallel BT_seed) \\ \sigma_D &= (v \times PrK1) + (w \times PrK2) \end{aligned} \tag{3}$$

Finally, D sends <ID, MT, TimeStamp, Chk, BT_seed, σ_D > to the aggregator node. The aggregator node then performs aggregate verification. The smart contract checks the validity of the aggregator's signature; if correct, the smart contract records (MT, TimeStamp, Chk, BT_seed) onto the blockchain.

4.4 Aggregate Verification

In our implementation of aggregate verification, we apply a new signature verification method that allows for superior efficiency and zero-knowledge proof. The algorithm for verifying the signature is proposed in Algorithm 1. An aggregator node receives a new transaction sent from D in the following format: <ID, MT, TimeStamp, Chk, BT_seed, σ_D >, and the aggregator node can then verify the signature σ_D with the following equation [32]: $e(\sigma_D, P) \stackrel{?}{=} e(v \cdot SID1, PK_{MAC1}) \cdot e(w \cdot H1(SID1 \oplus SID2), PK_{MAC2})$, where e is the bilinear mapping function [40] as in Table 1. The verification procedure is shown below:

$$\begin{aligned} &e(\sigma_D, P) \\ &= e(v \cdot PrK_1 + w \cdot PrK_2, P) \\ &= e(v \cdot PrK_1, P) \cdot e(w \cdot PrK_2, P) \\ &= e(v \cdot c_1 \cdot SID_1, P) \cdot e(w \cdot c_2 \cdot H_1(SID_1 \oplus SID_2), P) \\ &= e(v \cdot SID_1, c_1 \cdot P) \cdot e(w \cdot H_1(SID_1 \oplus SID_2), c_2 \cdot P) \\ &= e(v \cdot SID_1, PK_{MAC1}) \cdot e(w \cdot H_1(SID_1 \oplus SID_2), PK_{MAC2}) \end{aligned} \tag{4}$$

Algorithm 1: Verify the signature D

Data: <ID, MT, TimeStamp, Chk, BT_seed, σ_D >
Result: *true* or $false$
if $e(\sigma_D, P) == e(v \cdot SID1, PK_{MAC1}) \cdot e(w \cdot H1(SID1 \oplus SID2), PK_{MAC2})$ **then**
  if *Chk of DM is correct* **then**
    upload hash of BT_seed on the blockchain;
    return *true*;
  else
    return $false$;
else
  return $false$;

In the case that an aggregator node receives a large quantity of new data in a short time period, the aggregator node must verify multiple or numerous signatures. Under these circumstances, aggregate verification is applied as shown below. Let us suppose that n distinct transactions must be verified. Each transaction is denoted as < MT^n, $TimeStamp^n$, Chk^n, BT_seed^n, σ_D^n > . The aggregator node can perform aggregate verification on all signatures with the following calculation: $e(\sum_{i=1}^{n} \sigma^i, P) \stackrel{?}{=} e(\sum_{i=1}^{n} v_i \cdot SID_1^i, PK_{MAC1}) \cdot e(\sum_{i=1}^{n} w_i \cdot H1(SID_1^i \oplus SIDi_2^i), PK_{MAC2})$. The verification procedure is shown below:

$$
\begin{aligned}
& e(\sum_{i=1}^{n} \sigma^i, P) \\
&= e(\sum_{i=1}^{n} (v^i \cdot PrK_1^i + w^i \cdot PrK_2^i), P) \\
&= e(\sum_{i=1}^{n} (v^i \cdot Prk_1^i, P) \cdot e(\sum_{i=1}^{n} (w^i \cdot PrK_2^i), P) \\
&= e(\sum_{i=1}^{n} (v^i \cdot c_1 \cdot SID_1^i, P) \cdot e(\sum_{i=1}^{n} w^i \cdot c_2 \cdot H_1(SID_1^i \oplus SID_2^i), P)) \\
&= e(\sum_{i=1}^{n} v^i \cdot SID_1^i, c_1 \cdot P) \cdot e(\sum_{i=1}^{n} w^i \cdot H_1(SID_1^i \oplus SID_2^i), c_2 \cdot P) \\
&= e(\sum_{i=1}^{n} v^i \cdot SID_1^i, PK_{MAC1}) \cdot e(\sum_{i=1}^{n} w^i \cdot H_1(SID_1^i \oplus SID_2^i), PK_{MAC2})
\end{aligned}
\tag{5}
$$

4.5 Performance Simulation

We evaluate the performance of aggregate verification and other cases presented [41]. Cases 1-5 consist of two types of operations: the first is response from a verification node to a requesting node (R1). The second is the response from a response node to a requesting node (R2). R1 is then divided into two sub-cases (Case 1 and Case 2), while R2 is similarly separated into Case 3, Case 4, and Case 5. Each case depends on the type of response node which receives the version-check request message sent from the requesting node. Their proposed blockchain scheme has five different operation cases:

- Case 1: an IoT device sends a request to a verification node that has the newest data to verify whether its data is the newest.
- Case 2: an IoT device sends a request to a verification node that has the newest data to confirm its data is old and then downloads the newest data.
- Case 3: an IoT device first asks its neighbor IoT device to verify the data version and then verifies whether it has the newest data.
- Case 4: an IoT device first asks its neighbor IoT device to verify whether the data has a newer version and sends a download request to a verification node if it does not have the newest data.
- Case 5: an IoT device broadcasts a join verification message to a blockchain network to ask other nodes to join the verification process to check whether the data version of a device is equal to that of its neighbor node.

We define the cost of cryptographic operations required for each verification procedure. The parameters in Table 1 let S be the time of scalar multiplication in G1, P the time of bilinear pairing operation, and the time of MapToPoint hash operation is H. The pairing operation is the most time-consuming of those operations. We only consider these operations, which determine the speed of signature verification, omitting all other procedures such as one-way hash and point addition.

We use parameter size selection for the elliptic curve cryptography scheme [42] to ensure a security level of the 1024-bit RSA algorithm as a benchmark. G is an additive cyclic group of order q (160-bit prime number) on the elliptic curve, and P is the generator of G. The processing time of Tate pairing on an MNT curve with an 80-bit security level, 160-bit q, and embedding degree k=6 running on an Intel i7 3.07 GHz machine in experiment [43]. Table 2 shows the symbol and execution time in milliseconds.

Table 2: Symbols and Execution Time

Symbol	Description	Execution Time
P	Bilinear pairing operation	3.25
S	Scalar multiplication	0.41
H	MapToPoint hash operation	0.13

Table 3: The Comparison of Operations

Case	Operations
Case1	$2S + 1P$
Case2	$2S + 1P$
Case3	$5S + 3P$
Case4	$5S + 3P$
Case5	$6H + 3S + 2P$
Our scheme	$2H + 3S$

Our computational cost-performance comparison and overhead are shown in Fig. 2. In our scheme, there are five components (RID, q, P, G1, and e) that are passed into the smart contract for updating the data. Next, we explain the update message calculation of Case 1 to Case 5 based on [41], which requires more operations than in Case 1; Case 2 includes two decryption operations and one verification operation. The time in Case 3 and Case 4 to process n messages

in five decryption and three verification processes. In Case 5, the data is considered verified if the original verification is confirmed by other nodes; this is done via a six-block confirmation with PoW consensus. The request nodes need to receive six verification messages from other blockchain nodes to confirm the integrity of the data. As a result, the total computational cost is three decryption operations, two verification operations, and six ECDSA verifications.

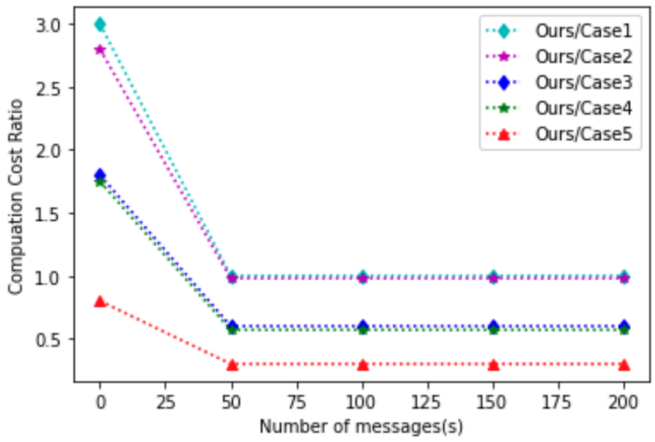


Figure 2: Computation Ratio

5 Management Framework and System Architecture

This section will focus on our blockchain management framework operation procedure and the details of our system architecture. Traditional paper contracts are replaced with smart contracts in the practical process of smallholder agricultural supply chain finance. We digitize the loan application and the information that may prove an upstream producer can reimburse the lender (operational plans, historical transaction records, agricultural production status, current delivery schedule, etc.) and upload this information to the blockchain system. We describe the relationship of the three layers that compose the system architecture.

5.1 Objects and Nodes

There are three main roles: 1. Smallholder (loan applicant) 2. Agricultural financial institution (lending institution) 3. Distribution channel (supermarket or hypermarket)

There are two separate procedures for signing the transaction contract and for the payment verification by the distribution channel: I. The buyer (distribution channel) and the seller (smallholder) sign a smart contract and upload it to Ethereum. II. Smallholders supply to large distributors. After the distribution channels confirm acceptance, they upload the acceptance records to the blockchain. The distribution channel regularly settles the sales volume and pays the smallholders, and at the same time, the sales volume data and payment information are uploaded to the blockchain. An order is not complete until the final payment has been processed.

5.2 Operation Procedure

The operation structure is shown in Fig. 3. The distribution channel (DC) first signs a contract with the smallholder through a smart contract and uploads it to the blockchain after a certain number of parties confirm its correctness. The smallholder must provide other information to financial institutions that may prove their ability to repay the loan, such as business plan books, historical transaction records, and agricultural production conditions to the agricultural financial institution (AFI). After the AFI confirms eligibility and approves the loan, the smallholder begins production. If the DC has paid the deposit, the deposit is then transferred to the smallholder's account in the AFI.

While the smallholder grows crops, production status is regularly uploaded to the blockchain through IoT devices for inspection by AFI and DC. For example, at the i'th production checkpoint, the IoT device uploads photos of the production progress to the blockchain via aggregate verification. Therefore, AFI can make phase i payment to the smallholder based on the production progress ledger. If the photos of the production status do not meet the mark and the smallholder cannot complete production on time, the AFI will recover the funds.

After the smallholder has completed production, they will upload the supply orders to the system when they deliver the produce to the DC. After the DC accepts the produce, the supply orders are uploaded to the blockchain to complete the acceptance of produce. Finally, the AFI makes the final payment to the smallholder, deducting interest and handling fees of the loan according to the previous production progress account book, and settles the remaining receivables.

On our agricultural supply chain blockchain financing platform, all kinds of transaction contracts, supply documents, and production activities are recorded in the blockchain system to allow for agricultural production and sales activities to be transparent and immutable. Relevant information from farmers stored in the blockchain system also proves their repayment ability, providing them with credentials for loans.

5.3 System Architecture

Fig. 4 shows the architecture of the blockchain system for the proposed financing platform. We choose Ethereum as our blockchain system because smart contract functionality is built-in. As shown in the user group located at the bottom of Fig. 4, the IoT device uploads new photos over a set interval. The original data is stored in the database while the following hash value is uploaded to the blockchain: <ID, CT, TimeStamp, Chk, BT_seed, σF >. Other data such as transaction contracts, supply records, and sales records are uploaded to the blockchain in smart contracts. After applying for permission, distribution channels may confirm contract information through the Inspection API. Agricultural financial institutions and agricultural loan reviewing entities may also view the field of agricultural product production materials and inspection documents.

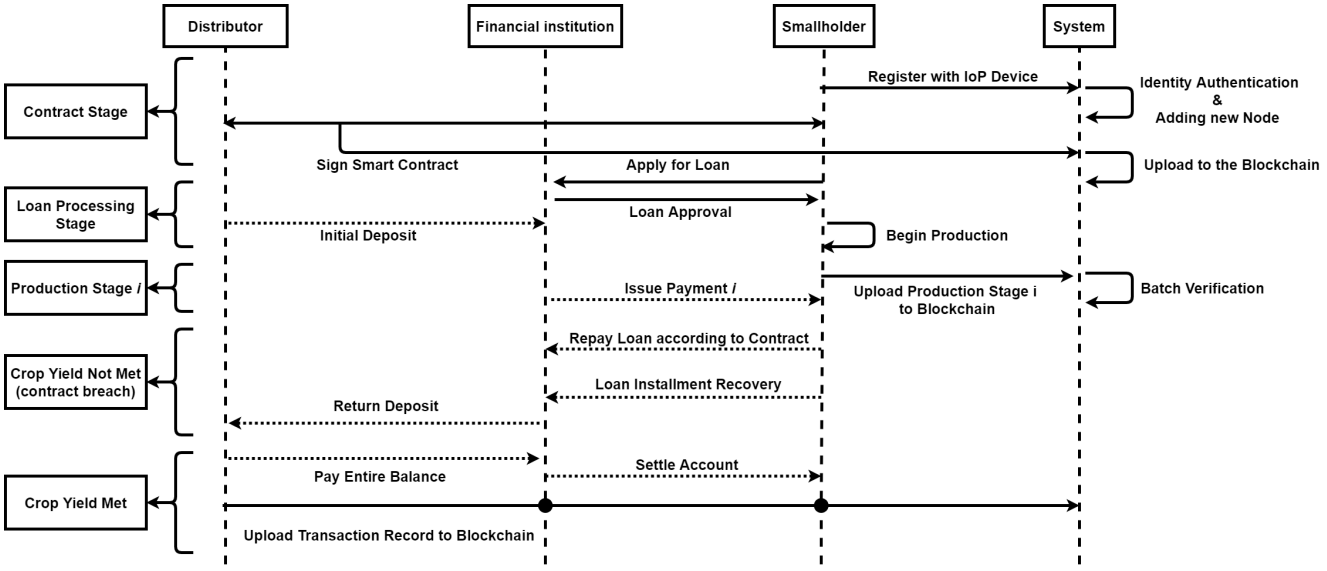


Figure 3: Operation Procedure

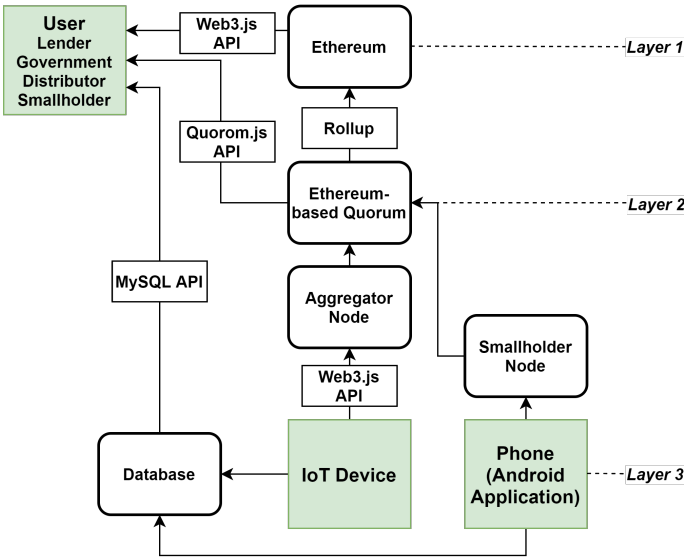


Figure 4: System Architecture

6 Implementation

Layer 2 is built on Quorum (version 2.7.0) with a modified version of the RAFT consensus mechanism [44]. It was initially built on three nodes in different data centers, where each node has 2GB of memory. In addition to fast handling and verification of transactions, this blockchain system must also support frequent on-chain data and contract status queries. Therefore, according to our evaluation results, the on-chain data must be stored in an SSD with at least 5000 read and write IOPS to handle more than 1,000 transactions and queries per second.

In the current situation, each node requires about 10 GB of disk space; it may also be stored in RAM when there is a backup power supply, but the cost is higher. Regarding the parameter settings of the blockchain, to avoid frequent retransmission of transactions between nodes, we set both global slots and global queue block numbers to 76798.

The need for a cheap and easy-to-manufacture device to monitor crops from a close range has arisen from a lack of availability of sensors suited for camera monitoring that have environmental resistance. Most smart farming IoT sensors do not employ the use of cameras. We create a new smart farming IoT device with a simple but robust design that may allow for crop monitoring using a mini-spherical camera: Hi3516CV300 with auto-zoom lens (3x optical zoom) *PTZ 355 degrees left and right, 90 degrees up and down. The camera is mounted to the top of a 1m x 1m x 1m transparent plastic box that crops can grow in.

Our IoT devices include (1) Wireless temperature and humidity sensing devices: measure temperatures ranging from -50C to 120C, and Humidity measurements ranging from 0 percent to 100 percent.; (2) 2.4GHz wireless sensing control device; (3) Weighing transducer: the force or mass is converted into a measurable electrical signal, and the detected weight data is uploaded to the cloud platform through the communication device; (4) Carbon monoxide temperature and humidity sensor (with routing function): Carbon dioxide sensing range: 400 10,000ppm; (5) The above-mentioned camera mount.

The number of transactions per month is shown in Fig. 5. In the case of high transaction volume in the future, it is necessary to adjust the data ratio of each layer of the LevelDB database and use batch parallel reading to increase the access efficiency of the SSD. In addition, to accelerate smart contracts, we also need to rewrite and compile Quorum to use evmos.

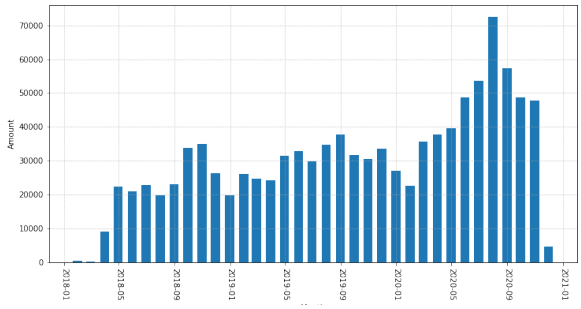


Figure 5: 2018/01 - 2020/12 Number of Transactions

7 Conclusion

In this paper, we propose secured MLAV for an IoT Blockchain architecture. We implement a high-efficiency and high-security agriculture 4.0 framework to allow for smallholders to join the agricultural supply chain blockchain. First, we create the Agricultural Supply Chain Finance operational structure and corresponding smart contracts which define the rules and stipulate the responsibilities and read and write permissions of the three types of nodes in the network (farmers, distribution channels, and financial institutions); second, we design an IoT blockchain system and an IoT camera sensor that allows for upstream producers' production activity data and historical order information to be uploaded to the blockchain. This data will provide guarantees to help them gain trust and qualify for loans from financial institutions; third, through aggregate verification technology, we successfully allow for multiple IoT devices to upload large amounts of data to the blockchain system in a short time. To guarantee security, smallholders can join the blockchain network with high efficiency, high transaction speed, and ID-based signature verification. Through the application of IoT Blockchain, the upstream producer's production activity data and historical order information may be provided to third party entities for evaluation. This ultimately provides a means by which they may gain trust and become eligible for loans from financial institutions.

7.1 Future Work

This paper provides a framework for securely and efficiently generating and storing data on a novel IoT blockchain system in an agricultural supply chain setting. Future work on this research may detail different application scenarios of our IoT blockchain framework in both agricultural and non-agricultural settings. The supply chain considerations, the corresponding smart contracts, and the Operation Procedure in Fig. 3 would need to be changed to fit the application scenario accordingly; however, the IoT blockchain infrastructure and batch verification algorithm may require little to no altering. In addition, future work could focus further on blockchain oracle implementations that allow for guided decisions based on automatic data analysis. An example application scenario in an agricultural setting could be in using our IoT sensor for monitoring of insects on crops. This may hold valuable implications in defining a crop's organic status because in many regions, in order to attain organic status, it must be ensured that insecticides were not used. The number of insects and the type of insects on crops may indicate whether insecticides were used, and therefore confirm the organic status of the crop.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment Ming-Fong Sie, Jingze Wu, and Seth Austin Harding contributed equally to this paper.

References

[1] J. Wu, M.-F. Sie, S. A. Harding, C.-L. Lin, S.-T. Wang, S.-w. Liao, "Multi-Layer Aggregate Verification for IoT Blockchain," in 2021 3rd Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS), 43–44, IEEE, 2021, doi:10.1109/BRAINS52497.2021.9569817.

[2] Z. Zhai, J. F. Martínez, V. Beltran, N. L. Martínez, "Decision support systems for agriculture 4.0: Survey and challenges," Computers and Electronics in Agriculture, **170**, 105256, 2020, doi:10.1016/j.compag.2020.105256.

[3] V. Saiz-Rubio, F. Rovira-Más, "From smart farming towards agriculture 5.0: a review on crop data management," Agronomy, **10**(2), 207, 2020, doi:10.3390/agronomy10020207.

[4] M. De Clercq, A. Vats, A. Biel, "Agriculture 4.0: The future of farming technology," Proceedings of the World Government Summit, Dubai, UAE, 11–13, 2018.

[5] R. Kamath, "Food traceability on blockchain: Walmart's pork and mango pilots with IBM," The Journal of the British Blockchain Association, **1**(1), 3712, 2018.

[6] H. Haswell, M. Storgaard, "Maersk and IBM unveil first industry-wide cross-border supply chain solution on blockchain," IBM, http://www-03. ibm. com/press/us/en/pressrelease/51712. wss, 2017.

[7] J. Nurgazina, U. Pakdeetrakulwong, T. Moser, G. Reiner, "Distributed ledger technology applications in food supply chains: a review of challenges and future research directions," Sustainability, **13**(8), 4206, 2021, doi:10.3390/su13084206.

[8] A. Ellebrecht, "Chain of Custody and Transparency in Global Supply Chains," in Sustainable Global Value Chains, 227–237, Springer, 2019, doi:10.1007/978-3-319-14877-9_13.

[9] S. Nakamoto, A. Bitcoin, "A peer-to-peer electronic cash system," Bitcoin.–URL: https://bitcoin. org/bitcoin. pdf, 2008.

[10] G. Wood, et al., "Ethereum: A secure decentralised generalised transaction ledger," Ethereum project yellow paper, **151**(2014), 1–32, 2014.

[11] V. Buterin, "Ethereum: Platform Review," Opportunities and Challenges for Private and Consortium Blockchains, 2016.

[12] H. Kalodner, S. Goldfeder, X. Chen, S. M. Weinberg, E. W. Felten, "Arbitrum: Scalable, private smart contracts," in 27th USENIX Security Symposium (USENIX Security 18), 1353–1370, 2018.

[13] N. Szabo, "Formalizing and securing relationships on public networks," First Monday, **2**(9), 1997.

[14] M. Alharby, A. Van Moorsel, "Blockchain-based smart contracts: A systematic mapping study," arXiv preprint arXiv:1710.06372, 2017.

[15] Y. Wang, J. H. Han, P. Beynon-Davies, "Understanding blockchain technology for future supply chains: a systematic literature review and research agenda," Supply Chain Management: An International Journal, 2019.

[16] S. Mansfield-Devine, "Beyond Bitcoin: using blockchain technology to provide assurance in the commercial world," Computer Fraud & Security, **2017**(5), 14–18, 2017, doi:10.1016/S1361-3723(17)30042-8.

[17] X. Lin, S.-C. Chang, T.-H. Chou, S.-C. Chen, A. Ruangkanjanases, "Consumers' intention to adopt blockchain food traceability technology towards organic food products," International Journal of Environmental Research and Public Health, **18**(3), 912, 2021, doi:10.3390/ijerph18030912.

[18] F. Casino, V. Kanakaris, T. K. Dasaklis, S. Moschuris, S. Stachtiaris, M. Pagoni, N. P. Rachaniotis, "Blockchain-based food supply chain traceability: a case study in the dairy sector," International Journal of Production Research, **59**(19), 5758–5770, 2021, doi:10.1080/00207543.2020.1789238.

[19] J. Xu, S. Guo, D. Xie, Y. Yan, "Blockchain: A new safeguard for agri-foods," Artificial Intelligence in Agriculture, **4**, 153–161, 2020, doi:10.1016/j.aiia.2020.08.002.

[20] A. Park, H. Li, "The effect of blockchain technology on supply chain sustainability performances," Sustainability, **13**(4), 1726, 2021, doi:10.3390/su13041726.

[21] H. Chen, Z. Chen, Y. Cheng, X. Deng, W. Huang, J. Li, H. Ling, M. Zhang, "Poster: An Efficient Permissioned Blockchain with Provable Reputation Mechanism," in 2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS), 1134–1135, IEEE, 2021, doi:10.1109/ICDCS51616.2021.00123.

[22] Y. Guo, C. Liang, "Blockchain application and outlook in the banking industry," Financial Innovation, **2**(1), 24, 2016, doi:10.1186/s40854-016-0034-9.

[23] R. Kumar, R. Tripathi, N. Marchang, G. Srivastava, T. R. Gadekallu, N. N. Xiong, "A secured distributed detection system based on IPFS and blockchain for industrial image and video data security," Journal of Parallel and Distributed Computing, **152**, 128–143, 2021, doi:10.1016/j.jpdc.2021.02.022.

[24] A. S. Rajawat, R. Rawat, K. Barhanpurkar, R. N. Shaw, A. Ghosh, "Blockchain-based model for expanding IoT device data security," in Advances in Applications of Data-Driven Computing, 61–71, Springer, 2021, doi:10.1007/978-981-33-6919-1_5.

[25] J. Lian, S. Wang, Y. Xie, "Tdrb: An efficient tamper-proof detection middleware for relational database based on blockchain technology," IEEE Access, **9**, 66707–66722, 2021, doi:10.1109/ACCESS.2021.3076235.

[26] X. Fu, H. Wang, P. Shi, "A survey of Blockchain consensus algorithms: mechanism, design and applications," Science China Information Sciences, **64**(2), 1–15, 2021, doi:10.1007/s11432-019-2790-1.

[27] M. Tripoli, J. Schmidhuber, "Emerging Opportunities for the Application of Blockchain in the Agri-food Industry," FAO and ICTSD: Rome and Geneva. Licence: CC BY-NC-SA, **3**, 2018.

[28] N. Kshetri, "Can blockchain strengthen the internet of things?" IT professional, **19**(4), 68–72, 2017, doi:10.1109/MITP.2017.3051335.

[29] A. Dorri, S. S. Kanhere, R. Jurdak, P. Gauravaram, "Blockchain for IoT security and privacy: The case study of a smart home," in 2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops), 618–623, IEEE, 2017, doi:10.1109/PERCOMW.2017.7917634.

[30] S. Varshney, P. Vats, S. Choudhary, D. Singh, "A Blockchain-based Framework for IoT based Secure Identity Management," in 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), volume 2, 227–234, IEEE, 2022, doi:10.1109/ICIPTM54933.2022.9753887.

[31] K. Hakuta, Y. Katoh, H. Sato, T. Takagi, "Batch verification suitable for efficiently verifying a limited number of signatures," in International Conference on Information Security and Cryptology, 425–440, Springer, 2012, doi:10.1007/978-3-642-37682-5_30.

[32] J.-W. Hu, L.-Y. Yeh, S.-W. Liao, C.-S. Yang, "Autonomous and malware-proof blockchain-based firmware update platform with efficient batch verification for Internet of Things devices," Computers & Security, **86**, 238–252, 2019, doi:10.1016/j.cose.2019.06.008.

[33] E. K. Kogias, P. Jovanovic, N. Gailly, I. Khoffi, L. Gasser, B. Ford, "Enhancing bitcoin security and performance with strong consistency via collective signing," in 25th {usenix} security symposium ({usenix} security 16), 279–296, 2016.

[34] J. Polge, J. Robert, Y. Le Traon, "Permissioned blockchain frameworks in the industry: A comparison," Ict Express, **7**(2), 229–233, 2021, doi:10.1016/j.icte.2020.09.002.

[35] A. Shwetha, C. Prabodh, "A Comprehensive Review of Blockchain Based Solutions in Food Supply Chain Management," in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 519–525, IEEE, 2021, doi:10.1109/ICCMC51019.2021.9418305.

[36] Q. Song, Y. Chen, Y. Zhong, K. Lan, S. Fong, R. Tang, "A supply-chain system framework based on internet of things using blockchain technology," ACM Transactions on Internet Technology (TOIT), **21**(1), 1–24, 2021, doi:10.1145/3409798.

[37] J. Liu, H. Zhang, L. Zhen, "Blockchain technology in maritime supply chains: applications, architecture and challenges," International Journal of Production Research, 1–17, 2021, doi:10.1080/00207543.2021.1930239.

[38] J. Tholen, D. de Vries, A. Daluz, C.-C. Antonovici, W. V. Brug, R. Abelson, D. Lovell, "Is there a role for blockchain in responsible supply chains?" in OECD Global Blockchain Policy Forum, 2019.

[39] J. Kurose, K. Ross, "Computer Networking: A Top Down Approach, 2012," .

[40] D. Hankerson, A. J. Menezes, S. Vanstone, Guide to elliptic curve cryptography, Springer Science & Business Media, 2006.

[41] B. Lee, J.-H. Lee, "Blockchain-based secure firmware update for embedded devices in an Internet of Things environment," The Journal of Supercomputing, **73**(3), 1152–1167, 2017, doi:10.1007/s11227-016-1870-0.

[42] E. Barker, A. Roginsky, et al., "Transitions: Recommendation for transitioning the use of cryptographic algorithms and key lengths," NIST Special Publication, **800**, 131A, 2011.

[43] N.-W. Lo, J.-L. Tsai, "An efficient conditional privacy-preserving authentication scheme for vehicular sensor networks without pairings," IEEE Transactions on Intelligent Transportation Systems, **17**(5), 1319–1328, 2015, doi:10.1109/TITS.2015.2502322.

[44] D. Ongaro, J. Ousterhout, "In search of an understandable consensus algorithm," in 2014 USENIX Annual Technical Conference (Usenix ATC 14), 305–319, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 5, 113-128 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Automatic Counting Passenger System using Online Visual Appearance Multi-Object Tracking

Javier Calle*, Itziar Sagastiberri, Mikel Aramburu, Santiago Cerezo, Jorge García

Fundación Vicomtech, Basque Research and Technology Alliance (BRTA), Donostia-San Sebastián, 20009, Spain

ARTICLE INFO

Article history:
Received: 31 August, 2022
Accepted: 08 October, 2022
Online: 25 October, 2022

Keywords:
People-counting
Multi-object tracking
Fisheye camera
Public transport
Visual appearance modelling

ABSTRACT

In recent years, people-counting problems have increased in popularity, especially in crowded indoor spaces, e.g., public transport. In peak hours, trains move significant numbers of passengers, producing delays and inconveniences for their users. Therefore, analysing how people use public transport is essential to solving this problem. The current analysis estimates how many people are inside a train station by using the number of people entering and leaving the ticket gates or estimating the train occupancy based on conventional CCTV cameras. However, this information is insufficient for knowing the train occupancy. The required data includes vehicle usage: how many people enter or leave a vehicle or which door is the most used. This paper presents a solution to the stated problem based on a multi-object tracker with a sequential visual appearance predictor and a line-based counting strategy to analyse each passenger's trajectory using an overhead fisheye camera. The camera selection inside the train was made after profoundly studying the railway environment. The method proposes a module to compute the total train occupancy. The solution is robust against occlusions thanks to the selected tracker and the fisheye camera field of view. This work shows a proof of concept dataset containing pseudo-real case scenarios of people's affluence in train doors recorded by fisheye cameras. Its purpose is to prove the system's functionality in these scenarios. The proposed approach achieved an overall accuracy of counting people getting on and off of 90.78% in the pseudo-real dataset, proving that this approach is valid.

1 Introduction

In the context of security surveillance, people counting is one of the most interesting tasks in analysing crowds. Although it seems like a simple task, when the crowd density is high, the number of occlusions increases the problem's complexity. During the last few years, this problem has risen significantly in the security field because it allows estimating the number of people in a room, thus limiting access to it. However, people-counting algorithms are also used in outside environments to estimate the number of pedestrians in a specific area.

There are two main circumstances where it is practical to count people in an area. The first one is in an outdoor space, where it is usual to use a CCTV camera on a building wall to overlook pedestrians. This case and the technologies related to it are called crowd counting. For this purpose, it is common to use crowd-density approaches. They behave exceptionally well when dealing with a highly dense crowd. On the contrary, they do not work well with small groups of people, where an error of a single person can mean a significant change in accuracy.

This paper is an extension of the work initially presented at the IEEE International Conference on Advanced Video and Signal-Based Surveillance [1]. In this case, the multi-object tracker analyses the people's trajectory to count the number of passengers inside a train. Even though the experiments were held in a train-related environment, the offered solution can work in situations where there is a need to count people crossing a narrow space and a limited height to locate the camera. Examples of this may include building doors (banks, shopping centres, stadiums, companies, etc.), other transport (buses) or multitudinous events (festivals, sports events, etc.).

As the objective is to count the number of people inside the train and analyse the influx of people using it, the crowd density approaches are not accurate enough, so other options are considered. The second case is an indoor area CCTV or fisheye cameras. The choice of the camera will depend on the characteristics of the room. Another approach is to know the flow of people getting in and out of a room, especially if there are multiple doors in an area that a single

*Corresponding Author: Javier Calle, email: jcalle@vicomtech.org

camera can not cover. In this case, it is common to use flow-based or detection-based methods to track the people.

In the context of a train, there is a very particular situation; people can only get on and off at train stops. This means that to know the number of people on the train, it is not necessary to count the number of people at all times, just at the train stops.

Another condition to be considered is that, on a train, it is common to find people blocking the door or entering and stopping in the middle of the door. This causes more occlusions and can generate counting errors if the chosen method is not prepared for having a person at the edge of the door.

The next encountered challenge was choosing the counting method. For this, the strategy used is based on route estimation, thus being possible to conclude whether people enter or leave through the door. A multitracking method with a flow door counting strategy has been chosen. All the people in the scene are tracked. And the method can add those who are entering and leaving. So the passenger flow is calculated by comparing the number of people crossing the train door.

Another important point to tackle this issue is the study of the different cameras that could be used. The study was conducted to have the best possible view of the door at the time of entry and exit to ease the video analysis.

The camera that best suited all of the needs was the overhead fisheye camera, as explained in more detail in Section 3. However, using a fisheye camera entails solving a problem of distortion in the shape of people. It is especially challenging at the train's entrance, as it is common to find a different height between the train and the station. That step causes an acceleration in the person's movement through the camera, exacerbated by the camera's distortion. This causes tracking methods based on movement modelling to lose some targets as the complexity of the target's movement increases in this situation.

Given this, the best-suited tracking method is the one that uses visual appearance to aid in the tracking, but it also needs to be robust against the distortion of the cameras. For the detection and tracking of people, the online multi-object tracking approach with an affinity model from previous work [1] is proposed.

This method is based on FairMOT [2]. Still, the addition of the affinity model has been inspired to deal with the visual appearance transformation of the object while it goes through the camera's field of view. This method will allow accurate tracking of people on the scene, as it is fully prepared to deal with the deformation caused by the fisheye cameras.

The model uses a convolutional LSTM encoder-decoder architecture to learn the space-time transformation metric between consecutive re-ID embeddings extracted from the object trajectory. This allows obtaining the next re-ID embedding by considering the long-term appearance information. Furthermore, the tracking algorithm can also handle temporal occlusions in video sequences by feeding back predicted re-ID embeddings into the affinity model.

With the contributions of this work, four problems are solved:

1. Camera positioning and type of camera. After realising a synthetic simulation-based analysis, an overhead fisheye camera is selected to be installed inside the train.

2. The addition of the visual modelling for the tracking to help the Kalman filter with the issues caused by the distortion of the fisheye camera.

3. Problems caused by the coach step and the train's environment. Solved by the visual appearance-based tracker.

4. Counting people method. The solution is based on all train doors' flow using the tracking results and a counting line.

The report is organised as follows. Section 1 is the current introduction followed by the related work (section 2). This section studies different avant-garde methods to count people based on estimating via area occupancy or trajectory analysis. Additionally, it inspects various people-tracking solutions. Section 3 analyses the railway environment, i.e., it considers the different types and settings of the cameras inside a train car and selects the best option for people counting. Section 4 describes the approach chosen to count people inside a train based on overhead fisheye cameras and the passenger's trajectory. It explores the main stages of the counting people solution, detection, tracking, flow control and occupancy computation. Section 5 presents the obtained results. For this purpose, it first introduces the used datasets for performance evaluation. Straightaway the experimental and global results are given along with a comparison with state-of-the-art works. Section 6 is the conclusion. Here the results are discussed, and conclusions about the people-counting solution are presented. The future work, section 7, describes conceivable improvements and suggestions. Lastly, some acknowledgements for APPRAISE European project are given in Section 8.

2 Related work

As the main method of this article specialises in trajectory-based people counting, this section will first analyse previous SOTA work on occupancy, detection, and trajectory analysis. And it will also present different tracking methods from the SOTA.

2.1 Area occupancy

During the first months of the COVID-19 pandemic, occupancy problems raised their importance as it was essential to prevent disease transmission in closed environments. The main objective was to know how many people were inside an area at a particular moment. These methods are divided into object-detection or density map-based methods. Each person is detected in the first case, while a general estimation is done in the second.

2.1.1 Object based

The approach proposed in [3] uses an overhead video camera. They transform each video frame into a grey-scale picture and subtract the background using the empty scene knowledge. Once verified that the tracked object is a person, they keep track of it, knowing if the person is inside the area.

Another option is the one proposed in [4], where a thermal camera was used to obtain the semantic segmentation for the human. After that, they use a classification model using Adaptive Boosting

(AdaBoost) and a regression model using a shallow neural network to estimate the occupancy.

2.1.2 Density map based

When the pedestrians' density is high, the number of occlusions is also high, and person detection is almost impossible.

In these cases, state of the art is to use methods that estimate a density map and calculate its integral, obtaining the number of objects in the image as [5]–[7]. Also, some "density map" based strategies focus on the most visible part of a human in a crowd, the head, reducing the occlusion effect. The approach made in [8] not only estimates the number of people in a group but localises them with a point in the middle of their heads. They propose a new metric to achieve high localisation errors and counting performances called density Normalised Average Precision (nAP).

2.2 Trajectory analysis

These methods use different strategies to know how many people have entered or left a room. They are divided into image-based or non-image-based methods.

2.2.1 Image based methods

One of the image-based methods is to use a multi-object tracker and analyse the trajectory to count how many people have entered or left a room, as the authors do in [9]. They use a line as a frontier between the inside and the outside of the train. They focus on head detection using a standard overhead camera placed on a train platform roof over the train's door. This method is based on a detector and tracker. Once the trajectory is known, the counting module verifies if the person has crossed the limit line. If the person crosses the line, the counting method checks the direction to determine if the person is entering or leaving the train.

Another similar approach is [10], which tracks heads and uses a frontier line to count people. Their contribution is the analysis of the reference line's height. Alternatively, it is possible to use a region of the image to do the tracking as in [11].

There are different methods to analyse the trajectory. The approach proposed in [12] uses multiple independent lines for counting people. Each line counts the number of people crossing without analysing the direction, just the number of people crossing the lines.

Other methods like [13] propose a strategy that obtains the direction of the trajectories generated by the tracker by computing the angle between the position of the mass centre in the actual frame and its position in a previous frame. In [14], they propose a method consisting of two lines that define an Area of Interest that the person has to cross to be counted. In this case, the direction is determined by which line was crossed first.

In [15], the author defines a Region of Interest (ROI) where they track the objects and obtain the direction by looking at the increasing or decreasing of the y-axis coordinate of the mass centre.

2.2.2 Non image based methods

Some approaches do not use RGB cameras, such as [16], where they use an infrared array sensor. The sensor is equipped with a wide-angle lens that covers 110º and 75º on each axis, so, if the sensor is placed on the ceiling at $2.6m$ the sensor will cover $6x3.2meters$. The method allows one to count and locate the person in each frame with a margin of 0.3 meters. The dataset used contained data from up to three people simultaneously.

Another method was proposed in [17] using two infrared sensors located at a distance of 15 cm between them. The idea is to use a device that counts how many people have passed through a door. Depending on which IR detect the object first, it is possible to know the direction of that object. This method works with environments that assure that only one person is crossing that door, for example, at ticket gates.

2.3 Tracking

A tracker is needed to be able to analyse the trajectories of people, so different options for tracking were studied. Multi-object tracking approaches are typically categorised into offline and online methods. Offline methods can use all frames in a sequence (present, past and future), whereas online methods only use past and current frames for inference. So the only option to analyse people's trajectory currently getting in or out of the train is to use online methods.

Online methods frequently use the tracking-by-detection approach [18, 19]. In order to connect tracklets or detections between frames, recent online multi-object tracking systems use an affinity model in the data association step. Using pairwise affinity ratings, affinity models attempt to account for occlusions and changes in appearance. These methods can be divided into two categories: (i) robust and discriminant re-ID embeddings-based methods and (ii) sophisticated scoring functions-based methods. For the first, siamese or triplet networks are frequently suggested, as appearance cues are essential. The first approach introduced by [20] proposes a combination of the standard region loss with a triplet loss for maximising and minimising the distance between similar and dissimilar identities. In [21], authors propose deep collaborative reinforcement learning under a unified network.

Sophisticated scoring functions-based methods simultaneously output detected objects and their associated re-ID embeddings, and they focus the attention on an affinity metric design. In [22], a siamese network is explicitly designed to estimate the affinity between the detections by adding the object's appearance. In [23], a quadruplet loss is proposed to emphasise both the object's appearance and its temporal proximity. A more recent proposal is the UMA triplet network, proposed in [24] to learn the single object tracking and affinity prediction tasks simultaneously, creating a unified multi-task learning framework.

Regardless of the approach, it should be noted that all of these algorithms only consider short-term temporal appearance information between successive frames. By teaching an appearance transformation metric, our approach promotes the use of long-term temporal appearance information.

3 Railway environment

Passenger counting in train carriages is a very characteristic problem, as people can only enter and exit at train stops. They also move

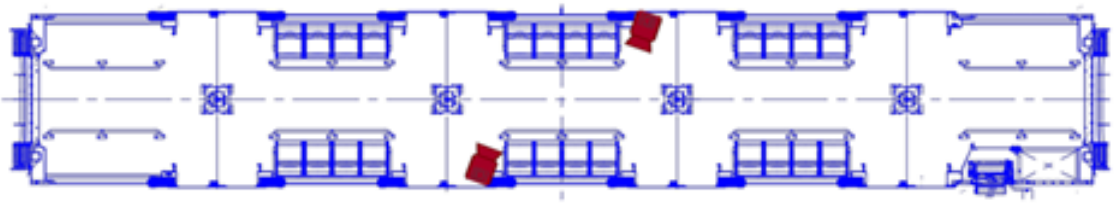


Figure 1: Schema for CCTV use.

together in groups, cross paths, wait at the door until it opens, and move all together at once. Therefore, different situations should be analysed, not only for choosing the best counting strategy but also for setting up the proper type of camera in the correct position to make this task less challenging.

It is important to note that, at certain times of the day, the density of people inside the train can be very high. This implies that the chosen method must be able to operate with a high number of simultaneous objects. In addition, the space inside a coach is minimal, implying that there will be many occlusions. Consequently, finding a counting method that works with high and low densities will be necessary.

While the train is moving, monitoring or tracking people is not necessary, as the train doors are closed, meaning there is no change in the influx of people. Therefore, the trajectory analysis of the users is only necessary when the train stops and the doors to the train open.

As stated before, there is a flow of passengers between carriages in a train, and there is an increased number of occlusions in situations of high occupancy. It would be necessary to count the number of people changing carriages to know the total count in each carriage. The door between carriages is an extra door, which implies a higher number of cameras.

Finally, it is essential to note that it is not necessary to strictly keep each person's ID throughout their journey. It is only needed during their entry or exit of the train, as the objective is to count entries and exits of the carriage through the door. Therefore, considering all this, the number of people will be controlled by focusing on door analysis. To know the total number of people inside the train, counting the number of people entering and exiting at each door is enough. Each door can be analysed separately and then added or subtracted from the people that joined or left the train.

It is crucial to select the best type of camera and its position to make the task easier. So we deeply analysed the possible camera positions and configurations in the wagon to choose the most appropriate for our study.

3.1 Explored camera configurations

It was decided to perform a study based on synthetic simulations to represent the most common and problematic situations in access to a train car. All these simulations were performed with different camera configurations inside the train. Outside will produce more useless tracking due to the people walking near the train but not going in or out. The most representative of which are shown here.

3.1.1 Train's CCTV camera

The first approach is using the CCTV cameras that are already available on the train. For security reasons, these cameras are angled, so the two front doors are checked, and the actions of users are seen to be able to act in case of an emergency, robbery, etc. The layout schematic of the cameras' positions in the coach can be seen in Figure 1.

To check the type of images that would be obtained by this type and the position of the camera, some 3D simulations were carried out. The simulated situations were (i) an empty train with one person entering, (ii) a train with a certain number of passengers already in and a group of people entering through the door.

In the situation where only one passenger is entering, Figure 2, they can be easily monitored. Through analysing images, it is easy to know whether the target is inside or not using this camera.



Figure 2: Example of a CCTV image.

In the second case, Figure 3, there is a greater number of occlusions. Passengers already on the train may occlude the entering people, depending on the distribution of the passengers. If rush hour images were analysed, where the train is almost fully packed before people enter, door visibility would be lost.



Figure 3: Example of a crowded space captured by a CCTV camera.

However, at the moment the group enters the carriage, it is not possible to differentiate whether the passengers in the back have already entered the train or not. There is no visibility of the people in the back of the group. Occlusions with this view will be a big problem, so the position and type of camera must be changed to get a more suitable angle of vision.

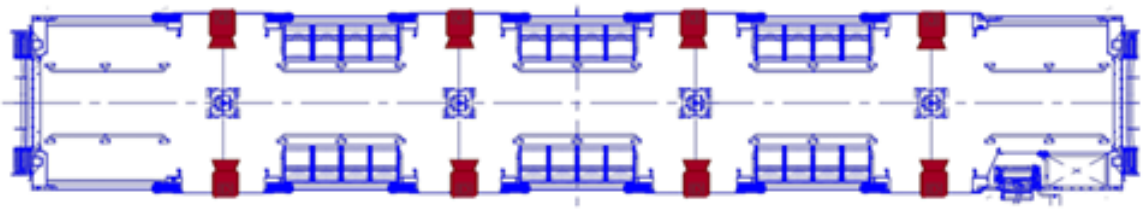


Figure 4: Schema for overhead camera use.

3.1.2 Standard overhead camera

The use of overhead cameras positioned over the doors is considered next. The exact position of the camera is inside the roof of the coach instead of directly above the door, as cameras can not be placed where the door's mechanism is. This means the camera is slightly displaced towards the inside of the carriage. This position also has the advantage of having the camera at a higher altitude to get a wider field of view of the scene. This kind of view can allow seeing the heads of every person going through the door, reducing the number of occlusions, and it is easier to extract the passenger's position with respect to the door. A schematic representing the distribution of the cameras can be seen in Figure 4.

The number of cameras is doubled with respect to the previous case, as there is one camera per door.

Figure 5 illustrates a situation where a single person enters the train. The placement of the camera allows for complete visualisation of the passenger. Having complete visibility of the door line allows for distinguishing whether people entered the train or not. However, the field of vision is relatively small, and we can only see the lower part of the door, but the middle and upper parts are not visible. Moreover, the field of view may not be enough to see the entire door if the door is wider or with a lower ceiling.

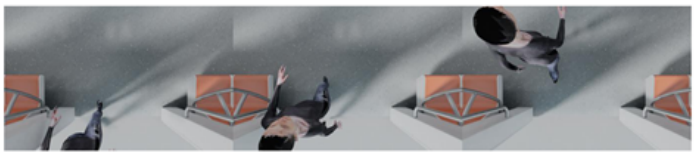


Figure 5: Example of an overhead camera.

For the second case, in Figure 6, it can be observed that the passengers already in the carriage do not generate any occlusion, as most of them are not visible. Passengers entering the carriage do not occlude each other either, but due to the limited field of vision, we have fewer frames to analyse the trajectory of the passengers.



Figure 6: Example of a crowded space captured by an overhead camera.

3.1.3 Fisheye overhead camera

An overhead camera allows a better perspective of the passengers' flow through the door. But this solution with a standard camera has a narrow field of view. So, the fisheye camera has been proposed to solve the lack of visibility. The placement is the same as in the previous case (Figure 4), but omnidirectional 180º field of view cameras are used in this one.

Figure 7 illustrates the first experiment, where the person is completely visible. In addition, the image shows a more extensive area inside and outside the train compared to a regular overhead camera. Seeing a larger area of the outside and the coach allows us to analyse the trajectory of the passengers more robustly, as there are more frames where they are visible.

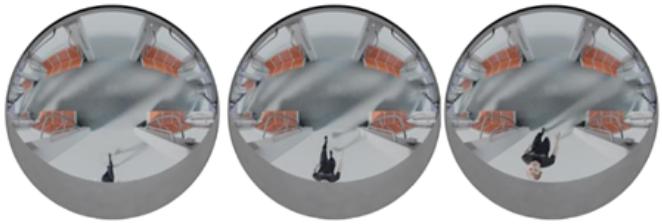


Figure 7: Example of a fisheye overhead camera.

In the second experiment (Figure 8), it can be observed that, similarly to previous cases, occlusions only appear in the case of the group entering the door when they are away from it. At the moment of getting in or out of the carriage, there are no occlusions, although the height of both the person and the camera will affect the perceived distortion.

It should be noted that the camera can only count correctly at the door below the camera, as there will be many occlusions at the opposite door. Hence one camera per door is needed.



Figure 8: Example of a crowded space captured by a fisheye overhead camera.

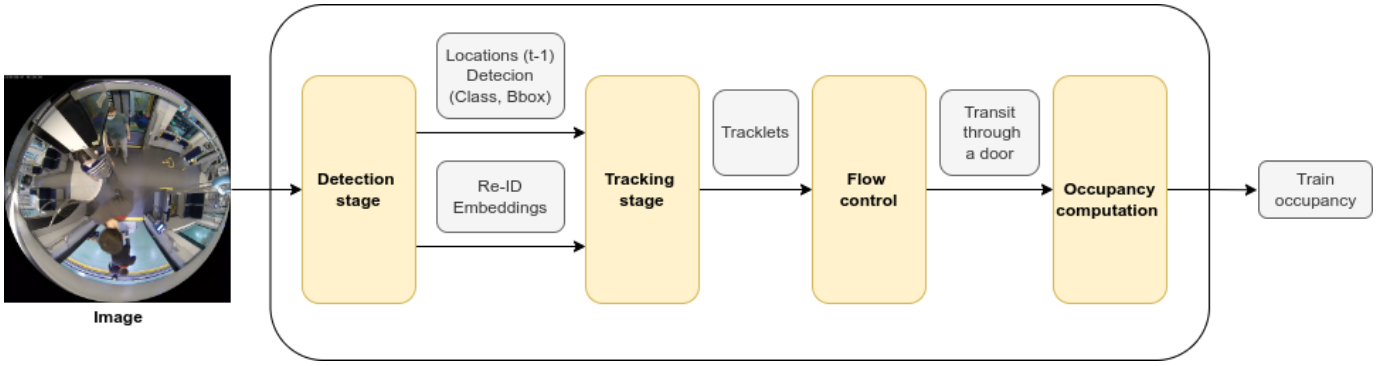


Figure 9: Overview of our overall approach.

3.2 Final configuration of the cameras

The first conclusion is that the camera must be above the door to have a good perspective. Otherwise, the door area is occluded, and the solution can not accurately analyse if a person has boarded or left the train in areas close to the door.

Secondly, after comparing the different types of cameras, it is concluded that both overhead camera options are suitable for the task. As the height of the train ceiling limits the camera location, the best option is to use a fisheye camera to obtain an omnidirectional view, even though the image will be distorted. This camera configuration allows greater flexibility when analysing each person's trajectory since they appear in a more significant number of frames.

In conclusion, the final configuration of our cameras is the one in Figure 4, using overhead fisheye cameras next to each train door.

4 Selected Approach

This section proposes a method to count people inside a train based on overhead fisheye cameras and the passenger's trajectory. Figure 9 shows the schema related to the method.

Both detection and tracking modules are the same as in our previous paper [1], specifically designed to track objects in videos recorded by omnidirectional cameras like the ones in this work. It relies on two main steps for each frame: (i) the detection of the object instances and (ii) the matching of detections to their corresponding tracklets. Once the tracking stage is done, the next step is the counting strategy for each door, where we know the passengers' transit over each doorway. The last step is to count the total number of passengers on the train.

4.1 Detection stage

In keeping with the prior work, the detection pipeline is based on the FairMOT object detection work [2]. This paper's solution uses a network structure which consists of two homogeneous branches to detect the objects and obtain re-ID embeddings in a single step. These embeddings are the feature vector of the detected object, which should ideally give us a smaller distance between the detections of the same object and maximise the space to the other things. For this, the detection network has a convolutional layer with 128 kernels that extract the re-ID features, explained with further detail in the FairMOT work [2].

The essential advantage of this network is that because it does not prioritise object detection over re-ID, the features in the output are ideal for both detection and recognition tasks. Although their work is based on DeepSORT [25], the authors present the results.

4.2 Tracking stage

The structure of this stage can be seen in Figure 10. Notably, the processing pipeline is as follows. Periodically, for each time instant t, any detected object in the scene is represented by $\{c_t^i, r_t^i, e_t^i\}$, where c_t^i is the object class, r_t^i represents the bounding box coordinates $\{x_t^i, y_t^i, w_t^i, h_t^i\}$ and e_t^i is the visual appearance representation of the object, known as re-ID embedding. The bounding box coordinates are given in MOT format [26]; (x,y) represent the top left point of the bounding box, and w and h are the width and height.

Then, depending on the detections in the first frame, a set of tracklets is initialised. As new items enter the scene and new tracklets are created, we continue to associate detections with the scene's existing tracklets for consecutive frames. In the association step, both motion (m) and affinity (a) models are used to calculate the pairwise matching scores s_{ij}^m and s_{ij}^a for every detected object i with every object that has previously been tracked j across successive time instances. A final score s_{ij}^g is calculated by a global scoring function that combines s_{ij}^m and s_{ij}^a scores.

As noted earlier, to link detections with active tracklets, we combine a motion model and an affinity model. This post-processing step is based on DeepSORT [25]. We contribute to this phase by having a single network to predict the embeddings necessary to link active tracklets and detections. As we shall demonstrate in Section 5, the features our network learns are more robust than those of the re-ID module of DeepSORT. Our visual appearance transformation network also learns the appearance evolution over time and estimates the embedding in the next frame once we get the embeddings from detections.

To estimate where the future tracklets will be, the motion model combines a constant velocity model and the Kalman Filter [27]. The scores s_{ij}^m are computed by utilising the squared Mahalanobis distance [28] to calculate the proximity between the predicted and detected bounding boxes. Finally, the unlikely matches are ruled out by thresholding the inverse chi distribution to a 95% confidence interval. In a similar process, the scores s_{ij}^a are computed utilising

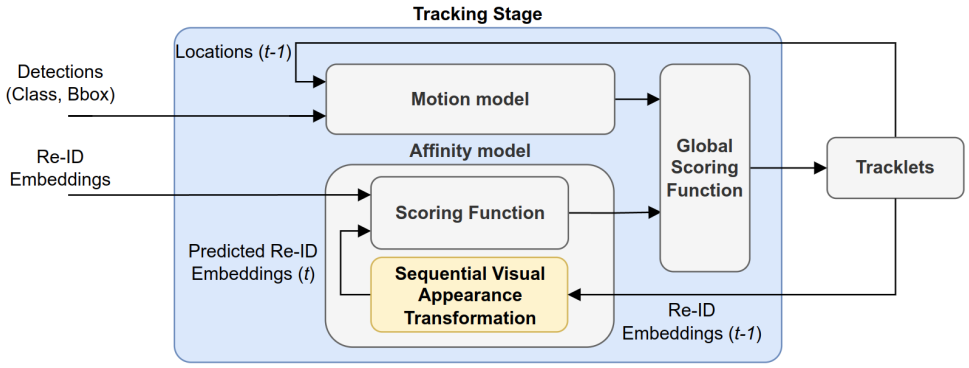


Figure 10: Overview of our online multi-object tracking approach.

the cosine distance to link the visual appearance information between the predicted and detected re-ID embeddings, all of which are based on DeepSORT [25]. The sequential visual appearance transformation network outputs the predicted embeddings online. Finally, using the following equation [25], we combine both distances:

$$s_{ij}^{g} = \lambda s_{ij}^{m} + (1 - \lambda) s_{ij}^{a} \quad (1)$$

where the parameter λ is utilised to balance the impact of the motion and affinity models.

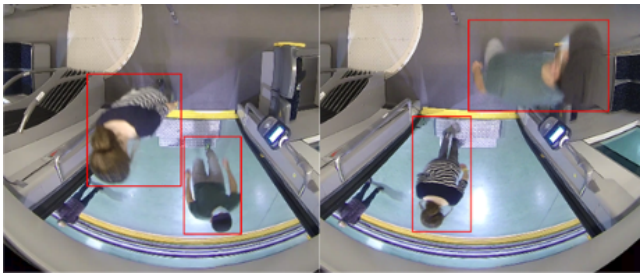
After obtaining the confusion matrix that describes the distances between detections and tracklets, we utilise the Jonker-Volgenant [29] implementation of the linear assignment problem (LAP) algorithm to match pairings that minimise the overall distance. This is a faster implementation of the Hungarian method for LAP. We save a pool of potential active tracks that we keep for n frames when none of the current detections can be matched to them. Although our method is quite similar to DeepSORT [25], we maintain both the new and lost tracks until we get a matching detection for 30 frames. For both new and lost tracks, the number of frames is the same (30). Once the target is detected again, we ought to be able to match it to one of the active tracks visually.

It is crucial to balance the effects of the Kalman and the re-ID embeddings when matching detections with tracks to get accurate results, When examining the image, the Kalman filter can effectively track the targets. Nonetheless, due to the camera's distortion, its constant velocity model causes some problems. The camera is centred on the door, the main area of interest, but that part of the camera is the one that suffers from the highest deformation. This means the target appears to move faster in the central part of the image, so constant velocity can not be assumed. The visual appearance allows us to obtain robustness against occlusions, especially against the high deformation in the door, where we want to know if the passenger is coming into or leaving the train.

4.2.1 *Sequential visual appearance transformation network*

The importance of this model comes from the cameras and environment in which we are developing this work. In the sequence in Figure 11a we can observe the deformation suffered by two people due to the step at the train's entrance. The deformation is more significant when the target is closer to the camera. The height difference created by the step changes the visual appearance of people more than under normal circumstances. We can see in Figure 11b the deformation in a regular fisheye image taken from the HABBOF [30] public dataset.

In both cases, the people in the sequence have taken only two steps and walked roughly the same distance. But if we look at the person with the green shirt on the right side of the image in Figure 11a, they seem to have moved a lot, and their appearance has changed more when compared to the person in the striped sweatshirt in the right side of the image in Figure 11b.



(a) Position variation by walking on the step.



(b) Position variation by walking on a flat floor.

Figure 11: Compare movement distance due to step

This is the main reason why the visual appearance transformation network is beneficial to model better the change in the person's appearance. The change in appearance is so significant that simply using the embeddings is not enough. We need a dedicated model that learns the visual transformation. On the other hand, we already mentioned that the Kalman filter is not robust enough in this situation. However, it's not only because the main area of interest is the

central part of the image. The step between the train and the ground magnifies the "acceleration" effect. Due to the distortion and the step in the door, our main area of interest, the target's velocity in the image is far from constant when getting in or out of the train.

Now, let us explain the architecture of the proposed network, which we maintained from the previous work. We are working with time data; thus, it makes sense to think about using a recurrent neural network (RNN). RNNs have an internal state known as the hidden state that holds the data relevant to what has been seen, and they sequentially process the new input data to maintain track of the temporal information.

Our sequential visual appearance transformation network has an encoder-decoder structure of deep convolutional long short-term memory (convLSTM) architectures [31]. We can see a graphical representation of a single cell in Figure 12.

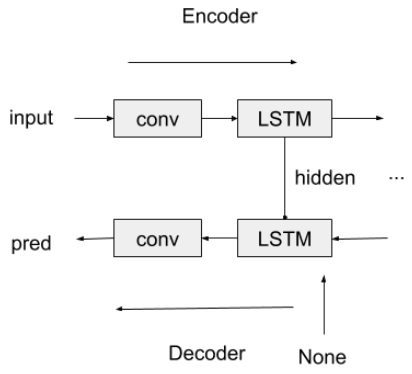


Figure 12: ConvLSTM with an encoder-decoder structure.

As noted, both the encoder and decoder parts are represented by a convLSTM cell. The convLSTM decoder uses the convLSTM encoder's hidden states as its hidden states. The suggested network may be implemented by iteratively concatenating many encoder-decoder structures. We decide to utilise convLSTM since it provides a method for transferring information across sequences. This keeps track of the appearance information for later and stops earlier signals from progressively vanishing. [32]. For this, the encoder captures the context of the visual appearance information, represented by our re-ID embeddings, summarising the previous states of the object trajectory. Conversely, the decoder uses the sequential accumulated transformation, in this instance, the anticipated re-ID embedding for the current frame, to generate the future object appearance.

Using ground truth object sequences, we learn the offline transformation of the visual appearance. We produce a set of re-ID embeddings $e^i_{gt,j}$ for each object identity i that are taken from ground truth detections of the object trajectory. For every $e^i_{gt,j}$, we infer the predicted re-ID embedding $e^i_{p,j}$ by using the proposed network. Finally, we compute the affinity loss as follows:

$$L_{\text{affinity}} = 1 - |\frac{1}{N}\sum_{i=1}^{N} D_c(e^i_{p,j}, e^i_{gt,j})| \quad (2)$$

where D_c represents the cosine distance and N is the batch size. We split the ground truth trajectories into sub-trajectories with a single length j because the length of the ground truth trajectories is unequal. As a result, we can do the training in batch mode.

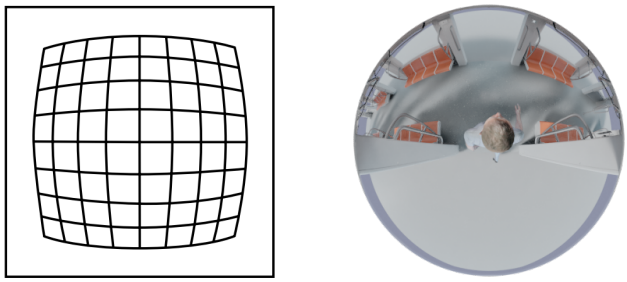
4.3 Counting strategy for flow control

Once explained how people are tracked during training, the following section describes how counting people is done. For this purpose, two main factors are taken into account. Firstly, the position and effect of the fisheye camera and how this affects the appearance of people in the image and, therefore, the selection of an appropriate reference point for a person. And, secondly, the logic behind the people counting module.

4.3.1 Fisheye camera effect

Fisheye cameras generate large deformations in objects as they move through space, as explained in section 4.2.1. They specifically produce barrel distortion (Figure 13a), which has more effect at the wide-angle end of the range of the image. The deformation may cause difficulties in developing a good people-counting strategy. So selecting a suitable reference point for a person to identify when he has entered or left a train door is key to obtaining a high hit rate.

Just in the centre of the image of a fisheye camera, an object's appearance will be the least deformed. This can be seen in Figure 13a: the squares located in the middle of the image have no distortion, while the extremes of the image suffer a great distortion. In a real case of a person crossing the door line, only their head and shoulders will be visible in the centre of an image, with no distortion. (Figure 13b).



(a) Barrel distortion [33]

(b) Overhead fisheye image with the person above the camera, and camera in line with the door.

Figure 13: Fisheye camera distortion

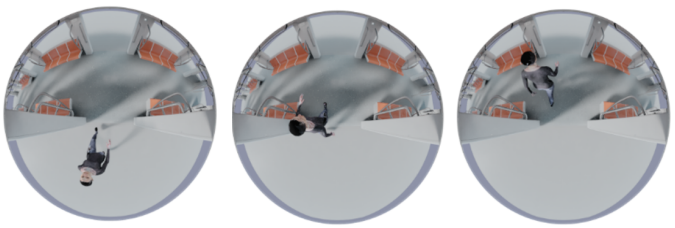


Figure 14: Simulation of an overhead fisheye camera in line with the door. *Left: Person approaching the door. Centre: Person below the camera and above the door line. Right: Person walking away from the door.*

When an object is further from the centre, the object's image suffers more deformation. For a person, once they are walking away from the camera's centre point, their entire body will appear in the image as long as there are no occlusions. An example can be seen in the left and right images from Figure 14. So the distortion is related to the object's position or person respective to the camera.

Ideally, the best place to locate the camera and detect peoples' positions would be just above the limit line; in this case, the train door axis, as shown in Figure 14. Using the central point of the bounding box would work perfectly as a reference point for the person's position. However, the camera is placed inside the train because of design and structural reasons. Therefore, placing the fisheye camera outside the door axis results in a person's appearance distortion. This makes it somewhat more complicated to define its reference point and to know whether a person is inside or outside the carriage in borderline situations, as in Figure 15. Moreover, this distortion is even increased by some trains' steps that may have at the doors.

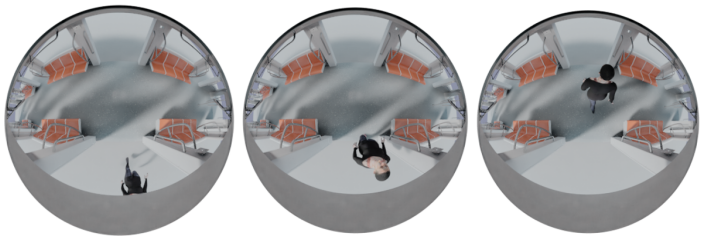


Figure 15: Simulation of an overhead fisheye camera inside the train. *Left: Person approaching the door. Centre: Person just above the door line. Right: Person walking away from the door.*

The position variation of an object with a constant height moving towards or away from the centre of an overhead fisheye camera is different for the highest and lowest part of the object. This means that when an overhead camera captures the movement of a person, the head's location suffers more variation than the feet' location. For example, in Figure 16 the variation of the head doubles the variation of the feet of the person.

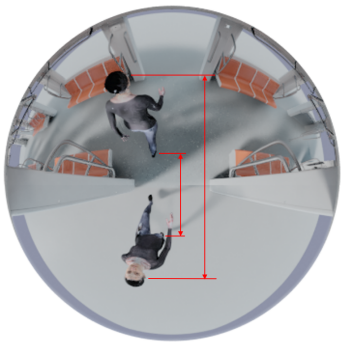
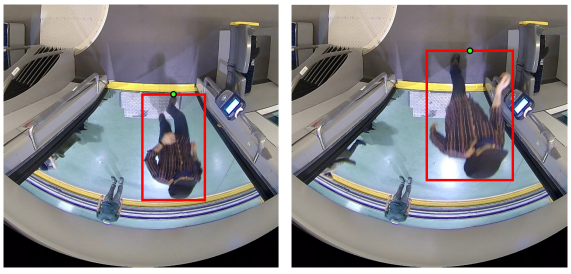


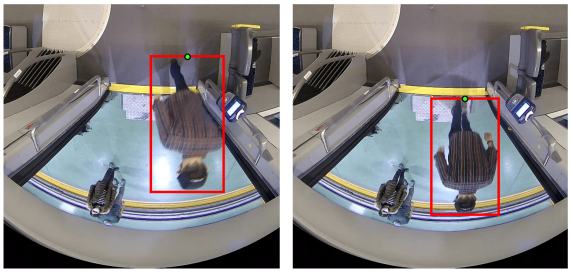
Figure 16: Deformation of top and bottom parts of an object moving in a fisheye camera. In this case, the person's head suffers more variation than the feet.

Usually, with standard CCTV cameras, it is common to use the bounding box centre or the person's head as a reference point, as the head is the most visible part. However, based on the problems stated before, with this camera configuration, the most suitable reference for defining the position of a person is their feet. As for the bounding box, the selected reference point is the upper-centred point related to the feet. Figure 17 shows the entering and exiting conditions picking the feet, located in the upper-centred point of the bounding box, as a reference point to consider that a person has crossed the door line axis.

While entering the train door (Figure 17a) the person is first entirely outside the train. The right image shows that most of the body appears outside the train, but the feet are inside. That means that the person has entered the train. This happens due to the previously explained distortion that depends on the height of an object. While exiting the train (Figure 17b) the left image shows that most of the body is outside the train, but is not considered that the person has left the train until the feet are out of it. This condition is fulfilled in the right image, where the feet are finally outside the train door limit.



(a) Entering condition



(b) Exiting condition

Figure 17: Entering and exiting conditions of a person through a train door (The train door limit is the yellow line).

This way, if a person is standing right at the door's limit, even if their body's bounding box is mostly outside the train, the person will be considered inside the train, as in Figure 18.

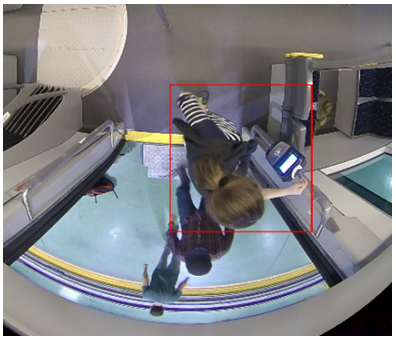


Figure 18: Example of a person standing right under the door.

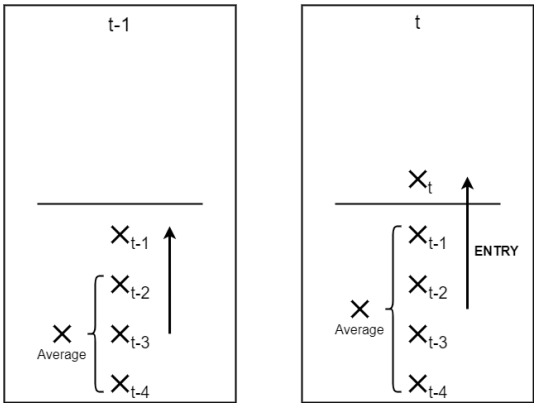
4.3.2 *Counting logic*

The counting logic is based on the state-of-the-art of [34]. The main idea is to determine a line and count how many people have crossed it in each direction. The line is defined as a horizontal line in the centre of the image.

To determine if an object has crossed a limit line, the author of the paper mentioned above uses the vertical (y-axis) coordinate of the centre of the bounding box of a person in each frame. Then, the current position and the mean value of the previous positions of the tracked person are considered to evaluate the crossing. As it states, the mean of all the previous points is selected because:

> "The reason we take the mean is to ensure our direction tracking is more stable. If we stored just the previous centroid location for the person, we leave ourselves open to the possibility of false direction counting. [...] by taking the mean, we can make our people counter more accurate."

In the current case of this paper, the limit line will be defined as the limit of the door (notice the yellow line in the Figures 17). So it will also be a horizontal line in the image. A person will be considered that has crossed the line if the limit line's vertical (y-axis) value is between the current position and the mean of the previous ones. A representation of the stated logic can be seen in Figure 19.



(a) Entering condition logic

(b) Exiting condition logic

Figure 19: Entering and exiting logic of a person crossing the line.

- **Enter condition logic** (Figure 19a): A person has entered the train if:
  - The current reference point is above the limit line.
  - The average of the previous points is below the line.

  (The person is moving to the upper part of the image, representing the inside of the train.)
- **Exit condition logic** (Figure 19b): A person has exited the train if:
  - The current reference point is below the limit line.
  - The average of the previous points is above the line.

  (The person is moving to the lower part of the image, representing the outside of the train.)

Some changes were made to the approach mentioned above due to some requisites specific to using the fisheye camera. Instead of using the bounding box centre, the reference point is the top middle point. This point coincides with the location of the feet in the fisheye images, as explained in Section 4.3.1; this reference point is more suitable to determine if a person is inside or outside the train.

The objective of the counting logic is to analyse the flow through the door. If a person goes in or out of the train multiple times in the same recording, the system must count every time the person crosses the line limit (see Figure 20). This may carry problems related to the counting logic, especially with the average of the previous reference points. If the person goes through the door several times, the average point will be unsuitable for the stated logic. That is why the previous tracking points are reset after a person crosses the line limit to avoid further problems. This way, the average position is kept on one side of the limit line, and if the person crosses again, even with the same tracking ID, the logic will still work.

The paper proposes a solution to reset the trajectories based on separating the whole trajectory into sub-trajectories. Every time a tracked person crosses the limit line in either direction, a sub-trajectory containing the previous points of the tracking is saved and set aside. Then, a new sub-trajectory is started, containing only the first reference point of the moment of entering or exiting the door. Figure 21 shows how the trajectory of Figure 20 is separated into four sub-trajectories.

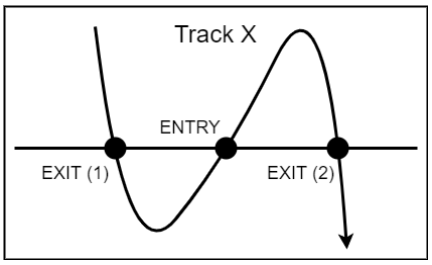


Figure 20: Example of the trajectory of a person entering and exiting multiple times.

This way, the average of the previous points is kept on one side of the line, and as soon as a person crosses the line, the module will count as the entry or exit and reset the previous points for the next sub-trajectory. For example, for the first exit in Figure 21 (Sub-track X_1) the average of the previous points is on the upper side of the

line. Once the person crosses the line (green circle), another sub-trajectory is started (Sub-track X_2). This sub-track ends once the person crosses the line and enters (orange circle). The average point in red is kept on the downside of the line. Similarly happens with the next sub-trajectory (Sub-track X_3), but the average is kept on the upper side. The trajectory ends when the person is no longer tracked (Sub-track X_4).

Finally, the flow of a door is composed of two positive integer numbers:

- **Enter flow**: each person that enters the train counts as a **positive enter**.
- **Exit flow**: each person that exits the train counts as a **positive exit**.

Separating the enter and exit values gives more information about the flow of a door. If a unique number was given as a flow, the information of people who entered or exited the door would be lost. For example, it would be the same for the cases where two people enter and two people exit and for no one entering and exiting the door (flow equal to zero in both cases).

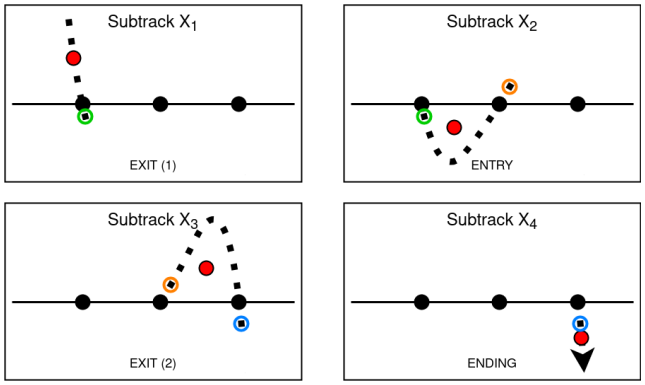


Figure 21: Considered sub-trajectories to evaluate if a person has crossed or not the line in multiple crossings within the same recording. The red dot indicates the average point of the sub-trajectory. The dotted line indicates each person's step. The coloured circle over a dot indicates the end/beginning of a sub-track.

4.4 Occupancy computation

The methodology for calculating the occupancy of a train is formed by the sum of the people entering minus the sum of the people leaving the train at each door; see Equation 3. It's important to notice that a particular door's flow can have either positive, zero or negative value. But the overall train occupancy will be either zero or positive.

$$Total_flow_{door} = Enter_flow_{door} - Exit_flow_{door} \quad (3)$$

Once the total flow of each door is known, to calculate the overall train occupancy, all door flow values are added together, as shown in 4. This particular value will indicate the number of passengers on the train at a given time.

$$Total_occupancy_{train} = \sum_{d=0}^{N_{door}} Total_flow(d) \quad (4)$$

$flow(d)$ is the flow door value, and d is the door number.

5 Results

5.1 Datasets

The next section presents the datasets used for the training and testing of the detection and tracking models and the use case of people counting through trajectory analysis. The datasets are separated into two sections: Public datasets (Section 5.1.1), which were used to train the detection and tracking models; and Proof of concept dataset (Section 5.1.2), which consists of video recordings of a real train fisheye camera where people enter and exit a train.

5.1.1 Public datasets (Detection and Tracking oriented)

The performance of our method is assessed over a certain amount of data obtained from omnidirectional cameras. We considered seven publicly available datasets specifically created to address the problem of people tracking and detection. They combine multiple indoor environments, including a wide range of challenging scenarios: crowded room, severe body occlusions, various body poses, head camouflage (e.g., hoods, hats) and low-light conditions.

Table 1 provides a summary of those datasets, including the number of frames, the number of people (IDs), and the data source for each of them.

It should be noted that the HABBOF, FES, and PIROPO databases lack tracking information.

We added some data captured in trains with the specific camera setting we proposed for these datasets.

Table 1: Public omnidirectional datasets.

Dataset	Images	IDs	Data Source
CEPDOF [35]	25.5k	51	Real
HABBOF [30]	5.8k	-	Real
FES [36]	301	-	Real
Bomni [37]	12.9k	85	Real
THEODORE [36]	100k	2307	Synthetic
Mirror World [38]	7k	63	Real
PIROPO [39]	3k	-	Real
PATHTRACK [40]	276.4k	16,287	Real
Total	430.9k	18,793	-

5.1.2 Proof of concept dataset (People counting)

The dataset for obtaining the results of counting people consists of several video clips. The recordings were captured by a fisheye camera located at the door of a train coach. The videos contain different cases of people entering and exiting the door:

- Cases of one person to multiple people entering and exiting the door one by one.
- People entering or exiting together in groups on the train.
- Video recordings with two or more people, usually crossing their paths.

- Cases where people enter and exit the train at varying walking speeds (Fastly and slowly moving passengers).
- Other actions such as stopping in the middle of the door.

All the clips were classified into three categories depending on the difficulty of the video stream:[1]

- **"Low"**: In these videos, there is usually one person who enters and/or exits the train door.
- **"Medium"**: There are usually 2 or 3 people who enter and/or exit the train.
- **"High"**: There are more than five people who enter and/or exit the train.

Once all the clips are grouped in the three types of difficulty, the dataset remains as in Table 2.

Table 2: Number of clips separated into difficulties.

Difficulty	No. of clips
Low	41
Medium	39
High	22

The ground truth data of the different clips consists of the tally of how many people enter and exit the train door, i.e., the total count of people crossing the limit line (door flow). Alongside the difficulty level and the distinction between people entering and exiting the train door, the ground truth data is composed as shown in Table 3.

Table 3: Ground truth data of the count of people crossing the train door line. The difference between entry and exit and the three difficulty levels of clips are reflected.

	People crossing count (GT)		
	Low	**Medium**	**High**
Entry	27	47	68
Exit	23	49	68
Total	50	96	136

In total, the dataset contains **282** door flows or entries and exits, which are divided into "Low", "Medium", and "High" difficulties.

5.2 *Experimental evaluation*

Our previous work explained how we obtain the best tracker configuration to achieve a state-of-the-art tracker [1].

Although our main contribution to the tracking stage was proposed in the association part, we also trained the FairMOT object detector network [2] following the same implementation details as the authors. The original network is trained on frontal view images, so its performance on omnidirectional data drops dramatically. Consequently, we have used all omnidirectional datasets to fine-tune the pre-trained network.

We select trajectories with a sequence length of $j = 40$ frames for our sequential visual appearance transformation network This value was chosen because in a camera recording at 15-20 fps, 40 frames are approximately 2 seconds of the recording, and a person will approximately take that time to enter a train; thus, we have 40 consecutive re-ID embeddings for each person's identity.

The network can keep the identification of the tracked individual for as many frames as they are in the scene, although it was designed to function with as little as 40 frames in training.

The tracklet is retained in the pool of potential identities for 30 frames if the system loses track of a person. If, after 30 frames, the system cannot match the tracklet with any further detections, we assume the individual has departed the scene. This part is important for evaluating the public datasets. However, as the goal is to know whether a person has walked through a door and the direction they walked in, keeping track of them for longer while they are inside the train is not essential.

We have used the same configuration as in our previous work [1] where we prove that using a high λ value and, therefore, giving more weight to the score of the Kalman filter results in a significantly lower outcome. This shows the importance of using appearance-based matching, adding robustness against occlusions. The best option is to use a low λ value but different from zero, so we use $\lambda = 0.1$.

The system's ability to re-ID people comes from visual information. In case of occlusion, if the target moves through the image, a motion-only model would discard the detection and start a new tracklet. That means we would be unable to count the action if the occlusion happens during an entry or exit if it wasn't for the visual model.

Following the standard practices in multi-object tracking, we use the multi-object tracking accuracy (MOTA) [41] and the ratio of correctly identified detections over the average number of ground-truth and computed detections (IDF1) for rigorously evaluating re-ID features with ground-truth detections.

To evaluate the performance of our tracker, we have considered a unique scenario containing the test set from all datasets that include ground truth tracking information, meaning: CEPDOF, MWR, and Bomni, with six different sequences in total. Also, to achieve a better performance, we train our tracker with the PATHTRACK dataset [40]. This dataset is one of the largest publicly available multiple objects tracking data sets.

Following the same implementation details as in our previous work [1], we obtain three proposed networks. Table 4 shows the MOTA and IDF1 results for the baseline FairMOT method and our three proposed networks. The convLSMT-Enc-Dec-3 achieves the best MOTA and IDF1 results with 88.21% and 87.77%, respectively.

[1] *Even though the video difficulties are separated into three subsets, the "Low" and "Medium" video complexities do not differ too much. The major dissimilarity is with the "High" difficulty videos.*

Table 4: Evaluation of our online tracking by detection approach including PATH-TRACK dataset [40]

Model	MOTA (%)	IDF1 (%)
FairMOT	82.94	80.25
FairMOT + convLSMT-Enc-Dec-1	87.85	87.73
FairMOT + convLSMT-Enc-Dec-3	**88.21**	**87.77**
FairMOT + convLSMT-Enc-Dec-5	87.15	87.55

We compare the performance of our approach on the Bomni [37] and Mirror World [38] datasets as we only found results for these two datasets. Table 5 demonstrates the benefits of our affinity model concerning state-of-the-art methods. As state-of-the-art methods do, we also include multi-object tracking precision (MOTP) for this comparison. Results show that our approach outperforms all existing methods considering the omnidirectional perspective. In particular, it improves the best performance (93.5%) obtained by BTLD [42] by almost 2% using the Bomni dataset. Similarly, we outperform the best performance (38.4%) obtained so far by SORT [38] by more than 40% using the Mirror World dataset.

Due to the lack of recent results for omnidirectional datasets, we decided to compare ourselves with DeepSORT [25]. We use the detections of our model and perform the tracking with their approach. Table 5 and 6 show results across the six test sequences. We used the model with $\lambda = 0$ for comparison, as that is the value they used in their final implementation. Table 5 shows that using FairMOT instead of DeepSORT yields slightly better results, proving the re-ID embeddings of their work are better than those of the original DeepSORT. Table 6 shows that our method improves tracking given the increase in the number of mostly and partially tracked people.

Table 5: Comparison with state-of-the-art methods.

Approach	MOTA (%)	MOTP (%)	IDF1 (%)
Bomni Dataset			
FTMO [37]	73.52	72.00	-
FTMO (updated) [43]	86.27	72	-
RTMOT [42]	78.55	76.74	-
BTLD [44]	93.5	-	-
Ours	**94.27**	**92.14**	**95.14**
Mirror World Dataset			
SORT [38]	38.4	-	32.1
Ours	**84.14**	**81.93**	**88.68**
Across all datasets (MW, Bomni and CEPDOF)			
DeepSORT[25]	81.37	80.15	79.98
FairMOT[2]	82.94	80.34	80.25
Ours	**88.03**	**84.25**	**86.87**

Finally, Figure 22 shows that the presented method is more robust to the appearance distortion caused by overhead cameras than DeepSORT. Moreover, it can recognise the target in the centre of the image, whereas DeepSORT changes its ID due to appearance distortion.

Taking all of these results into account, we prove that adding the sequential visual appearance transformation network helps with the reidentification task in the case of fisheye cameras. Due to the deformation, the appearance of the objects is less constant than with regular cameras, so we need the aid of a model specialised in predicting the object's appearance and comparing the embeddings. We can't use the embeddings directly because the appearance changes between frames.

Table 6: Tracking quality compared with DeepSORT.

Model	Mostly tracked	Partially tracked	Totally lost
DeepSORT	16	13	6
FairMOT + convLSMT-Enc-Dec-3	**20**	**11**	**4**

5.3 *Global results*

The proof of concept dataset defined in Section 5.1.2 is used to obtain the results of the people counting method. This dataset is exclusively used to get the results of the people counting module because it is the closest dataset to a real-case scenario, the main objective is to check the feasibility of the people counting logic from Section 4.3.2. The dataset is not considered significant enough to present the results of the detection and tracking module due to its size.

The accuracy results for each difficulty level were based on the total count of entries and exits from each video divided by its ground truth (GT), as shown in Equation 5. The following Table 7 shows the accuracy results of the people counting logic.

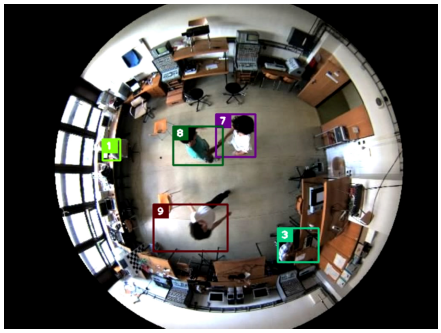
$$accuracy_{Entry/Exit}(\%) = \frac{\sum Output}{\sum GT} \tag{5}$$

Table 7: People counting accuracy (%).

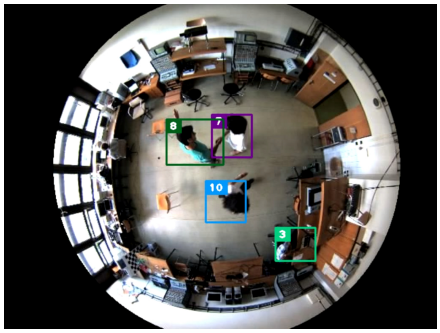
	People counting accuracy (%)		
	Low	Medium	High
Entry	100.00	97.87	91.18
Exit	100.00	89.80	79.41
Mean	100.00	93.75	85.29

The overall accuracy of counting people getting on and off is **90.78%**. The worst results are presented with "High" difficulty clips, as the detector may fail in crowded situations or when people cross their paths. Even though the dataset used to obtain these results is based on real case scenarios (recorded with fisheye cameras and with people entering and exiting the train door), the number of door flows is not representative enough for a generalization. So the results might not be reliable. Future work aims to create a more extensive dataset with more realistic scenarios to validate the people-counting module alongside the detection and tracking models.

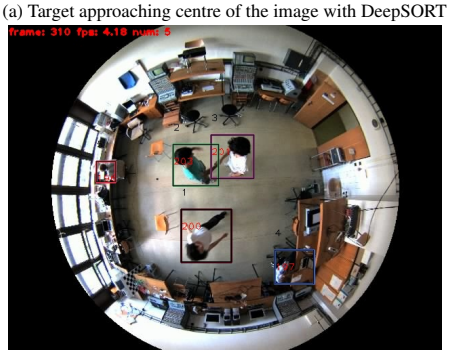
The approach proposed in [9] was similar to this project's strategy. As the author uses a tracker and a counting line to know how



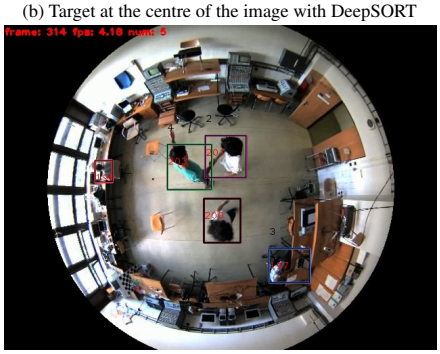
(a) Target approaching centre of the image with DeepSORT



(b) Target at the centre of the image with DeepSORT



(c) Target approaching centre of the image with our method



(d) Target at the centre of the image with our method

Figure 22: One of the scenarios where our method can keep track of a target compared to state-of-the-art methods.

many people entered or left the train. In this case, it uses a standard overhead camera, so there is no image distortion. This type of camera needs a higher position to have enough vision to analyze the passengers' flow, so the point of view (camera location) is outside the train, on the roof of a train platform. The accuracy metric used in this proposal is based on a confusion matrix. The ground truth data of the videos specifies when a person has entered or exited the train. The proposal achieves an accuracy of 92.01% and 92.47% in counting people leaving and entering the train, respectively.

The approach in this paper instead is based on an overhead fisheye camera inside the train. A fisheye camera can be located in a lower position maintaining a wide viewing angle, and, therefore, inside the train, close to the door. This location's advantage is that the cameras can follow the passengers' trajectory inside and outside the train at the expense of having distortion. But, as explained in section 4.3.1, the tracker solves the fisheye camera's image distortion problem. The accuracy metric used in this paper is based on counting entries and exits, achieving an overall accuracy of 90.78% against the proof of concept dataset.

Comparing the results and affirming which method is better does not make sense since the cameras and the databases are different. The only viable comparison is to evaluate the accuracy of the two solutions and check that the results are similar. It is impossible to affirm that one methodology is better than the other, as their goal is only to count people, but we also want to analyze the trajectory and the influx of people using the train.

6 Conclusions

This paper presents a method based on video sequences to count the number of people inside a train car based on the doors' flow. The technique can be applied in environments with narrow spaces and a limited height to locate the camera, such as buildings, transports or events where a door limits the entry. The proposed approach is based on an appearance-based multi-object tracker and a door's flow counting method that analyze the trajectory of each passenger, counting how many times the people enter or leave the coach.

A study of the best camera type and its location was done to obtain the best possible vision of the passengers at the door. After analyzing the train's environment and carrying out different simulations, the best results were given by an overhead fisheye camera. This camera causes distortion, which our appearance model can handle, improving the performance achieved by the motion modelling-only models. Moreover, the selected trackers can also handle the step acceleration problem.

Once the trajectory was known, a flow control module based on a counting line was developed to calculate the number of times the passengers crossed the line in each direction. Different options for the reference point of a bounding box were regarded, and the central upper point was selected as the best reference point to decide whether a person was on either side of the line. Finally, the overall train occupancy module knows the train occupancy based on all the door's flow information.

After all this work, it is possible to affirm that using fisheye cameras allows for better tracking of people thanks to their broader field of view. The selected tracker can solve the distortion problems caused by the fisheye camera. Moreover, the counting method is robust against the different situations found on a train, and it is

successfully tested on a proof-of-concept dataset. Finally, it is confirmed that the previous tracker [1] works as well as expected in a different and more complex environment.

7 Future work

It has been proven that the method works, but its accuracy cannot be confirmed because the dataset used is small and in a controlled environment. The future line of work is to obtain a dataset taken in a real environment, annotate it and check its absolute precision. As for the algorithm, the next step would be to try a Generative adversarial network (GAN) [45] to model the appearance of the detections, as this model has proven to be a very effective way of predicting visual appearance in a time sequence.

8 Acknowledgments

This work has been partially funded by the European Union's H2020 research and innovation programme under grant agreement 101021981, project APPRAISE (Facilitating public & private security Operators to mitigate terrorism scenarios against soft targets).

References

[1] I. Sagastiberri, N. v. d. Gevel, J. García, O. Otaegui, "Learning Sequential Visual Appearance Transformation for Online Multi-Object Tracking," in 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1–7, 2021, doi:10.1109/AVSS52988.2021.9663809.

[2] Y. Zhang, C. Wang, X. Wang, W. Zeng, W. Liu, "FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking," arXiv preprint arXiv:2004.01888, 2020.

[3] E. Khoumeri, H. Fraoucene, E. Khoumeri, C. Hamouda, R. Cheggou, People Counter with Area Occupancy Control for Covid-19, 405–415, 2021, doi: 10.1007/978-3-030-63846-7_38.

[4] A. Naser, A. Lotfi, J. Zhong, "Adaptive Thermal Sensor Array Placement for Human Segmentation and Occupancy Estimation," IEEE Sensors Journal, **21**(2), 1993–2002, 2021, doi:10.1109/JSEN.2020.3020401.

[5] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, A. Hauptmann, "Learning Spatial Awareness to Improve Crowd Counting," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 6151–6160, 2019, doi: 10.1109/ICCV.2019.00625.

[6] Y. Li, X. Zhang, D. Chen, "CSRNet: Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes," 1091–1100, 2018, doi:10.1109/CVPR.2018.00120.

[7] Y. Miao, Z. Lin, G. Ding, J. Han, "Shallow Feature Based Dense Attention Network for Crowd Counting," in The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, 11765–11772, AAAI Press, 2020.

[8] Q. Song, C. Wang, Z. Jiang, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, Y. Wu, "Rethinking Counting and Localization in Crowds:A Purely Point-Based Framework," 2021.

[9] S. A. Velastin, R. Fernández, J. E. Espinosa, A. Bay, "Detecting, Tracking and Counting People Getting On/Off a Metropolitan Train Using a Standard Video Camera," Sensors, **20**(21), 2020, doi:10.3390/s20216251.

[10] D. Kuplyakov, Y. Geraskin, T. Mamedov, A. Konushin, "A Distributed Tracking Algorithm for Counting People in Video by Head Detection," paper26–1, 2020, doi:10.51130/graphicon-2020-2-3-26.

[11] J.-W. Kim, K.-S. Park, B.-D. Park, S.-J. Ko, "Real-time vision-based people counting system for the security door," in Proceedings of the IEEK Conference, 1416–1419, The Institute of Electronics and Information Engineers, 2002.

[12] J. Barandiaran, B. Murguia, F. Boto, "Real-time people counting using multiple lines," in 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services, 159–162, IEEE, 2008.

[13] J. Ahmad, H. Larijani, R. Emmanuel, M. Mannion, A. Javed, "An intelligent real-time occupancy monitoring system using single overhead camera," in Proceedings of SAI Intelligent Systems Conference, 957–969, Springer, 2018.

[14] S. Yu, X. Chen, W. Sun, D. Xie, "A robust method for detecting and counting people," in 2008 International conference on audio, language and image processing, 1545–1549, IEEE, 2008.

[15] P. Chato, D. J. M. Chipantasi, N. Velasco, S. Rea, V. Hallo, P. Constante, "Image processing and artificial neural network for counting people inside public transport," in 2018 IEEE Third Ecuador Technical Chapters Meeting (ETCM), 1–5, IEEE, 2018.

[16] M. Bouazizi, C. Ye, T. Ohtsuki, "Low-Resolution Infrared Array Sensor for Counting and Localizing People Indoors: When Low End Technology Meets Cutting Edge Deep Learning Techniques," Information, **13**(3), 2022, doi:10.3390/info13030132.

[17] R. L. dos Santos, H. C. de Oliveira, M. C. de Almeida, D. F. Vieira, E. P. L. Junior, T. Ji, "A Low-Cost Bidirectional People Counter Device for Assisting Social Distancing Monitoring for COVID-19," Journal of Control, Automation and Electrical Systems, **33**(4), 1148, 1160, 2022, doi: 10.1007/s40313-022-00916-z.

[18] P. Bergmann, T. Meinhardt, L. Leal-Taixe, "Tracking Without Bells and Whistles," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[19] W. Tian, M. Lauer, L. Chen, "Online Multi-Object Tracking Using Joint Domain Information in Traffic Scenarios," IEEE Transactions on Intelligent Transportation Systems, **21**(1), 374–384, 2020, doi:10.1109/TITS.2019.2892413.

[20] W. V. Ranst, F. De Smedt, J. Berte, T. Goedemé, "Fast Simultaneous People Detection and Re-identification in a Single Shot Network," in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, doi:10.1109/AVSS.2018.8639489.

[21] L. Ren, J. Lu, Z. Wang, Q. Tian, J. Zhou, "Collaborative Deep Reinforcement Learning for Multi-object Tracking," in V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss, editors, Computer Vision – ECCV 2018, 2018.

[22] X. Dong, J. Shen, "Triplet Loss in Siamese Network for Object Tracking," in Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[23] J. Son, M. Baek, M. Cho, B. Han, "Multi-object Tracking with Quadruplet Convolutional Neural Networks," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3786–3795, 2017, doi:10.1109/CVPR.2017.403.

[24] J. Yin, W. Wang, Q. Meng, R. Yang, J. Shen, "A Unified Object Motion and Affinity Model for Online Multi-Object Tracking," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[25] N. Wojke, A. Bewley, D. Paulus, "Simple Online and Realtime Tracking with a Deep Association Metric," 2017.

[26] P. Dendorfer, H. Rezatofighi, A. Milan, J. Shi, D. Cremers, I. D. Reid, S. Roth, K. Schindler, L. Leal-Taixé, "MOT20: A benchmark for multi object tracking in crowded scenes," CoRR, **abs/2003.09003**, 2020.

[27] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," Transactions of the ASME–Journal of Basic Engineering, **82**(Series D), 35–45, 1960.

[28] P. C. Mahalanobis, "On the Generalised Distance in Statistics," Proceedings of the National Institute of Sciences of India, **2**(1), 49—55, 1936.

[29] R. Jonker, A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," Computing, **38**(4), 325–340, 1987.

[30] S. Li, M. Tezcan, P. Ishwar, J. Konrad, "Supervised People Counting Using An Overhead Fisheye Camera," in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1–8, 2019, doi: 10.1109/AVSS.2019.8909877.

[31] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. kin Wong, W. chun Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," 2015.

[32] S. A. Rahman, D. A. Adjeroh, "Deep Learning using Convolutional LSTM estimates Biological Age from Physical Activity," Scientific Reports, 2019, doi:10.1038/s41598-019-46850-0.

[33] G. Vass, "Applying and removing lens distortion in post production," 2003.

[34] A. Rosebrock, "OpenCV People Counter," https://pyimagesearch.com/2018/08/13/opencv-people-counter/, 2018, [Online; accessed 12- Aug- 2022].

[35] Z. Duan, M. O. Tezcan, H. Nakamura, P. Ishwar, J. Konrad, "RAPiD: Rotation-Aware People Detection in Overhead Fisheye Images," 2020.

[36] T. Scheck, R. Seidel, G. Hirtz, "Learning from THEODORE: A Synthetic Omnidirectional Top-View Indoor Dataset for Deep Transfer Learning," 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, doi:10.1109/wacv45572.2020.9093563.

[37] B. E. Demiroz, I. Ari, O. Eroglu, A. A. Salah, L. Akarun, "Feature-based tracking on a multi-omnidirectional camera dataset," in 2012 5th International Symposium on Communications, Control and Signal Processing, 1–5, 2012, doi:10.1109/ISCCSP.2012.6217867.

[38] R. B. Knapp, N. F. Polys, J.-B. Huang, A. Ibrahim, N. Ma, C. Hurt, Y. xiao, "MW-18Mar Dataset," .

[39] U. P. d. M. G.-U. Grupo de Tratamiento de Imágenes, "PIROPO Database: People in Indoor ROoms with Perspective and Omnidirectional cameras," .

[40] S. Manen, M. Gygli, D. Dai, L. V. Gool, "PathTrack: Fast Trajectory Annotation with Path Supervision," 2017.

[41] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa, P. Soundararajan, "The CLEAR 2006 Evaluation," in R. Stiefelhagen, J. Garofolo, editors, Multimodal Technologies for Perception of Humans, 1–44, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[42] H. S. P. Brauer, Camera based Human Localization and Recognition in Smart Environments, Ph.D. thesis, University of the West of Scotland, 2014.

[43] B. Demiroz, A. Salali, L. Akarun, "Multiple person tracking using omnidirectional cameras," 1231–1234, 2014, doi:10.1109/SIU.2014.6830458.

[44] G. Gemignani, BTLD+:A Bayesian Approach to Tracking Learning Detection by Parts, Ph.D. thesis, UNIVERSIT'A DEGLI STUDI DI MILANO, 2013.

[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial nets," in Advances in neural information processing systems, 2672–2680, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 4, 114-120 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Estimating a Minimum Embedding Dimension by False Nearest Neighbors Method without an Arbitrary Threshold

Kohki Nakane[1,*], Akihiro Sugiura[2], Hiroki Takada[1]

[1]*Graduate School of Engineering, University of Fukui, Fukui City, 910-8507, Japan*

[2]*Department of Radiological Technology, Gifu University of Medical Science, Seki City, 501-3892, Japan*

ARTICLE INFO

Article history:
Received: 30 May, 2022
Accepted: 10 August, 2022
Online: 24 August, 2022

Keywords:
False Nearest Neighbors
Attractor
Chaos
Van der Pol Equation
Lorenz Equation
Electrooculogram

ABSTRACT

The false nearest neighbors (FNN) method estimates the variables of a system by sequentially embedding a time series into a higher-dimensional delay coordinate system and finding an embedding dimension in which the neighborhood of the delay coordinate vector in the lower dimension does not extend into the higher, that is, a dimension in which no false neighbors or neighborhoods exist. However, the FNN method requires an arbitrary threshold value to distinguish false neighborhoods, which must be considered each time for each time series to be analyzed. In this study, we propose a robust method to estimate the minimum embedding dimension, which eliminates the arbitrariness of threshold selection. We applied the proposed approach to the van der Pol and Lorenz equations as representative examples of chaotic time series. The results verified the accuracy of the proposed variable estimation method, which showed a lower error rate compared to the minimum dimension estimates for most of the thresholding intervals set by the FNN method.

1. Introduction

In our previous studies, we have studied the development of mathematical models and feature extraction using artificial intelligence for biological signals. We simulated biological signals using GAN [1,2], a deep-learning generative model. Then, from the trained generative model, we have statistically analyzed a huge number of optimized parameters and investigated the possibility of extracting new features from the biological signals that have not been discovered so far. Then, since deep learning handles a large number of parameters, the generated model can be said to be a multivariable-dependent system. On the other hand, it has been pointed out that the biological signals to be generated are not multivariable in nature, but may be relatively simple systems that depend on a few variables at most [3,4]. To investigate the variables on which the system depends, several methods have been proposed to estimate the minimum embedding dimension of the time series using attractors.

The false nearest neighbor (FNN) method [5–7] estimates the variables of time series obtained from dynamical systems. The concept of false neighbors may be understood more clearly by reference to the Lorenz equation [8], as shown in Figures 1-3. A and B in the 1D attractor are no longer close to each other with increasing dimensionality of the attractor [9]. In the FNN method, the minimum embedding dimension is obtained using the following four steps. (i) Sequentially embed the time series into a higher-dimensional delay coordinate system. (ii) Compare the neighborhood distance of the lower-dimensional delay coordinate vector with that of the next higher-dimensional delay coordinate vector. (iii) If the rate of change of the neighborhood distance of the delay coordinate vector with increasing dimensionality of the delay coordinate system exceeds a given threshold value, the vector is considered a false neighborhood. (iv) The minimum embedding dimension is that in which all delay coordinate vectors are not false neighbors for the first time with increasing embedding dimensions.

The primary advantage of this method is that it uses all samples and thus is not an estimator, and is computationally fast owing to the simplicity of the necessary calculations. However, the FNN method requires an arbitrary threshold value to determine whether

*Corresponding Author: Kohki Nakane, kohki.pc@gmail.com

https://dx.doi.org/10.25046/aj070415

neighborhoods are false. Although observed time series are typically mixed with noise, the FNN method requires the selection of a threshold value depending on the level of noise, even for a single series. That is, the value of this threshold must be considered for each time series to be analyzed. In this study, we propose a robust minimum embedding dimension estimation method that eliminates this arbitrariness in threshold selection.

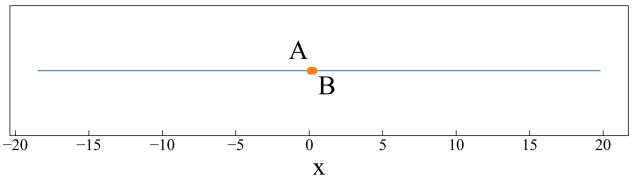


Figure 1: 1D attractor for the Lorenz equation

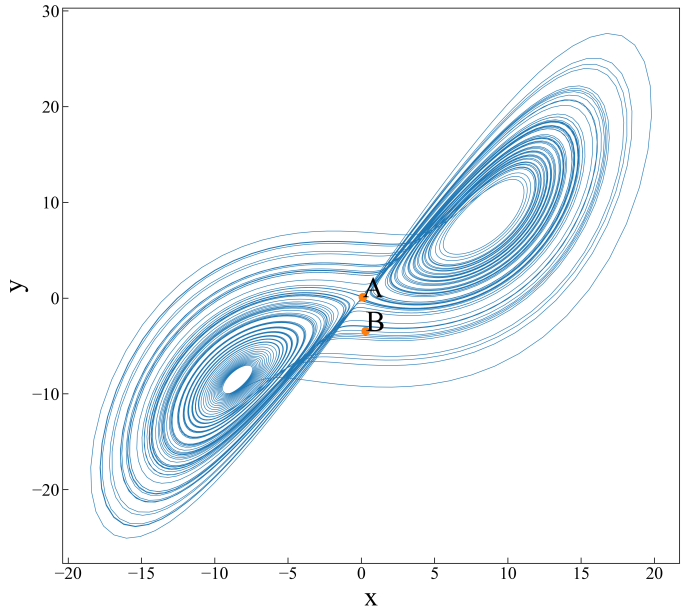


Figure 2: 2D attractor for Lorenz equation

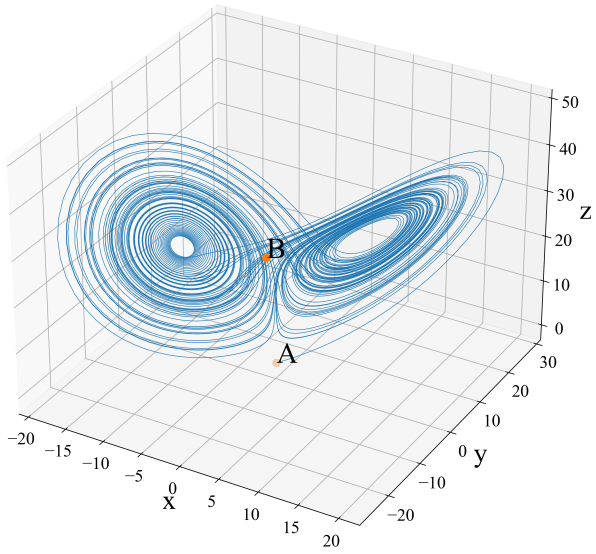


Figure 3: 3D attractor for Lorenz equation

2. FNN Methods

2.1. Conventional Methods

The authors proposed a method to estimate the minimum embedding dimension in 1992 [5]. Although this approach is simple in principle and works well for many nonlinear systems, it requires an appropriate threshold for every problem. The computational procedure of the FNN method is described in below.

Step 1: Reconstruct the attractor of the time series to be analyzed using Takens' embedding method [10]. Assuming that the target time series with N samples is x, the lag time is τ, and the embedding dimension is d, the d-dimensional attractor for x is reconstructed by $\{y_d(t)\}$ as follows.

$$\boldsymbol{y}_d(t) = \big(x(t), x(t+\tau), \cdots x(t+(d-1)\tau)\big)\,. \qquad (1)$$

Step 2: Calculate the time t' of the nearest neighbor vector of $\boldsymbol{y}_d(t)$, $\boldsymbol{y}_d^n(t)$, according to the distance criterion D.

$$D_d(t,t') = |\boldsymbol{y}_d(t) - \boldsymbol{y}_d(t')|\,,\ \ 0 \le t' \le N-(d-1)\tau\,,\ \ t \ne t', \qquad (2)$$

$$t' = \underset{0\le i\le N-(d-1)\tau}{argmin}\ D_d(t,i)\,, \qquad (3)$$

$$\boldsymbol{y}_d^n(t) = \boldsymbol{y}_d(t')\,. \qquad (4)$$

However, the set of x values for which the function $f(x)$ on some set $\boldsymbol{A}$ is minimal is denoted as follows.

$$\underset{x\in\boldsymbol{A}}{argmin}\, f(x) = \{x \in \boldsymbol{A} \mid f(x) = \underset{y\in\boldsymbol{A}}{min} f(y)\}\,. \qquad (5)$$

Step 3: Determine the distance $D_d(t,t')$ between $\boldsymbol{y}_{d+1}(t)$ and $\boldsymbol{y}_{d+1}^n(t)$ when the embedding dimension is $d+1$.

$$\boldsymbol{D}_{d+1}(t,t') = |\boldsymbol{y}_{d+1}(t) - \boldsymbol{y}_{d+1}^n(t)|\,. \qquad (6)$$

Step 4: Calculate the percentage of vectors in the attractor that are greater than or equal to the given threshold R_{tol} (false neighbor rate) relative to the rate of change in the neighbor point distance $D_{d+1}(t,t')/D_d(t,t')$ with increasing embedding dimensions.

$$\frac{\boldsymbol{D}_{d+1}(t,t')}{\boldsymbol{D}_d(t,t')} > R_{tol}\,. \qquad (7)$$

Step 5: Repeat the operations in **Step 1** to **4** for the given maximum number of embedding dimensions. Then, count from the lowest dimension to the first dimension for which the false neighbor ratio is zero and set the latter as the minimum embedding dimension.

2.2. Proposed Method

According to this procedure, a neighborhood is considered false when the rate of increase exceeds the threshold value R_{tol} in **Step 4**. However, this value must be set for each time series to be analyzed. Therefore, in this study, we propose a false neighborhood reduction rate M_{tol} as a new criterion to replace the threshold R_{tol}. In **Step 4**, the rate of increase in the nearest neighbor distance with increasing dimensionality is calculated, and the percentage of vectors the rates of increase of which exceed the threshold value R_{tol} is calculated. In the proposed approach, the analysis is performed in the same manner up to the point at which the rate of increase of the nearest neighbor distance is obtained, except that the following calculation procedure is used for computable times: $\boldsymbol{T} = \{t | 0 \leq t \leq N - (d-1)\tau\}$.

Step 4': For the vector $\boldsymbol{y}_d^n(T) = \boldsymbol{y}_d(\boldsymbol{T}')$ that is the nearest neighbor to the vector $\boldsymbol{y}_d(\boldsymbol{T})$ of the attractor at time $\boldsymbol{T}$, find the rate of change of the nearest neighbor distance with increasing embedding dimensions. After applying the ordinary logarithm, the median value is M_{tol}, as given below.

$$M_{tol} = Med\left(\log_{10}\frac{\boldsymbol{D}_{d+1}(\boldsymbol{T},\boldsymbol{T}')}{\boldsymbol{D}_d(\boldsymbol{T},\boldsymbol{T}')}\right). \quad (8)$$

Step 5': With increasing embedding dimension d, the dimension in which M_{tol} first becomes less than 1 is considered the minimum embedding dimension, being that in which embedding without false neighbors is established.

Generally, time series may contain noise. Even for time series generated from systems of the same type, noise levels may vary considerably depending on the observation method. For example, consider the case in which the minimum embedding dimension is estimated using the FNN method for a time series with a higher noise level. Owing to the higher noise level, the trajectory of the reconstructed attractor becomes unstable. In this case, the variance of the rate of change of the distance to the nearest neighbor increases with increasing dimensionality. Therefore, the number of points that exceed the given threshold (false nearest neighbors) increases, and the estimated minimum embedding dimension is expected to be high compared to observed time series with less noise. However, because the proposed method refers to the median of the rate of change of the nearest neighbor distance with increasing dimensionality, the effect of increasing variance is expected to be small. Therefore, the proposed approach could serve as a more objective evaluation criterion by eliminating the arbitrariness of the threshold value.

3. Evaluation Method

To compare the FNN method with the proposed method, we used a time series of numerical solutions calculated from a mathematical model in which the variables of the system were known in advance. We considered the van der Pol equation, which depends on two variables, and the Lorenz equation, which depends on three, to investigate whether the estimated embedding dimension was affected by changing the nonlinearity of the time series by setting multiple decay coefficients for the van der Pol equation. In addition, Gaussian noise of 0%, 1%, 10%, and 100% of the standard deviation of the signal was added to the numerical solutions computed from both equations to comprehensively examine their robustness to noise. The minimum embedding dimension was estimated for the time series generated with each parameter setting using both the FNN method and the proposed method, and the difference from the expected minimum embedding dimension was evaluated using the error rate described below.

To demonstrate its applicability to general time series, we applied the proposed approach to time series data recorded from electrooculograms to evaluate how many factors or variables could be identified as underlying nystagmus movements in the system.

3.1. Van der Pol Equation

The van der Pol oscillator is among the first examples of deterministic chaos discovered [11]. Van der Pol found a stable oscillation in an electrical circuit using vacuum tubes, which he referred to as the relaxation oscillation. The van der Pol equation is a two-variable differential equation in x and y described by the following equations (9) and (10) with the damping coefficient μ.

$$\dot{x} = \mu\left(y + x - \frac{x^3}{3}\right), \quad (9)$$

$$\dot{y} = -\frac{x}{\mu}. \quad (10)$$

In this work, we evaluated whether the numerical solution of the van der Pol equation could be calculated as a second-order system with a minimum embedding dimension for the numerical solution of van der Pol equation. A total of 80 van der Pol time series were generated, including 20 with decay coefficients $\mu = 0.1$, 0.2, ... 2.1, and 4 with noise levels of 0%, 1%, 10%, and 100% of the standard deviation of the original time-series signal. For all generated time series, the initial values of x and y were set to 0.1, with a time-step width $dt = 0.01$, and a time-series length $l = 10000$ (Figure 4).

3.2. Lorenz Equation

The Lorenz equation [8] is a nonlinear ordinary differential equation of three variables x, y, and z that exhibits chaotic behavior, as given below. The equations depend on three variables, as shown in the following equations (11), (12), and (13). Lorenz presented this classical equation in 1963 while working on a model of atmospheric variability as a meteorologist at the Massachusetts Institute of Technology [8].

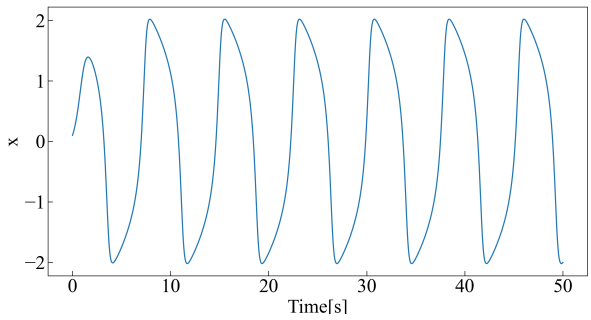


Figure 4: Typical example of van der Pol time series

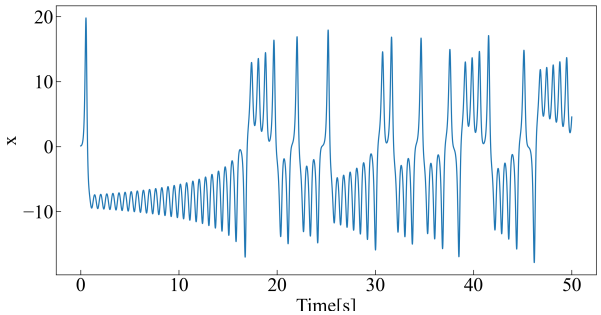


Figure 5: Typical example of Lorenz time series

$$\dot{x} = -px + py, \tag{11}$$

$$\dot{y} = -xz + rx - y, \tag{12}$$

$$\dot{z} = xy - bz, \tag{13}$$

where p, r, and b are constants that determine the behavior of the system.

Numerical solutions of the fourth-order Runge-Kutta method for the Lorenz equation were also obtained for reference, and the displacement x over time was plotted. For the parameter settings p=10, r=28, and b=8/3, the time interval was set to 0.01 second, and the initial values of x, y, and z were set to 0.1.

In this study, the numerical solution of the Lorenz equation was used to evaluate whether it could be calculated as a third-order system with a minimum embedding dimension. To evaluate the Lorenz time series, each parameter value was set to p=10, r=28, b=8/3, and four time series were generated with noise levels of 0%, 1%, 10%, and 100% of the standard deviation. For all generated time series, the initial values of x, y, and z were set to 0.1, the width of the time step dt = 0.01, and length of the time series l = 10000, respectively (Figure 5).

3.3. Electrooculogram data

An electrooculograms(EOG) is a time series recording changes in the distribution of electric potential around the eye caused by eye movements in two variables x and y. EOG are commonly used to examine the function of the retinal pigment epithelium and analyze eye movements. The electrical characteristics of the data involve a positive potential produced in the cornea closer to the environment and a negative potential produced closer to the retina. During measurement, single electrodes are affixed to a subject's inner and outer cornea, and their eye movements are recorded electrically via the potential difference between the electrodes [12].

Eye movements are divided into four types, including fixation, in which the eye concentrates on a single spot, pursuit, in which the eye slowly follows an object in the central fossa, saccadic and impulsive movements, in which the eye immediately detects anomalies, and vestibular movements, in which the eye responds to body movements when the body moves. The ocular muscles attached to the eyeballs move the eye and include the lateral rectus, medial rectus, superior rectus, inferior rectus, superior oblique, and inferior oblique muscles, as well as the upper eyelid elevator muscle, which is responsible for eye-opening action. Of the ocular muscles, the lateral rectus muscle is innervated by the abducens nerve, the superior oblique muscle by the pulmonic nerve, and all other muscles by the oculomotor nerve [13–15].

The EOG data used in this study were obtained from 11 healthy subjects (three men and eight women) aged 20-23 years. The subjects viewed experimental 3D images for 180 s each, and the angular velocity of their eye movement during this time was measured [16] (Figure 6).

3.4. Evaluation: definition of error value

We defined the error value as an index to evaluate the minimum embedding dimension obtained using the FNN algorithm and the proposed method. In this study, we used formula (14), where $\boldsymbol{D}$ is the minimum embedding dimension expected for the time series generated for each parameter setting, $\boldsymbol{D_{est}}$ is the minimum embedding dimension estimated from each method for those time series, and K is the total number of parameters when varying the attenuation coefficient and the amount of noise added using the error value E.

$$E = \frac{1}{K}\sum_{i=0}^{K-1}\left|\boldsymbol{D_i} - \boldsymbol{D_{est_i}}\right|. \tag{14}$$

4. Results

For Van der Pol Equation and Lorenz Equation, we report the estimated minimum embedding dimension and the error values of the expected minimum embedding dimension and the estimated values from the FNN method and the proposed method. However, for the Van der Pol Equation and Lorenz Equation, multiple time series were prepared under various conditions, so the estimated minimum embedding dimension is reported as an aggregate using the median value for each noise level. For the error between the expected minimum embedding dimension and the estimated value, we also report the error values aggregated by R_{tol} for the FNN method. The proposed method reports one error value per equation, since there is no threshold such as R_{tol}.

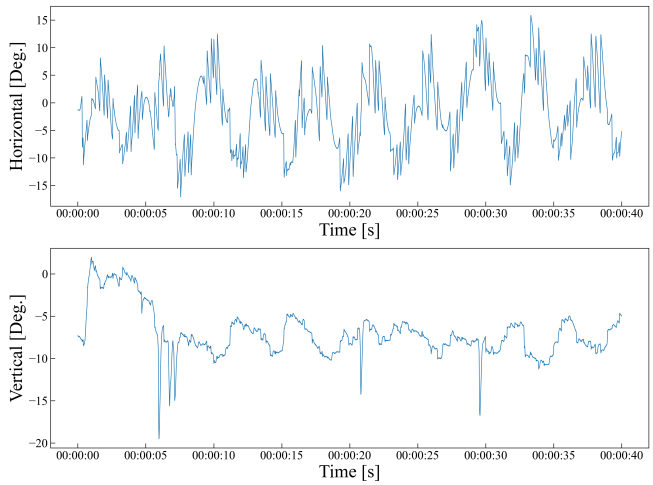


Figure 6: Typical example of an electrooculogram time series

Next, we report the results of estimating the minimum embedding dimension by each method for EOGs as an example of actual observed data.

4.1. Van der Pol Equation

First, we report on the estimated minimum embedding dimension for each method: the FNN method shows two dimensions when the noise level is 0%. The FNN method shows 2 dimensions when the noise level is 0%, 4 dimensions when the noise level is 1%, 5 dimensions when the noise level is 10%, and 6 dimensions when the noise level is 100%. On the other hand, the proposed method shows two dimensions when the noise level is 0% and 1%, and three dimensions when the noise level is 10% or higher.

Next, we report the error values for the estimated and expected minimum embedding dimension. When R_{tol} was set in the range of 1, 2, 4, and 1024 in the FNN method, the error value was minimized when R_{tol}=1024. In addition, the proposed method exhibited the same error value as when R_{tol}=1024 (Figure 7).

4.2. Lorenz Equation

First, we report on the estimated minimum embedding dimension for each method: the FNN method shows 6 dimensions for all noise levels from 0% to 100%, while the proposed method shows 2 dimensions for noise levels of 0% and 1%, and 3 dimensions for noise levels above 10%. On the other hand, the proposed method shows two dimensions for noise levels of 0% and 1%, and three dimensions for noise levels above 10%.

Next, we report the error values for the estimated and expected minimum embedding dimension. When R_{tol} was set in the range of 1, 2, 4, and 1024 in the FNN method, the error value was the smallest when R_{tol}=512,1024. In the proposed method, the error value was smaller than R_{tol}=256 and larger than R_{tol}=512, 1024 (Figure 8).

4.3. Electrooculogram data

For the variation in the horizontal axis, the minimum embedding dimensions were estimated using the FNN and proposed methods. With the FNN method, R_{tol} was estimated to be 1, 2, 4..., 1024. Assuming that the error value of the estimated minimum embedding dimension was smallest when R_{tol}=1024 in 4.1 and that 4.2 and can be applied to EOG data, the horizontal variation of EOG data by the FNN method was estimated and was considered as a system consisting of approximately 4 to 6 variables. In the vertical direction, the same can be considered as a system consisting of 8 to 11 variables. In contrast, using the proposed method, the minimum embedding dimension was estimated as two to three dimensions in the horizontal direction and two dimensions in the vertical direction, suggesting that the system is composed of approximately two to three variables (Figure 9-10).

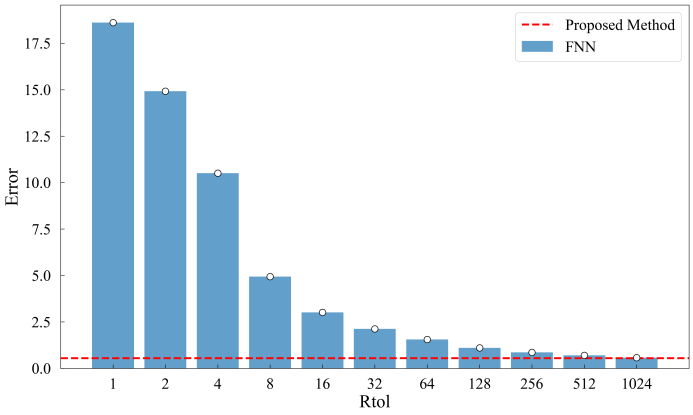


Figure 7: Error rate in estimating the minimum embedding dimension for van der Pol time series using the FNN method and the proposed method

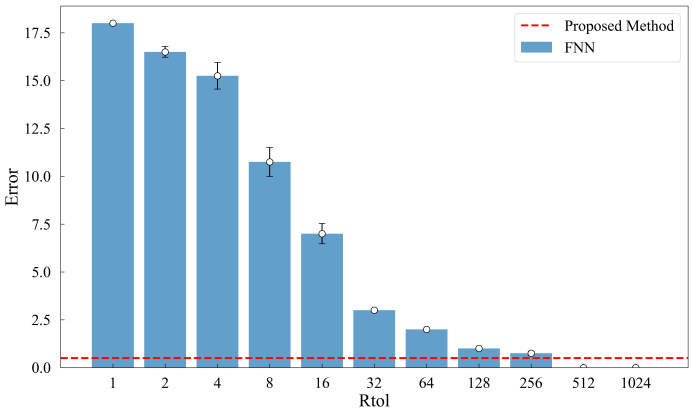


Figure 8: Error rate in estimating the minimum embedding dimension for Lorenz time series using the FNN method and the proposed method

5. Discussion

To compare the FNN method with the proposed approach, we considered a time series of numerical solutions calculated from a mathematical model in which the number of variables constituting the system is known in advance. In this study, the van der Pol equation consisting of two variables and the Lorenz equation

consisting of three variables were used, and the difference between the expected minimum embedding dimension and the minimum embedding dimension calculated using each method was evaluated as an error value. First, we discuss the estimation results of the minimum embedding dimension by the proposed method and the FNN method. Next, we discuss the error values between the minimum embedding dimension estimated by each method and the actual minimum embedding dimension. Finally, we discuss the results of estimating the minimum embedding dimension by each method for EOGs as an example of actual observed data.

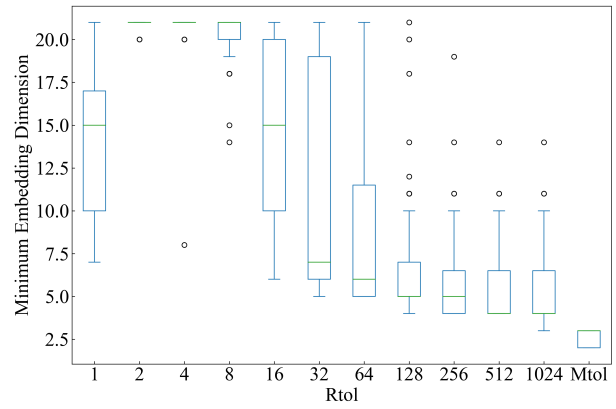


Figure 9: Estimation of the minimum embedding dimension for the horizontal-axis electrooculogram data

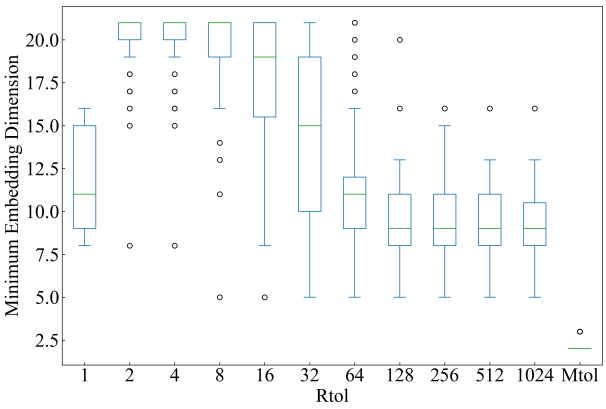


Figure 10: Estimation of the minimum embedding dimension for the vertical-axis electrooculogram data

5.1. Van der Pol Equation

The estimates of the minimum embedding dimension are discussed: both the FNN method and the proposed method show a minimum embedding dimension of 2 dimensions when the noise level is 0%. This is also the case for the FNN method in [17]. However, the FNN method suggested that the estimate may be vulnerable to noise, since an increase in the noise level to 1% resulted in a three-dimensional estimate. On the other hand, the proposed method did not affect the estimates until the noise level exceeded 10%, suggesting that it may be more robust against noise when compared to the FNN method.

Next, we discuss the error between the estimated first embedding dimension and the expected minimum embedding dimension for each method: the FNN method showed the smallest error when R_{tol}=1024. At this time, the proposed method also showed a comparable error. In other words, the proposed method does not require any threshold adjustment and shows the same error as the optimal value in the FNN method. Of course, since the error tends to decrease as R_{tol} increases, it is conceivable that the error would decrease more when the threshold is further increased. However, since the decrease in error stalls when R_{tol} exceeds 256, the decrease in threshold value is considered to be limited even if the threshold value is increased. In addition, it is difficult to determine the optimal value of R_{tol} because there are situations where the nature and data noise level are unknown in the actual time series to be analyzed. Even if the threshold is set in advance with a large value, it is believed that there may be cases where the minimum embedding dimension is underestimated.

5.2. Lorenz Equation

The estimates of the minimum embedding dimension are discussed: the FNN method estimated the minimum embedding dimension to be 6 under all noise levels of 0%, 1%, 10%, and 100%. It can be seen that even when the noise level is 0%, the results are very different from what one would expect. To reiterate, this estimate uses the median of the output values estimated for various R_{tol} settings, so it may output 3, the expected minimum embedding dimension, when the appropriate R_{tol} is set. Conversely, if an appropriate R_{tol} cannot be set, it is difficult to estimate the appropriate minimum embedding dimension. On the other hand, the proposed method is considered to be tolerant to noise because it could output the expected minimum embedding dimension even when relatively large noise levels of 10% and 100% were introduced. On the other hand, when the noise levels were relatively low (0% and 1%), the proposed method estimated one dimension less than the expected minimum embedding dimension. This may indicate that the minimum embedding dimension may be underestimated for time series with little disturbance.

Next, we discuss the error between the estimated embedding dimension and the expected minimum embedding dimension for each method: the FNN method showed the smallest error when R_{tol} = 512 and 1024. In this case, when R_{tol} is set to a value of 512 or 1024, the optimal minimum embedding dimension can be obtained even after considering the effects of various noises. On the other hand, if R_{tol} is set to a value higher than 1024, there is a possibility that the minimum embedding dimension will be underestimated. On the other hand, the proposed method had a smaller error value than the FNN method when R_{tol} was less than 512, but the FNN method had a smaller error value when R_{tol} was greater than 512. From these results, it can be said that if the optimal R_{tol} can be set, the FNN method has a smaller error value,

i.e., a more accurate estimation of the minimum embedding dimension, but in most cases, the proposed method has a more accurate estimation value.

5.3. Electrooculogram data

To demonstrate an application to general time series, we applied both the FNN method and the proposed method to EOG data to determine the minimum embedding dimension and examined how many variables the eye movements consist of in the system. The results show that the FNN method found 4 to 6 dimensions in the horizontal direction and 8 to 11 dimensions in the vertical direction. In contrast, the proposed method suggested 2 to 3 dimensions in the direction of the horizontal axis and 2 dimensions in the vertical direction. Several mathematical models of eye movements have been devised, and the Westheimer model s[18–20] is a representative example. In this model, human eye movements are described by a third-order differential equation, which is also close to the minimum embedding dimension calculated by the method proposed in this work.

6. Conclusion

In this study, we have proposed a method to solve the arbitrariness of the threshold used in the FNN method for time series. We have also applied the proposed method to EOG data as an example, and compared the performance of the FNN method and the proposed method in estimating the number of factors that determine eye movements. Compared to the FNN method, the minimum embedding dimension calculated by the proposed method exhibited a lower error value than expected under various conditions. The proposed approach mitigates the difficulty of setting appropriate parameters for a time series of unknown nature and obviates the need for an arbitrary threshold value. In future research, we intend to verify the effectiveness of the proposed method by applying it to a wider range of general time series.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems, 2014, doi:10.3156/jsoft.29.5_177_2.

[2] A. Radford, L. Metz, S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings, 2016.

[3] F. Kinoshita, Y. Mori, Y. Matsuura, H. Takada, M. Miyao, "A study of numerical analysis and mathematical modeling of electrogastrograms in the young subjects," IEEJ Transactions on Electronics, Information and Systems, **136**(9), 2016, doi:10.1541/ieejeiss.136.1261.

[4] Y. Jono, T. Tanimura, F. Kinoshita, H. Takada, "Evaluation of Numerical Solution of Stochastic Differential Equations Describing Body Sway Using Translation Error," Forma, **35**(1), 2020, doi:10.5047/forma.2020.006.

[5] M.B. Kennel, R. Brown, H.D.I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," Physical Review A, **45**(6), 1992, doi:10.1103/PhysRevA.45.3403.

[6] R. Hegger, H. Kantz, "Improved false nearest neighbor method to detect determinism in time series data," Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics, **60**(4 B), 1999, doi:10.1103/physreve.60.4970.

[7] S. Wallot, D. Mønster, "Calculation of Average Mutual Information (AMI) and false-nearest neighbors (FNN) for the estimation of embedding parameters of multidimensional time series in matlab," Frontiers in Psychology, **9**(SEP), 2018, doi:10.3389/fpsyg.2018.01679.

[8] E.N. Lorenz, "Deterministic Nonperiodic Flow," Journal of the Atmospheric Sciences, **20**(2), 1963, doi:10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2.

[9] J. Milnor, "On the concept of attractor," Communications in Mathematical Physics, **99**(2), 1985, doi:10.1007/BF01212280.

[10] F. Takens, Detecting strange attractors in turbulence, 1981, doi:10.1007/bfb0091924.

[11] Balth. van der Pol, " LXXXVIII. On 'relaxation-oscillations' ," The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, **2**(11), 1926, doi:10.1080/14786442608564127.

[12] W.C. Lara, B.L. Jordan, G.M. Hope, W.W. Dawson, R.A. Foster, S. Kaushal, "Fast Oscillations of the Electro-oculogram in Cystic Fibrosis," Investigative Ophthalmology & Visual Science, **44**(13), 4957, 2003.

[13] Z.M. Hafed, J.J. Clark, "Microsaccades as an overt measure of covert attention shifts," Vision Research, **42**(22), 2002, doi:10.1016/S0042-6989(02)00263-8.

[14] H. Kobayashi, Y. Watanabe, N. Ohashi, K. Mizukoshi, "Pendular Optokinetic Nystagmus Test in Patients with Central Nervous System Disorders," Practica Otologica, Supplement, **1989**, 1989, doi:10.5631/jibirinsuppl1986.1989.Supplement36_133.

[15] R. Engbert, R. Kliegl, "Microsaccades uncover the orientation of covert attention," Vision Research, **43**(9), 2003, doi:10.1016/S0042-6989(03)00084-1.

[16] A. Sugiura, R. Ono, Y. Itazu, H. Sakakura, H. Takada, "Analysis of Characteristics of Eye Movement While Viewing Movies and Its Application," Nihon Eiseigaku Zasshi. Japanese Journal of Hygiene, **77**, 2022, doi:10.1265/jjh.21004.

[17] E. Conte, A. Federici, G. Pierri, L. Mendolicchio, J.P. Zbilut, A brief note on recurrence quantification analysis of bipolar disorder performed by using a van der Pol oscillator model, 2009.

[18] G. Westheimer, "Mechanism of saccadic eye movements," A.M.A. Archives of Ophthalmology, **52**(5), 1954, doi:10.1001/archopht.1954.00920050716006.

[19] K. Shimono, M. Kondo, K. Shibuta, S. Nakamizo, "Psychophysical and vergence responses of normal and stereoanomalous observers to pulse-disparity stimulus," The Japanese Journal of Psychology, **53**(3), 1982, doi:10.4992/jjpsy.53.136.

[20] P.D.S.H. Gunawardane, R.R. Macneil, L. Zhao, J.T. Enns, C.W. de Silva, M. Chiao, "A fusion algorithm for saccade eye movement enhancement with eog and lumped-element models," IEEE Transactions on Biomedical Engineering, **68**(10), 2021, doi:10.1109/TBME.2021.3062256.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 4, 121-126 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Regularity of Radon Transform on a Convex Shape

Pat Vatiwutipong*

Department of Mathematics and Computer Science, Kamnoetvidya Science Academy, Rayong, 21210, Thailand

ARTICLE INFO

Article history:
Received: 18 June, 2022
Accepted: 16 August, 2022
Online: 24 August, 2022

Keywords:
Radon transform
Regularity property
Convex shape

ABSTRACT

Radon transform is a mathematical tool widely applied in various domains, including biophysics and computer tomography. Previously, it was discovered that applying the Radon transform to a binary image comprising circle forms resulted in discontinuity. As a result, the line detection approach based on it became discontinued. The d-Radon transform is a modified version of the Radon transform that is presented as a solution to this problem. The properties of the circle cause the Radon transform to be discontinuous. This work extends this finding by looking into the Radon transform's regularity property and a proposed modification to a convex shape. We discovered that regularity in the Radon space is determined by the regularity of the shape's point. This leads to the continuity condition for the line detection method.

1 Introduction

This paper is an extension of work originally presented in the 2022 14th International Conference on Knowledge and Smart Technology (KST) [1].

Radon transform is a well-known integral transform used in many fields that was originally invented in 1917 by Johann Radon, [2]. The Radon is used to create an image from the projection data associated with cross-sectional scans, [3]. Many other integral transforms, such as the Hough transform, Penrose transform, and Fourier transform, are significantly associated with the Radon transform.

Intuitively, Radon transformation is to integrate a given object and project the value onto the lower-dimensional space in various directions. In computer tomography, for example, the Radon transform was applied to a 3D object to obtain a 2D image. In practice, we acquire the 2D projected picture and must use the inversion of the Radon transform to reconstruct the 3D object, [4].

Hough transform is one of the line detection techniques used in image processing. By applying the Hough transform to an image, the image's point accumulation in each direction is computed. In the Hough space, a direction that passes through more points will have a greater value if the image comprises multiple points. The line direction in the image can be used to detect the maximal argument on the Hough space, [5].

In 2016, R. Aramini et al. found a unifying perspective of the Radon and Hough transformations [6]. The Radon transform is considered a continuous version of the Hough transform, so it was also applicable as a line detection method by changing the accumulation function to the integration of images in each direction instead. Since the Hough transform is a discrete transformation, the line detection method's continuity cannot be expected. However, it is different for the Radon transform.

In [1], we explained the benefits of the continuity of line detection method. It has been shown that the continuity of the line detection method using Radon transform depends on the regularity of the Radon space, but that work focused only on the image containing points. In practice, an actual image contains not only points but also shapes. For a grey-scale image as a function from a 2D image space to the interval $[0, 1]$, the value 0 and 1 represent white and black colors, respectively. If the image is continuous, we have that the Radon transform is also continuous. Nevertheless, image discontinuity is expected for a binary image whose range is just $\{0, 1\}$. This paper aims to study the regularity properties of the Radon transform of a binary image containing shapes, which lead to the continuity of the line detection method.

2 Definition and Elementary Properties

This section presents the definition of *Radon transform* and its basic properties.

Definition 2.1. (Radon transform) For a given function $m : W \subset \mathbb{R}^2 \to \mathbb{R}$, the Radon transform of m is defined, for each pair of real numbers θ and r,

$$\mathcal{R}[m](\theta, r) = \int_{r=x\cos\theta+y\sin\theta} m(x, y)ds, \tag{1}$$

*Corresponding Author: Pat Vatiwutipong, Email: pat.v@kvis.ac.th

whenever the integration is defined.

Lemma 2.2. *We have $(x, y) \in \ell_{\theta,r}$ if and only if there exists $s \in \mathbb{R}$ satisfying $x = r\cos\theta - s\sin\theta$ and $y = r\sin\theta + s\cos\theta$.*

According to this lemma, we can rewrite the definition of Radon transform as

$$\mathcal{R}[m](\theta, r) = \int_{-\infty}^{\infty} m(r\cos\theta - s\sin\theta, r\sin\theta + s\cos\theta)ds. \tag{2}$$

It was proven in [7] that the Radon transform $\mathcal{R}[m]$ exists for all $m \in L^1(\mathbb{R}^2)$.

For a function $h : E \to \mathbb{R}$, define a *norm* as in [7] by,

$$\|h\|_{\infty,1} = \max_{\theta\in[-\frac{\pi}{2},\frac{\pi}{2})} \int_{-\infty}^{\infty} |h(\theta, r)|dr. \tag{3}$$

It was also proven in [7] that, for $m \in L^1(\mathbb{R}^2)$,

$$\|\mathcal{R}[m]\|_{\infty,1} \leq \int_{\mathbb{R}^2} |m(x, y)|dxdy. \tag{4}$$

Since the Radon transform is linear, we have

$$\|\mathcal{R}[m] - \mathcal{R}[\tilde{m}]\|_{\infty,1} \leq \int_{\mathbb{R}^2} |m(x, y) - \tilde{m}(x, y)|dxdy, \tag{5}$$

which means that if there is a small change on the function m, the change on its Radon transform will also be small.

3 Regularity of Radon Transform

3.1 Regularity on Circle

First, let B_1 be a single circle in image space W centered at $P_1 = (x_1, y_1)$ with radius $R_1 > 0$. The Radon transform of this single circle is equal to the length of the intersection between the circle and $\ell_{\theta,r}$, so if the circle is centered at the origin, we have that

$$\mathcal{R}[B_1](\theta, r) = \begin{cases} 2\sqrt{R_1^2 - r^2} & \text{if } |r| \leq R_1 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Linearity and shifting property of the Radon transform are proven in [8]. From the shifting property,

$$\mathcal{R}[B_1](\theta, r) = \begin{cases} 2\sqrt{R_1^2 - (r - x_1\cos\theta - y_1\sin\theta)^2} \\ \qquad\qquad \text{if } |r - x_1\cos\theta - y_1\sin\theta| \leq R_1 \\ 0 \qquad\qquad \text{otherwise.} \end{cases} \tag{7}$$

Recall that we can represent $\mathcal{R}[P_1](\theta, r)$ by a sinusoidal curve $r_1(\theta) = x_1\cos\theta + y_1\sin\theta$, so we have that

$$\mathcal{R}[B_1](\theta, r) = \begin{cases} 2\sqrt{R_1^2 - (r - r_1(\theta))^2} & \text{if } |r - r_1(\theta)| \leq R_1 \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

In the Radon space E, the graph of $\mathcal{R}[B_1](\theta, r)$ can be seen as a "tunnel" centered at r_1 (see Figure 1).

When θ is fixed we can consider $\mathcal{R}[B_1](\theta, r)$ as a function of r which forms an upper-half of an ellipse centered at $r_1(\theta)$, with base length and height equal to $2R_1$ (see figure 2).

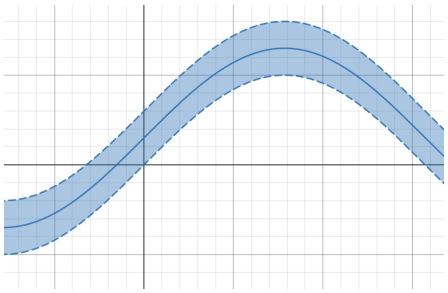


Figure 1: Graph of $\mathcal{R}[B_1](\theta, r)$ in E

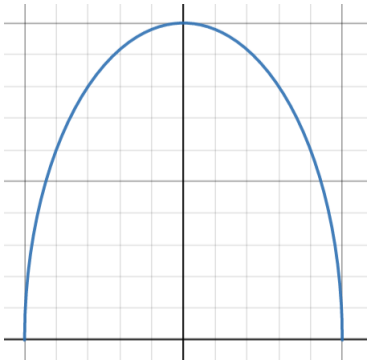


Figure 2: Cross section of $\mathcal{R}[B_1](\theta, r)$ for a fixed θ.

In this setting, the Radon transform $\mathcal{R}[B_1](\theta, r)$ is a continuous function on (θ, r) but it is not differentiable in r at $r = r_1(\theta) \pm R$.

To obtain more regularity, we modify the Radon transform by changing the integration domain from a line to a strip centered at that line instead.

Definition 3.1. (d-Radon transform) For $d > 0$ and $m \in L^1(\mathbb{R}^2)$, define the $d-$Radon transform of m, for each pair of real numbers θ and r, by

$$\mathcal{R}_d[m](\theta, r) = \int_{r-d}^{r+d} \mathcal{R}[m](\theta, \rho)d\rho. \tag{9}$$

We also have linearity and shifting properties for our new transformation. By the fundamental theorem of calculus and the fact that $\mathcal{R}[B_1](\theta, r)$ is continuous, the d-Radon transform $\mathcal{R}_d[B_1](\theta, r)$ is differentiable in r. Next, we will derive $\mathcal{R}_d[B_1](\theta, r)$ explicitly for each θ. For the sake of convenience, we shift the Radon space such that $r_1(\theta) = 0$. We have two cases.

Case I: If $0 < d \leq R_1$, we have that

$$\mathcal{R}_d[B_1](\theta, r) = \begin{cases} \int_{r-d}^{R_1} \mathcal{R}[B_1](\theta, \rho)d\rho & \text{if } R_1 \leq r \leq R_1 + d \\ \int_{r-d}^{r+d} \mathcal{R}[B_1](\theta, \rho)d\rho & \text{if } -R_1 \leq r \leq R_1 \\ \int_{-R_1}^{r+d} \mathcal{R}[B_1](\theta, \rho)d\rho & \text{if } -R_1 - d \leq r \leq -R_1. \end{cases}$$

That is $\mathcal{R}_d[B_1](\theta, r)$

$$= \begin{cases} \frac{\pi R_1^2}{2} - [(r-d)\sqrt{R_1^2-(r-d)^2} + R_1^2 \arctan(\frac{r-d}{\sqrt{R_1^2-(r-d)^2}})] \\ \qquad \text{if } R_1 \le r \le R_1 + d \\ [(r+d)\sqrt{R_1^2-(r+d)^2} + R_1^2 \arctan(\frac{r+d}{\sqrt{R_1^2-(r+d)^2}})] \\ \quad -[(r-d)\sqrt{R_1^2-(r-d)^2} + R_1^2 \arctan(\frac{r-d}{\sqrt{R_1^2-(r-d)^2}})] \\ \qquad \text{if } -R_1 \le r \le R_1 \\ \frac{\pi R_1^2}{2} + [(r+d)\sqrt{R_1^2-(r+d)^2} + R_1^2 \arctan(\frac{r+d}{\sqrt{R_1^2-(r+d)^2}})] \\ \qquad \text{if } -R_1 - d \le r \le -R_1. \end{cases} \quad (10)$$

The graph is shown in Figure 3.

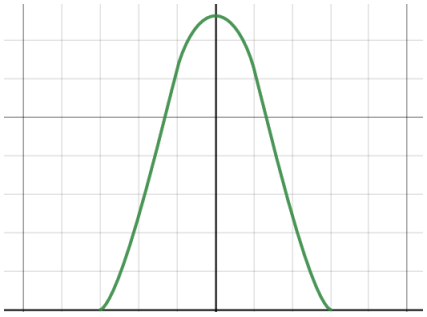


Figure 3: Cross section of $\mathcal{R}[B_1]_d(\theta, r)$ when $0 < d \le R_1$ for a fixed θ.

Case II: If $R_1 < d$, we have that

$$\mathcal{R}_d[B_1](\theta, r) = \begin{cases} \int_{r-d}^{R_1} \mathcal{R}[B_1](\theta, \rho)d\rho & \text{if } R_1 \le r \le R_1 + d \\ \int_{-R_1}^{R_1} \mathcal{R}[B_1](\theta, \rho)d\rho & \text{if } -R_1 \le r \le R_1 \\ \int_{-R_1}^{r+d} \mathcal{R}[B_1](\theta, \rho)d\rho & \text{if } -R_1 - d \le r \le -R_1. \end{cases}$$

That is $\mathcal{R}_d[B_1](\theta, r)$

$$= \begin{cases} \frac{\pi R_1^2}{2} - [(r-d)\sqrt{R_1^2-(r-d)^2} + R_1^2 \arctan(\frac{r-d}{\sqrt{R_1^2-(r-d)^2}})] \\ \qquad \text{if } R_1 \le r \le R_1 + d \\ \pi R_1^2 \qquad \text{if } -R_1 \le r \le R_1 \\ \frac{\pi R_1^2}{2} + [(r+d)\sqrt{R_1^2-(r+d)^2} + R_1^2 \arctan(\frac{r+d}{\sqrt{R_1^2-(r+d)^2}})] \\ \qquad \text{if } -R_1 - d \le r \le -R_1. \end{cases} \quad (11)$$

The graph is shown in Figure 4.

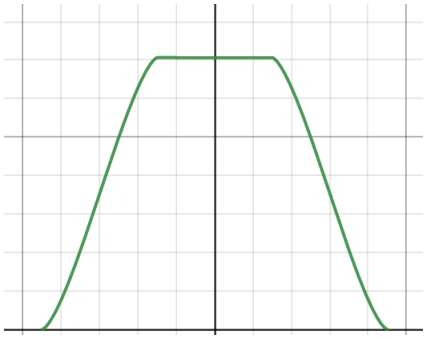


Figure 4: Cross section of $\mathcal{R}[B_1]_d(\theta, r)$ when $R_1 < d$ for a fixed θ.

For a fixed θ, $\mathcal{R}[B_1](\theta, r)$ reaches its maximum at $r = r_1(\theta)$. If $d \le R_1$, the maximum is $2d\sqrt{R_1^2 - d^2} + 2R_1^2 \arctan(\frac{d}{\sqrt{R_1^2-d^2}})$ and if $d > R_1$, the maximum is πR_1^2.

Lemma 3.2. *Let B_1 and B_2 be circles with same radius R, and fix $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2})$, we have that*

$$\max_{r\in\mathbb{R}} |\mathcal{R}[B_1](\theta, r) - \mathcal{R}[B_2](\theta, r)| = \min\{2R, 2\sqrt{2Ra_\theta - a_\theta^2}\} \quad (12)$$

and

$$\max_{r\in\mathbb{R}} |\mathcal{R}_d[B_1](\theta, r) - \mathcal{R}_d[B_2](\theta, r)| \le 2\sqrt{2}Ra_\theta \quad (13)$$

where $a_\theta = |r_1(\theta) - r_2(\theta)|$.

Proof. The proof of this lemma is straightforward. □

Now, we define another norm for the Radon space.
For a function $h : E \to \mathbb{R}$ define

$$\|h\|_\infty = \max_{(\theta,r)\in[-\pi,\pi)\times\mathbb{R}} |h(\theta, r)|. \quad (14)$$

Let B_1 be a circle and B_2 be another circle obtained by shifting B_1 with distance δ. We have that the value of $|\mathcal{R}[B_1](\theta, r) - \mathcal{R}[B_2](\theta, r)|$ will reach the maximum when $a_\theta = \|P_1 - P_2\| = \delta$. So, from Lemma 3.2, we get that $\|\mathcal{R}[B_1] - \mathcal{R}[B_2]\|_\infty = \min\{2R, 2\sqrt{2R\delta - \delta^2}\}$ which cannot be bounded by $C\delta$ for any constant C. Indeed, its derivative when δ is close to zero tends to infinity. So, in this case, a small change in the image will create a large change in its Radon transform.

For the d-Radon transform, the situation is totally different. Since we have a linear estimation in the previous lemma, we get that $\|\mathcal{R}_d[B_1] - \mathcal{R}_d[B_2]\|_\infty \le 2\sqrt{2}R\delta$. By linearity of d-Radon transform, we can extend this result to the case of images containing more than one circle. This leads to the next proposition.

Definition 3.3. Let $\mathfrak{B}$ be a set of all images containing a finite number of circles with the same radius $R > 0$. Define a distance function $\bar{D}$ on $\mathfrak{B}$ by $\bar{D}(m_1, m_2) = \bar{d}(m_1^P, m_2^P)$ where $\bar{d}$ is defined in Equation (4) and m^P is an image containing only the center of each circle in m.

Proposition 3.4. *Let $m_1, m_2 \in \mathfrak{B}$ be two images which contain N_1 and N_2 circles respectively, and $N_M = \max\{N_1, N_2\}$. We have that*

$$\|\mathcal{R}_d[m_1] - \mathcal{R}_d[m_2]\|_\infty \le 2\sqrt{2}RN_M\bar{D}(m_1, m_2). \quad (15)$$

3.2 Regularity on Convex Shape

Let m be an image that contains a single convex shape, that is, $m(x, y) = 1_A$, where A is a connected, compact and convex subset of W. Let curve C be the boundary of A. Suppose that C is continuous and piecewise differentiable.

Definition 3.5. Let P be a point on C. Let $\partial C(P)$ be the set of $\phi \in [-\pi, \pi)$ such that there exists r_ϕ which makes line $\ell_{\phi,r_\phi} : r_\phi = x\cos\phi + y\sin\phi$ intersects the set C only at point P or tangent of C in P.

Note that $\partial C(P)$ always exists for any P since C is convex, and it is a singleton when P is a differentiable point of C.

Remark 3.6. Since C is a closed curve, we get that the set $\{\phi : \phi \in \partial C(P), P \in C\}$ is equal to $[-\pi, \pi)$.

For all $\phi \in \partial C(P)$, let $\ell^{\perp}_{\phi,r_\phi}$ be the perpendicular line to ℓ_{ϕ,r_ϕ} passing through P. Denote the length of the segment of $\ell^{\perp}_{\phi,r_\phi}$ that intersected with A by D_ϕ. Parameterize $\ell^{\perp}_{\phi,r_\phi}$ by s. Set s to be zero at P and $s > 0$ in the direction that passes through A.

Definition 3.7. For a fixed P, define a function $f_\phi : \mathbb{R} \to \mathbb{R}$, for each ϕ, by letting $f_\phi(s)$ be the length of $\ell_{\phi,r_\phi+s}$ across A. Equivalently, $f_\phi(s) = \mathcal{R}[m](\phi, r_\phi + s)$. Note that $f_\phi(s) = 0$ for all $s \notin [0, D_\phi]$. When P is a differentiable point of C, since $\partial C(P)$ is a singleton, we may denote ℓ_{ϕ,r_ϕ} by ℓ_{θ_P,r_P} and f_ϕ by f_P.

All notations in definition 3.5 and 3.7 are shown in Figure 5.

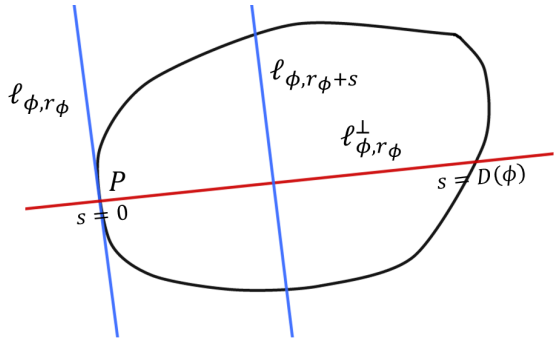


Figure 5: Point P on C and its ℓ_{ϕ,r_ϕ}, $\ell^{\perp}_{\phi,r_\phi}$ and D_ϕ

Definition 3.8. We say that point P is a *locally linear* point on C if there is a neighborhood of P where C is a line in it.

Lemma 3.9. *For all $s \in (0, D_\phi)$, the line $\ell_{\phi,r_\phi+s}$ is intersects with C in exactly two points.*

Proof. Suppose that there is $s \in (0, D_\phi)$ such that the line $\ell_{\phi,r_\phi+s}$ is intersects with C in more than two points. If one point is isolated, we can draw a line connecting that point to the others as in Figure 6 (b). It belongs to A, which contradicts the convexity of A. If there is no isolated point, there is a part of C that intersects with $\ell_{\phi,r_\phi+s}$ which is locally linear as in Figure 6 (c). Since C is the boundary of the shape, one side of it should not be in A. Without loss of generality, we can suppose that it is the left side. We pick the point on that part of C which is farthest from $\ell_{\phi,r_\phi+s}$ so that the line connecting that point and P will not be contained in A. Again, this contradicts the convexity of A. So the only case that can happen is that it crosses only two points as shown in Figure 6 (a). □

Proposition 3.10. *The function f_ϕ is continuous and piecewise differentiable on $(0, D_\phi)$.*

Proof. By previous lemma, we have that the value of $f_\phi(s)$ on $(0, D_\phi)$ is equal to the distance between the two intersection points of $\ell_{\phi,r_\phi+s}$ and C. Since C is continuous on $(0, D_\phi)$, so is f_ϕ. We can also get piecewise differentiability of f_ϕ by a similar argument. □

Proposition 3.11. *The curve C is locally linear at P if and only if f_P is not continuous at 0.*

Proof. First of all, note that $\lim_{s\to 0^-} f_P(s) = 0$. Suppose that C is locally linear at P, then C is differentiable at P. So, $f_P(0)$ is equal to the length of C that intersect with ℓ_{θ_P,r_P}, which is greater than zero. That is f_P is not continuous at 0. On the contrary, suppose that C is not locally linear at P. So, C must intersect ℓ_{θ_P,r_P} only at point P, that is $f_P(0) = 0$. Since C is continuous, we get $\lim_{s\to 0^+} f_P(s) = 0$. This proves continuity of f_P at $s = 0$. □

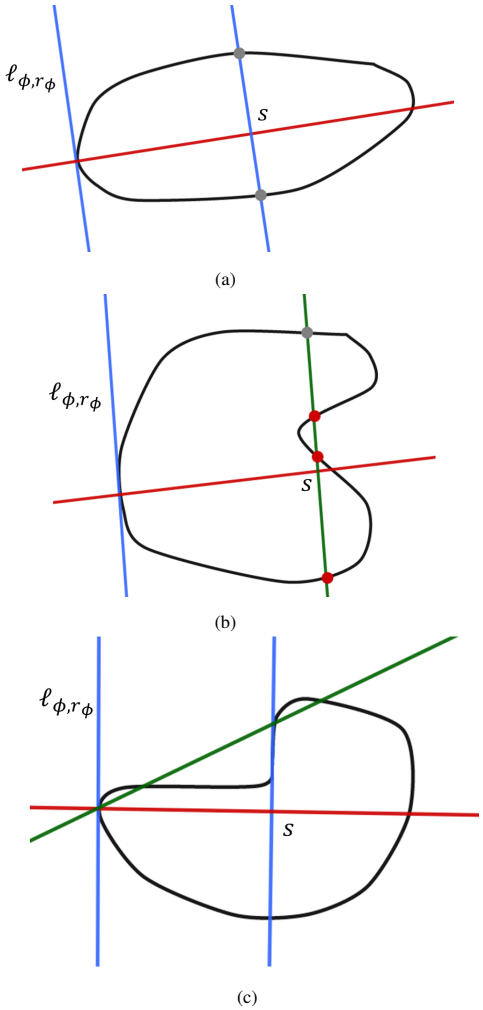


Figure 6: Non-differentiable point P on C and its ϕ^+, ϕ^-

Proposition 3.12. *If the curve C is differentiable at point P, then f_P is not differentiable at 0. Conversely, if f_ϕ is not differentiable at 0 for some $\phi \in \partial C(P)$, then C is differentiable at point P.*

Proof. Suppose that C is differentiable at P. Since ℓ_{θ_P,r_P} is tangent to C at P, we have $f'_P(0) = \infty$. On the contrary, suppose that C is not differentiable at P. There will be two distinct value $\phi^+ = \max\{\phi \in \partial C(P)\}$ and $\phi^- = \min\{\phi \in \partial C(P)\}$ as in Figure 7. Let $\phi \in \partial C(P)$. We get that $f'_\phi(s)$ will be bounded above and below by the slope of $\ell_{\phi^+,r_{\phi^+}}$ and $\ell_{\phi^-,r_{\phi^-}}$ evaluated with respect to ℓ_{ϕ,r_ϕ}. □

Hence, by Remark 3.6, we can separate $\theta \in [-\pi, \pi)$ into three types according to the property of its corresponding point P as: a locally linear, a differentiable but not locally linear and a non-differentiable points. The Radon transform is discontinuous for the locally linear point, and the d-Radon transform is continuous but not differentiable. For the differentiable point that is not locally linear, the Radon transform is continuous but not differentiable, and the d-Radon transform is differentiable. Lastly, at the non-differentiable point, both Radon and d-Radon transform are differentiable.

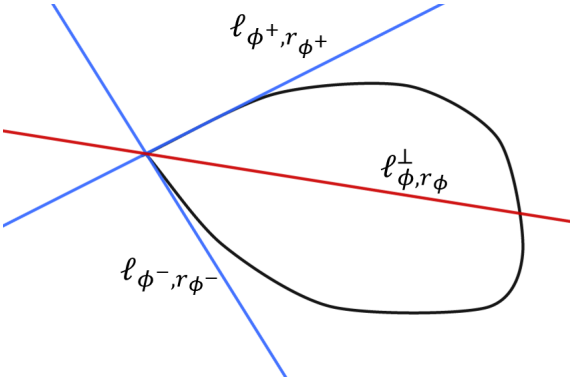


Figure 7: Non-differentiable point P on C and its ϕ^+, ϕ^-

To illustrated the result, take the following numerical examples.

Example 3.13. Consider a square

$$S = \{(x, y) : |x - y| + |x + y| \leq 2\}.$$

Only the corner points $(-1, -1), (-1, 1), (1, -1)$ and $(1, 1)$ are non-differentiable points of S, and the others are locally linear. So, the Radon transform of S is not continuous only in the horizontal and vertical directions, see Figure 8. For d-Radon transform, it is continuous in every direction.

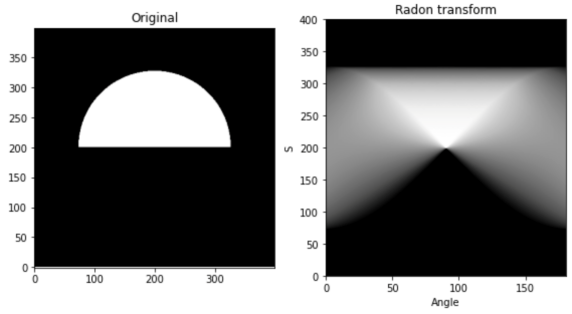


Figure 8: Image of the shape S in Example 3.13 and its Radon transform

Example 3.14. Consider a half circle

$$S = \{(x, y) : x^2 + y^2 \leq 1, y \geq 0\}.$$

Only the corner points $(-1, 0)$ and $(1, 0)$ are non-differentiable points of S. All point on a base of the half circle is locally linear and the other are non-differentiable. So, the Radon transform of S is not continuous only in the horizontal direction, $\theta = 0$. For the other directions, the Radon transform is continuous, see Figure 9. Again, for d-Radon transform, it is continuous in every direction.

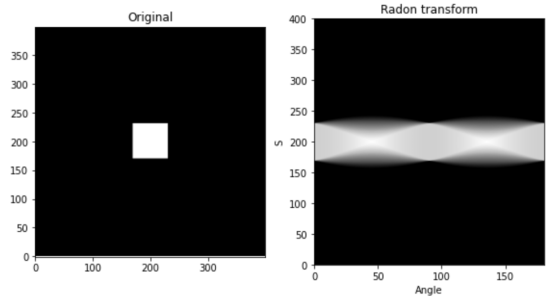


Figure 9: Image of the shape S in Example 3.14 and its Radon transform

4 Line Detection Methods

The line detection method is a procedure to indicate lines from a given image. In our setting, lines are parameterized by θ and r. So, line detection is a method to extract θ^{dt} and r^{dt} from each image. We can consider a line detection method as a set-valued function or correspondence that maps images to the set of detected parameters.

First, we consider the class $\mathfrak{B}_N$ of binary image consist of N circles on W. That is, $m \in \mathfrak{B}_N$ if $m : W \to \{0, 1\}$ in the form of

$$m(x, y) = \sum_{i=1}^{N} \delta_{B_{R_i}(P_i)} \tag{16}$$

where $B_R(P)$ denotes a circle centered at P with radius R.

We define a distance between images $m_1, m_2 \in \mathfrak{B}_N$ that consist of circles $B_{R_i}(P_i)$ for $i = 1, ..., N$ and $B_{S_i}(Q_i)$ for $i = 1, ..., N$ respectively by: $\bar{D} : \mathfrak{B}_N \times \mathfrak{B}_N \to [0, \infty)$ as $\bar{D}(m_1, m_2) =$

$$\min_{\sigma \in \Pi_N} \sum_{i=1}^{N} \left(2 \min\{R_i, S_{\sigma(i)}\} \|P_i - Q_{\sigma(i)}\| + \pi |R_i^2 - S_{\sigma(i)}^2| \right)$$

where Π_N denotes the set of permutations of $\{1, 2, ..., N\}$ and $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^2$. It was shown in [1] that the function $\bar{D}$ is a distance function on $\mathfrak{B}_N$. This distance is a specific case of the Wasserstein metric.

For the set of points, if all points are located on the given line $r^* = x \cos\theta^* + y \sin\theta^*$, we get that $(\theta^*, r^*) = \text{argmax}_{(\theta,r)\in E} \mathcal{R}[m](\theta, r)$ []. We can extend this result to the image consist of circles.

Theorem 4.1. *Let $m \in \mathfrak{B}_N$ and $d > 0$. If the center of all circles in m are on the line $r^* = x \cos\theta^* + y \sin\theta^*$, we get that*

1. $(\theta^*, r^*) = \text{argmax}_{\theta\in[-\frac{\pi}{2},\frac{\pi}{2})} \mathcal{R}[m](\theta, r)$,
2. $(\theta^*, r^*) \in \text{argmax}_{\theta\in[-\frac{\pi}{2},\frac{\pi}{2})} \mathcal{R}_d[m](\theta, r)$.

In particular, if $d < R_i$ for all radius R_i of circle consisted in m, we get that $(\theta^*, r^*) = \text{argmax}_{\theta\in[-\frac{\pi}{2},\frac{\pi}{2})} \mathcal{R}_d[m](\theta, r)$.

The proof of this theorem can be found in [1]. The next step is to show the continuity of the d-Radon transform. Here, the maximum theorem is needed.

Lemma 4.2. *(Maximum Theorem, see Theorem 3.5 in [9]) Let X, Y to be topological spaces. For a continuous function $f : X \times Y \to \mathbb{R}$ and a compact value correspondence $C : Y \rightrightarrows X$ such that $C(y) \neq \emptyset$ for all $y \in Y$. Define a function $f^* : Y \to \mathbb{R}$ by $f^*(y) = \sup\{f(x, y) : x \in C(y)\}$ and a correspondence $C^* : Y \rightrightarrows X$ by $C^*(y) = \operatorname{argsup}\{f(x, y) : x \in C(y)\}$. If C is continuous at y, then f^* is also continuous at y and C^* is upper hemi-continuous at y.*

Theorem 4.3. *For $d > 0$ and $m \in \mathfrak{B}_N$. A correspondence $m \mapsto (\theta^{dt}, r^{dt})[m] = \operatorname{argmax}_{(\theta,r)\in E}\mathcal{R}_d[m](\theta, r)$ is upper hemi-continuous under the norm induced by metric $\bar{D}$.*

Proof. Let $d > 0$. Consider $\mathcal{R}_d$ as a function from $E \times \mathfrak{B}_N$ to $\mathbb{R}$ and a correspondence C defined by $C(m) = [-\frac{\pi}{2}, \frac{\pi}{2})$ for all $m \in \mathfrak{B}_N$. To show that $\mathcal{R}_d$ is continuous, let $m_1, m_2 \in \mathfrak{B}_N$ and $(\theta_1, r_1), (\theta_2, r_2) \in E$. By linearity of the d-Radon transform, we have that

$$\begin{aligned}
&|\mathcal{R}_d[m_1](\theta_1, r_1) - \mathcal{R}_d[m_2](\theta_2, r_2)| \\
&= |\mathcal{R}_d[m_1](\theta, r) - \mathcal{R}_d[m_1](\theta_2, r_2) \\
&\qquad + \mathcal{R}_d[m_1](\theta_2, r_2) - \mathcal{R}_d[m_2](\theta_2, r_2)| \\
&\leq |\mathcal{R}_d[m_1](\theta, r) - \mathcal{R}_d[m_1](\theta_2, r_2)| \\
&\qquad + |\mathcal{R}_d[m_1](\theta_2, r_2) - \mathcal{R}_d[m_2](\theta_2, r_2)|
\end{aligned}$$

□

The last step follows from the fact that $\mathcal{R}_d$ is continuously differentiable. As a consequence, $\mathcal{R}_d : E \times \mathfrak{B}_N \to \mathbb{R}$ is continuous. Then, by maximum theorem, the correspondence $m \mapsto (\theta^{dt}, r^{dt})[m] = \operatorname{argmax}_{(\theta,r)\in E}\mathcal{R}_d[m](\theta, r)$ is upper hemi-continuous. So, if $\mathcal{R}[m]$ has a unique maximum, then a correspondence $m \mapsto (\theta^{dt}, r^{dt})[m] = \operatorname{argmax}_{(\theta,r)\in E}\mathcal{R}_d[m](\theta, r)$ is a continuous function.

Hence, we can extend this fact from an image consisting of circles to convex shapes. The main theorem can be restated as follows:

Theorem 4.4. *For $d > 0$ and image m consist of convex shape, if every point of edge of every shape in m are not locally linear, then correspondence $m \mapsto (\theta^{dt}, r^{dt})[m] = \operatorname{argmax}_{(\theta,r)\in E}\mathcal{R}_d[m](\theta, r)$ is upper hemi-continuous.*

Here, the metric used in this theorem is defined in the same way as $\bar{D}$ but on the image consists of convex shapes instead of circles.

5 Conclusion

The regularity of the Radon transform, its modification called the d-Radon transform, and line detection methods based on these two transforms have all been investigated. The line detection method based on the Radon transform is not continuous for an image consisting of a circle. After more analysis, we determined that the circle has a differentiable edge at each point, which is the source of discontinuity. The locally linear point, differentiable point, and non-differentiable point are three types of shape edges. The Radon transform's regularity is determined by the sort of point on edge in that direction. A non-differentiable point is the only situation in which the Radon transformation can preserve its continuity. To achieve Radon transformation continuity in all directions, the shape must not be differentiable at all points, which is impossible.

To increase the regularity of the line detection method, we introduce the d-Radon transform by changing the domain of integration from line to strip. Equivalently, it integrates the Radon space in the r direction. The d-Radon transform of a convex shape was explored for regularity. We discovered that the differentiable requirement of the transformation is that the shape's edge is not locally linear. This is also a condition for the line detection method's continuity.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The authors very much appreciate the support by my beloved native friend, Phatthamon Saenmuk, for editing my broken English.

References

[1] P. Vatiwutipong, "Continuity of line detection methods based on the Radon transform," in 2022 14th International Conference on Knowledge and Smart Technology (KST), 29–33, 2022, doi:10.1109/KST53302.2022.9729056.

[2] J. Radon, "On the determination of functions from their integral values along certain manifolds," in IEEE Transactions on Medical Imaging, 170–176, 1986, doi:10.1109/TMI.1986.4307775.

[3] S. R. Deans, The Radon Transform and Some of Its Applications, Springer, 2007.

[4] S. R. Deans, "Hough Transform from the Radon Transform," IEEE Transactions on Pattern Analysis and Machine Intelligence, **2**, 185–188, 1981, doi: 10.1109/TPAMI.1981.4767076.

[5] G. T. Herman, Fundamentals of Computerized Tomography, Springer, 2009.

[6] M. C. B. M. P. A. M. M. Riccardo Aramini, Fabrice Delbary, "Hough Transform from the Radon Transform," arXiv:1605.09201, 2016, doi:10.48550/arXiv.1605.09201.

[7] C. L. Epstein, Introduction to the Mathematics of Medical Imaging, Society for Industrial and Applied Mathematics, 2009.

[8] T. G. Feeman, The Mathematics of Medical Imaging: A Beginner's Guide, Springer, 2015.

[9] N. S. P. Shouchuan Hu, Handbook of Multivalued Analysis 1: Theory, Springer, 1997.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 3, 123-128 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Antenna System Design To Increase Power Transfer Efficiency with NFC Wireless Charging Technology

Jérémy Quignon[1,2,*], Anthony Tornambe[1], Thibaut Deleruyelle[3], Philippe Pannier[2]

[1] *STMicroelectronics, Rousset, 13106, France*

[2] *Université Aix-Marseille, IM2NP (UMR-CNRS 7334), Marseille, 13013, France*

[3] *Yncrea Mediterranée, IM2NP (UMR-CNRS 7334), Toulon, 83000, France*

ARTICLE INFO

Article history:
Received: 18 March 2022
Accepted: 25 May, 2022
Online: 13 June, 2022

Keywords:
NFC technology
Wireless Charging
Power Transfer Efficiency
IoT devices
Optimized Antennas

ABSTRACT

The NFC wireless charging feature is an extension of the NFC technology that can be implanted on wearables. The purpose of this paper is to show how to increase power transfer efficiency on both transmitter and receiver antenna systems. To demonstrate this problematic, firstly this paper gives an overview of how this NFC feature is implemented (architecture, power transfer, carrier frequency, communication bandwidth…), and can be complementary to Qi technology. Then, it provides a study on how to improve the power transfer efficiency on the antennas. To perform this result, the designer can adapt some antenna parameters as coupling coefficient, quality factor, matching method, and the antenna size. If these recommendations are respected, the power transfer efficiency between the antennas could reach between 70% and 80% with the NFC charging technology.

1. NFC Technology Overview

The NFC technology is initially used for contactless communication to exchange data. It is present in wearable applications such as authentication, payment, ticketing, pairing... However, these wearables are mechanical constraints because they embed a lot of components in a limited space. Moreover, for the wearables to be waterproof, and to avoid overheating issues, mechanical links (such as buttons or charger ports) must be removed, which leads to the use of wireless charging technology for batteries. To answer these problematics, a first paper was published [1], which described this new wireless charging technology as a solution.

The NFC wireless charging feature can use the same architecture (antenna, EMI filter, matching circuit, and NFC chip) as the NFC communication feature in order to operate [2]. So, a designer can use both NFC features on the same product [3].

Furthermore, the NFC antenna is a benefit to use NFC wireless charging in targeted devices. Indeed, since antennas are not standardized, they can take every possible size as shown on Figure 1. On the other hand, they do not have to be used with ferrite. So, the wearables that cannot use the Qi charging due to its features, like these standardized antennas with its ferrite, they can use the NFC wireless charging feature. Hence, all products can use wireless charging technology.

Nevertheless, the NFC technology has benefits and drawbacks. The issues that exist in NFC communication feature are also present in NFC charging feature. The main challenge for an NFC product is to be interoperable with other NFC devices available on the market because of their very characteristics, such as the antenna shape, waveform signal, or power signal. To avoid compatibility issues, the designer must check whether the new product, with its feature-related requirements, is compliant with the different existing standards: for instance, ISO 14443, NFC Forum, and EMVCo [4].

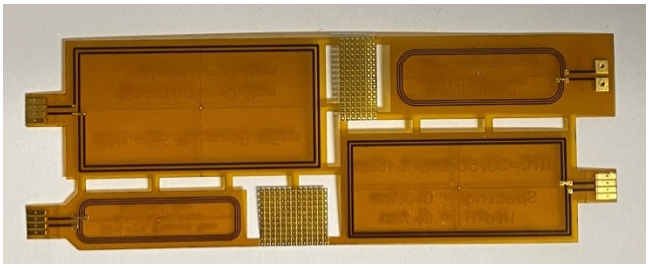


Figure 1: NFC wireless charging antenna with different sizes

*Corresponding Author: Jérémy Quignon, jeremy.quignon@st.com

Another issue is related to the NFC wireless charging, which was standardized only recently. The NFC Forum developed the standard for NFC charging at the end of 2021 to standardize the communication protocol, power supply, and product behavior in relation to issues such as foreign object detection [5]. This means that the products developed and marketed before this standardization operates with a proprietary mode.

The aim of the NFC wireless charging feature is to provide an alternative to Qi charging so that the user can benefit from better power transfer efficiency. This article describes this new wireless charging process and gives some recommendations to optimize the power transfer efficiency between the transmitter and receiver antennas, which is where the highest power losses occur.

2. NFC Wireless Charging Architecture

The NFC wireless charging feature is a radio frequency technology like NFC communication feature. It is based on a transmitter and a receiver, show on Figure 2.

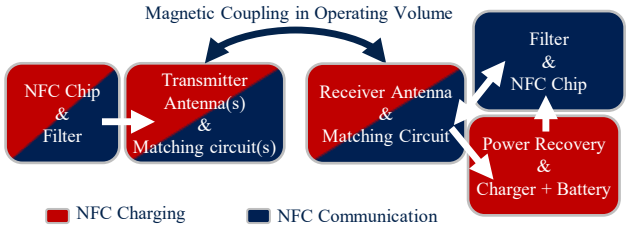


Figure 2: NFC wireless charging architecture

The transmitter uses the same architecture to perform both NFC features. It consists of:

- One NFC chip with its filter circuit, which provides power, and processes the exchanged data.
- One or two antenna(s) with its(their) matching circuit(s).

The difficulty on the transmitter board is on the matching circuit. Indeed, the calibration is different for the two NFC features [6]. So, to increase the power transfer efficiency, the transmitter must be matched in operating mode. This functionality is further described in part III.
On the other hand, to optimize the communication distance, the transmitter must be matched in the open-air mode (there is no metal object and/or antenna in the PCD's operating volume).

In this paper, we chose to use the same antenna and two different matching circuits to address both features.

The receiver, for its part, is composed of one antenna to receive the power and/or data signal with its matching circuit. The matching must be done in open air mode (without metal object in operating volume of the transmitter). Then, this antenna is connected to two different circuits:

- The NFC wireless charging circuit. It consists of a power recovery, in this architecture a diode bridge is used, and a charger chip with a battery.
- The NFC communication circuit. It consists of an EMI filter and an NFC chip to process the exchanged data.

The new NFC wireless charging uses a carrier frequency of fr=13.56MHz, as shown on Figure 3, which also gives the power limit that the device must not exceed. Indeed, the device must not generate a signal with carrier frequency higher than 60dBµA/m. Moreover, the power limit decreases with communication data rate (ΔF).

Furthermore, to enhance the power transfer efficiency, the system quality factor, Q, must be as high as possible. But, to optimize the communication bandwidth, the system quality factor must be limited.

So, if the transmitter and receiver use one antenna, it is important to find the best tradeoff in terms of quality factor to have the best power transfer efficiency sent, according to the resonance frequency fr, and to allow good communication, according to the bandwidth ΔF, between the two boards as shown on (1) and Figure 4.

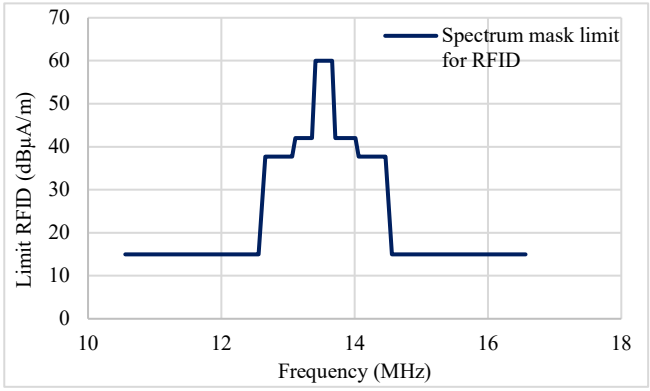


Figure 3: Power limit for RFID transmission

$$Q=\frac{f_r}{\Delta F} \quad (1)$$

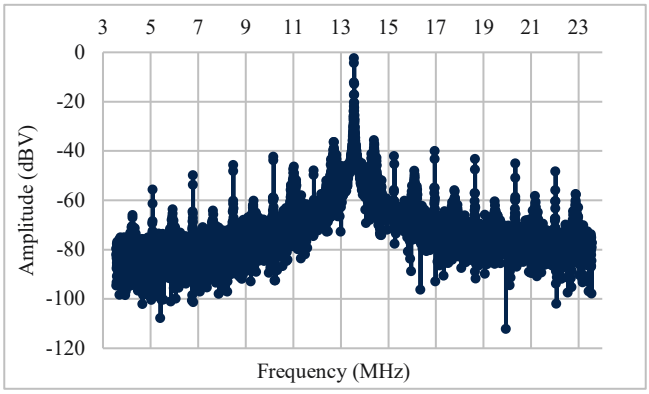


Figure 4: NFC signal with a quality factor allowing both NFC feature

The NFC Forum standardized the power transfer used for NFC charging. They indeed give several power classes as shown in Table 1. This power transfer is guaranteed at the receiver end, and the standard provides the maximum power as an indication. A designer can use this technology at higher power [7]. Moreover, we can see that NFC charging targets a lower power than the Qi power class shown in Table 2.

Table 1: NFC wireless charging power classes given by the NFC Forum.

	Power Classes	Minimum Power
Power Transfer	Class 0	250 mW
	Class 1	500 mW
	Class 2	750 mW
	Class 3	1000 mW

Table 2: Qi wireless charging power classes given by the WPC.

	Power Classes	Power
Power Transfer	BPP	5000 mW
	EPP	15000 mW

3. Optimization of antenna parameters to increase NFC power transfer efficiency

The purpose of the NFC wireless charging feature is to obtain best power transfer efficiency [8]. In order to reach this goal, it is important to mind one critical area, delimited by the transmitter and the receiver antennas themselves and the space between them. The power losses are indeed greatest in this area, and this is where the designer can increase the overall power transfer efficiency.

The Figure 5 shows the used boards to perform the test given on this chapter. The Figure 6 shows a block diagram of NFC wireless charging use to find the result exposed on this chapter.

The first parameter to increase the power transfer efficiency is to put the same antenna size on the transmitter board and the receiver board [9]. Furthermore, since this new technology is used in small devices, the selected antenna must be smaller than the size of the targeted wearable. For the tests described in this article, we used the class 6 antenna as the reference antenna for the transmitter and receiver board, as shown on Figure 7. This is the smallest antenna given by the ISO standard [10].

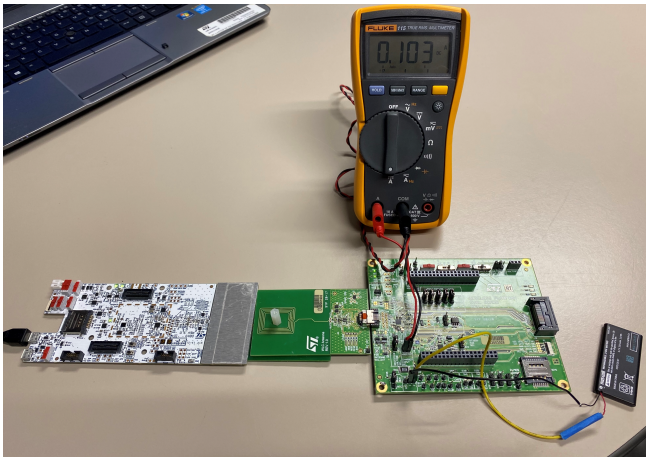


Figure 5: Power transfer efficiency measurement on test board

Therefore, if the transmitter's reference antenna is used in front of different receiver antenna sizes, as shown in Table 3, we can see, on Figure 8, the impact of the antenna design on the power transfer efficiency between the coupled antennas. The results show that the best efficiency is achieved when the PCD and PICC antennas are identical.

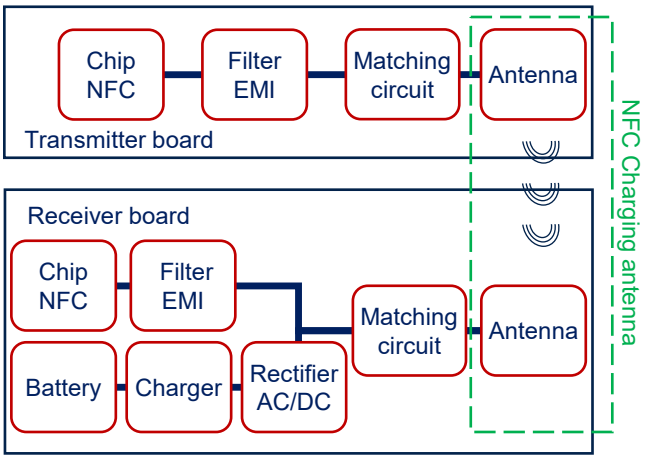


Figure 6: A block diagram of an NFC wireless charging system

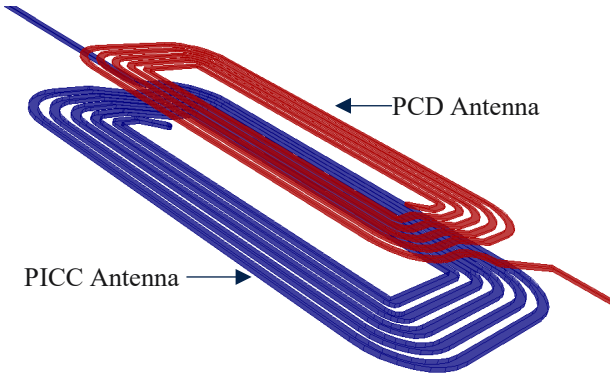


Figure 7: Transmitter and receiver antennas coupled

Table 3: Antenna design used for the different tests.

Receiver antenna size	Length	Width
Class 6 max size	25 mm	20 mm
Bigger than transmitter antenna	49 mm	30 mm
Smaller than transmitter antenna	20mm	20 mm

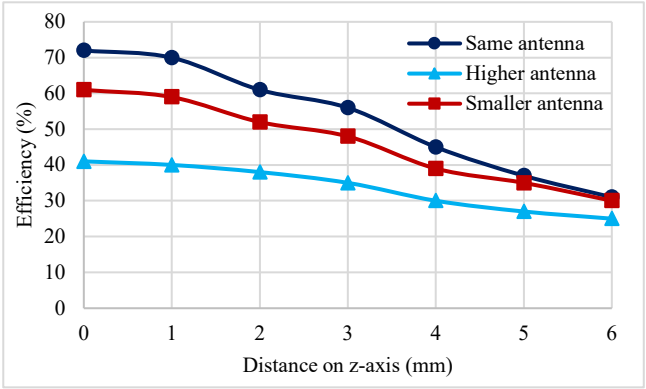


Figure 8: Power transfer efficiency simulated between the PCD and PICC antennas according to the distance along the z-axis

We can split the power losses between both transmitter and receiver antennas into three parts. The first two losses are on the real part of the PCD and PICC antenna impedance. To quantify these P_antenna losses, we measured the RMS voltage V_Antenna across the antennas. We then divided the measured value by the equivalent parallel resistance of the antenna, R_antenna, as shown on (2).

$$P_{Antenna} = \frac{V_{Antenna_{RMS}} \times V_{Antenna_{RMS}}}{R_{Antenna}} \quad (2)$$

The last loss is between the antennas due to the coupling coefficient. To decrease the impact of this parameter, the designer must use the same form factor antenna. Thus, the coupling coefficient k increases as shown on (3) [11]. It uses the equivalent radius of both the receiver (R_Rec) and transmitter R_Trans antenna. The two antennas are centered on a single x-axis.

$$k = \frac{R_{Trans}^2 \times R_{Rec}^2}{\sqrt{R_{Trans} \times R_{Rec}} \left(\sqrt{(R_{Trans}^2 + x^2)}\right)^3} \quad (3)$$

Figure 9 shows that, if we use the three receiver antenna shapes (defined in Table 3) in front of the transmitter with the reference antenna, the power transfer efficiency for a given coupling coefficient is better with the "same antenna" configuration.

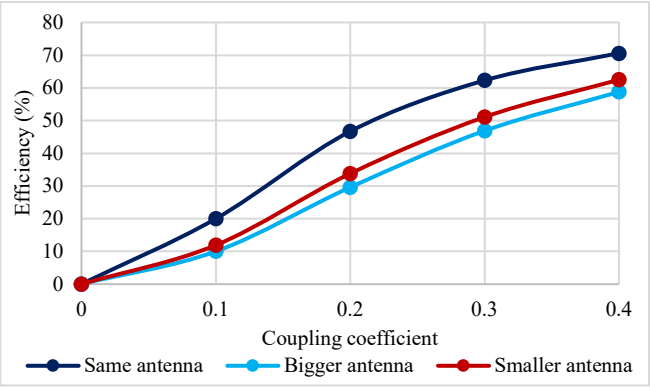


Figure 9: Power transfer efficiency simulated according to the coupling coefficient between the two antennas

The last antenna parameter that a designer can adjust to optimize the power transfer efficiency of the antenna is the antenna quality factor. Indeed, this parameter characterizes the resonance model and allows both the transmitted and received signal amplitudes to be increased [12]. Figure 10 presents the impact of the antenna quality factor at transmission (Tx) and reception (Rx) on power transfer efficiency. However, for the NFC communication use case, the transmitter and receiver quality factor must be adapted to keep a bandwidth large enough not to filter the data signal. The bigger the communication rate is, the smaller the transmitter and receiver quality is. In classical products, the transmitter and receiver quality factor are between 5 and 20.

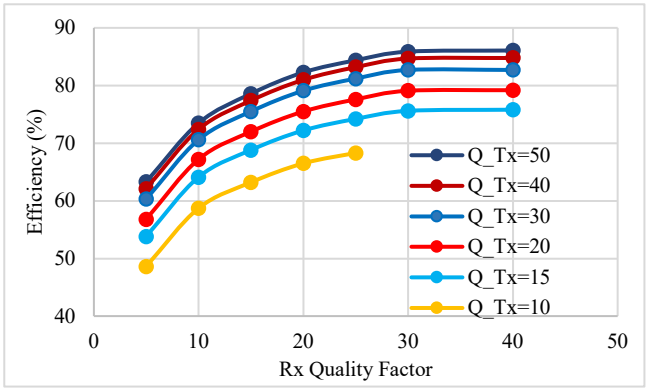


Figure 10: Impact of the quality factor on power transfer efficiency

In addition to the antenna parameters, the designer must keep in mind two other methods to optimize the power transfer efficiency. The first one is the space between both the transmitter and receiver antennas [13]. As announced in part II, both NFC feature use the same architecture. However, the operating volume is different. It is few centimeters for NFC communication whereas it is limited to only a few millimeters for NFC charging. To increase the performance, it is therefore preferable to reduce the operating volume to the location where the receiver receives the most power from the transmitter. Choosing this method also allows the transmitter and receiver boards to be in the matching position, and to increase the coupling coefficient. This again increases the power transfer efficiency.

To perform this test, a custom operating volume is created. Class 6 antennas are used on both transmitter et receiver for this test as shown on Figure 11.

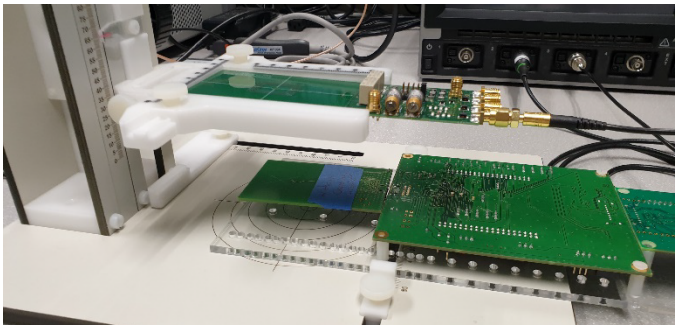


Figure 11. Operating volume test bench

Figure 12 shows power transfer efficiency curves in the custom operating volume. Initially both antennas are centered, and we apply a x offset for different z height. We can see that the power transfer efficiency is spatially limited to just a few millimeters or a single position.

The second method used to increase the power transfer efficiency even more, is to match the transmitter in operating mode. To perform this calibration, the receiver antenna is placed in operating volume of the transmitter antenna, where it receives the most power. Now, the transmitter is set up to send the maximum power when the receiver is near to it. This again shows that maximum power is delivered at the position where the

matching takes place and reduces the operating volume to a static position.

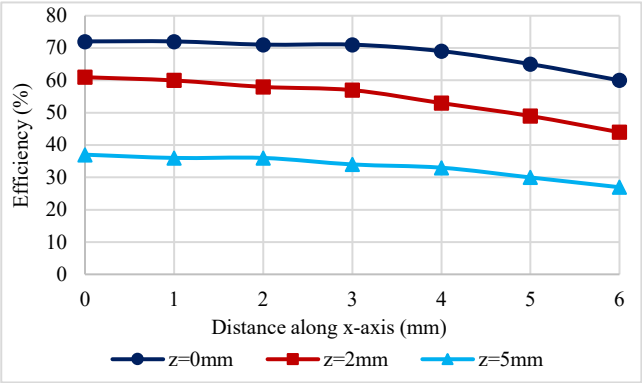


Figure 12: Power transfer efficiency in the operating volume according to the x- and z-axes

With this calibration method, the coupling effect has no more impact. Indeed, when the receiver antenna (L_Rec) is brought closer to the transmitter antenna (L_Trans) which leads to the detuning of the transmitter because the mutual inductance (M) changes as shown in (4). Because of this phenomenon the transmitter is no longer optimized to send maximum power [14]. It is why the transmitter must be matched in operating mode.

$$k=\frac{M}{\sqrt{L_{Trans} \times L_{Rec}}} \quad (4)$$

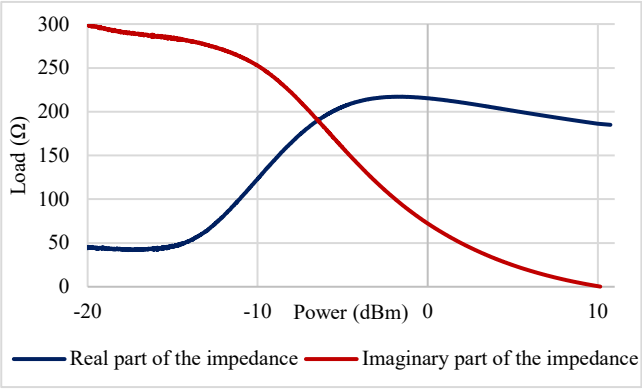


Figure 13: Diode bridge nonlinearity impact on the PICC impedance

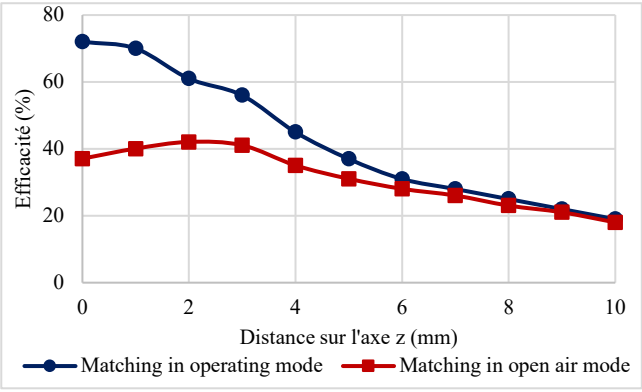


Figure 14: Power transfer efficiency depending on the matching mode

After determining the transmitter matching position, the designer must consider a second parameter to match the transmitter: the diode bridge non-linearity. Indeed, this parameter implies that the receiver impedance varies depending on the power that passes through it, as shown on Figure 13. To consider this nonlinearity, the designer must match the transmitter when it charges the receiver. Figure 14 shows the power transfer efficiency between both transmitter and receiver antennas with the impact of the matching mode used on the transmitter antenna.

4. Conclusion

This paper gives the architecture used by the NFC wireless charging feature. This technology is used to charge the embedded batteries in wearables. The designer can use the NFC products already available on the market to perform NFC charging. It must update the protocol to the new NFC Forum standard on both transmitter and receiver boards.

In addition to this protocol update, this article shows how to reach an optimized power transfer efficiency on the antennas. The results exposed in this article give the antenna parameters (antenna design, coupling coefficient, quality factor) and two additional methods (matching mode and operating volume) that a designer must pay attention to optimize the power transfer efficiency. If the advices are respected, the efficiency on antenna class 6 can reach between 70% and 80%.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

I would like to thank the co-authors who helped me in the writing of this paper with their patience, advice and experience. I would also like to thank the STMicroelectronics proofreading service for improving the English level of this paper.

Finally, I thank the ANRT organization which allowed the feasibility of my PHD in link with the laboratory IM2NP and the company STMicroelectronics.

References

[1] J. Quignon, A. Tornambe, T. Deleruyelle and P. Pannier, "Wireless charging using NFC technology," 6th International Conference on Smart and Sustainable Technologies", 1-4, 2021, doi: 10.23919/SpliTech52315.2021.9566468.

[2] S. Kolev, "Designing a NFC system," 56th International Scientific Conference on Information, Communication and Energy Systems and Technologies, 111-113, 2021, doi: 10.1109/ICEST52640.2021.9483482.

[3] H. Shibuya, T. Tsukuda, H. Suzuky, T. Shimizu, M. Dobashi, S. Nishizono, M.Baba, H. Sasaki, K. Terajima, "A wireless charging and near-field communication combination module for mobile applications," 64th Electronic Components and Technology Conference, 763-768, 2014, doi: 10.1109/ECTC.2014.6897371.

[4] N. S. S. Shobha, K. S. P. Aruna, M. D. P. Bhagyashree and K. S. J. Sarita, "NFC and NFC payments: A review," International Conference on ICT in Business Industry & Government, 1-7, 2016 doi: 10.1109/ICTBIG.2016.7892683.

[5] NFC Forum "Technical specification wireless charging version 2.0," 2021.

[6] L. Li, Z. Gao and Y. Wang, "NFC antenna research and a simple impedance matching method," Proceedings of 2011 International Conference on

Electronic & Mechanical Engineering and Information Technology, 3968-3972, 2011, doi: 10.1109/EMEIT.2011.6023890.

[7] https://www.nucurrent.com/nfc-wireless-charging-explained/ Accessed on January, 2022.

[8] M. Watanabe and K. Akatsu, "A study of high electrical power and high efficiency antenna in 13.56 MHz Wireless Power Transfer," Energy Conversion Congress and Exposition, 1027-1031, 2020, doi: 10.1109/ECCE44975.2020.9235829.

[9] B. Couraud, T. Deleruyelle, R. Vauche, D. Flynn and S. N. Daskalakis, "A Low Complexity Design Framework for NFC-RFID Inductive Coupled Antennas," IEEE Access, **8**, 111074-111088, 2020, doi: 10.1109/ACCESS.2020.3001610.

[10] ISO/IEC, 14443, "Cards and security devices for personal identification _Contactless proximity objects," 2018.

[11] K. Finkenzeller, RFID Handbook: Fundamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and Near-Field Communication, 1999.

[12] P. Baumgartner, O. Biro, W. Renhart, R. Torchio, C. Riener, C. Koger, T. Bauernfeind, "PEEC based design optimization of NFC antennas considering the full system quality factor," Tenth International Conference on Computational Electromagnetics , 1-4, 2019 doi: 10.1049/cp.2019.0123.

[13] W. Lee, J. Kang, J. Lee, Y. -J. Hong, J. Suh and S. J. Kim, "A Compact Base Station System for Biotelemetry and Wireless Charging of Biomedical Implants," Asia-Pacific Microwave Conference, 339-341, 2019, doi: 10.1109/APMC46564.2019.9038396.

[14] M. Meiller, Impact of antenna design tune and match on wireless range, High Frequency Electronics, IEEE Antennas and Propagation Magazine, **62**(5), 95-107. 2015.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 1, 125-130 (2023)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Northern Leaf Blight and Gray Leaf Spot Detection using Optimized YOLOv3

Brian Song[1], Jeongkyu Lee[2,*]

[1]Roslyn High School, Department, Institute, Roslyn Heights, 11577, USA

[2]Northeastern University, Khoury College of Computer Science, San Jose, CA, 95113, USA

ARTICLE INFO

Article history:
Received: 01 September, 2022
Accepted: 21 January, 2023
Online: 24 February, 2023

Keywords:
Northern leaf blight
Gray leaf
Corn disease
YOLOv3
CBAM

ABSTRACT

Corn is one of the most important agricultural products in the world. However, climate change greatly threatens corn yield, further increasing already prevalent diseases. Northern corn leaf blight (NLB) and Gray Leaf Spot are two major corn diseases with lesion symptoms that look very similar to each other, and can lead to devastating loss if not treated early. While early detection can mitigate the amount of fungicides used, manually inspecting maize leaves one by one is time consuming and may result in missing infected areas or misdiagnosis. To address these issues, a novel deep learning method is introduced based on the low latency YOLOv3 object detection algorithm, Dense blocks, and Convolutional Block Attention Modules, i.e., CBAM, which can provide valuable insight into the location of each disease symptom and help farmers differentiate the two diseases. Datasets for each disease were hand labeled, and when combined, the base YOLOv3, Dense, and Dense-attention had $AP_{0.5}$ NLB lesions/$AP_{0.5}$ Gray leaf spot lesions value pairs of 0.769/0.459, 0.763/0.448, and 0.785/0.483 respectively.

1. Introduction

Worldwide, 10 to 40% of crops die due to pests and diseases [1]. This issue will only get worse in the future, as temperature changes due to climate change will lead to more favorable conditions for pathogens [2]. In order to deal with this growing issue, it is necessary to come up with accurate diagnostic methods in order to quickly treat diseased plants. Focusing on maize, one of the most important crops in the world that accounts for two-thirds of the total volume of coarse grain trade globally in the past decade [3], in recent years, both Northern corn leaf blight, i.e., NLB, and Gray leaf spot disease has become more prevalent. In 2015, NLB was ranked first in most destructive corn disease in the northern United States and Ontario, Canada, up from seventh in 2012, with an estimated loss of 548 million bushels. For comparison, in 2015, the second-ranked most destructive corn disease, anthracnose stalk rot, had an estimated loss of 233 million bushels, less than half of NLB. In 2012, Gray leaf spot was ranked sixth in most destructive corn disease and has maintained its place as one of the top four most destructive from 2013-2015 [4]. NLB is caused by the fungus *Exserohilum turcium*, and the most distinguishing visual symptom of the disease is a cigar-shaped, tan lesion that can range from one to seven inches long, as shown in Figure 1. Gray leaf spot is a fungal disease caused by *Cercospora zeae-maydis* with rectangular lesions from two to three inches long, as shown in Figure 2., often leading to confusion by farmers due to its similarity to NLB lesions. Although fungicides can be used as a treatment for these two diseases, studies have shown that fungicides persist in aquatic systems and are toxic to organisms [5]. Early detection can mitigate the necessity of fungicides, traditionally through manual scouting [6]. However, this method is time-consuming and can result in inaccurate or missed diagnoses due to human error. As a result, several types of image-based machine learning solutions, such as convolutional neural networks, i.e., CNN, and object detection algorithms, have been proposed.



Figure 1: NLB infected maize images from Cornell CALS

CNNs are commonly used for image classification, which involves assigning an entire image to a single class label. However, in cases where the location of objects in an image is important,

*Jeongkyu Lee, San Jose, CA, 95113, jeo.lee@northeastern.edu

image classification can be difficult to interpret and verify. Object detection algorithms, on the other hand, can identify the specific location and extent of objects in an image, and can even draw bounding boxes around them to highlight their location. These algorithms can be useful in situations such as identifying disease symptoms in natural environment images, where it is important to know the exact location of the symptom. The typical symptoms of NLB and Gray leaf spot, which are brown, oval lesions, can be difficult to distinguish with the naked eye, but computer vision algorithms may be able to identify and differentiate them.



Figure 2: Gray leaf spot infected maize images from Cornell CALS

Many different object detection algorithms have been created, such as YOLOv3, which is both accurate and has low inference speed. Low inference speed is essential to speed up traditional manual practices and increases the likelihood of detecting a lesion on live video. However, one of the tradeoffs for its high inference speed is reduced accuracy. As a result, an optimized algorithm based on YOLOv3 was proposed in this paper by applying two methods: Dense blocks and convolutional block attention modules, i.e., CBAM. These optimizations were chosen to increase accuracy while maintaining or reducing inference speed compared to the base YOLOv3. In addition to proposing an improved algorithm, one of the main challenges of object detection is the need for more high-quality datasets, especially in niche areas such as plant disease detection. It is much easier to create image classification datasets because the label for an entire image is a single class or word. For object detection datasets, each image may contain more than one class, numerous objects per class, and requires the tedious work of locating all objects in an image and drawing bounding box labels around them. As a result, if there are no object detection datasets for a specific class, datasets originally for CNNs may be used instead by converting them to the correct format. In summary, this paper offers the following contributions:

- Application of machine learning to detect NLB and gray leaf spot lesions
- An optimized YOLOv3 with improved detection ability without significantly increased inference speed
- A NLB and grey spot dataset suited for object detection

This paper is an extension of work originally presented in IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC) [7], and the paper is structured as follows: Section 2 is the related works. Section 3 explains the YOLOv3 model and its optimizations. Section 4 shows how the dataset was created and explains the evaluation metrics. Section 5 shows the performance of the algorithms, and section 6 is the conclusion.

2. Related Works

Both image classification algorithms, such as CNNs and object detection algorithms, have been successfully applied for plant disease diagnosis. In [8], the authors used a CNN for plant disease diagnosis, training it on an extensive image dataset of close-up, individual diseased leaves using the Plant Village dataset. The dataset consists of images of twenty-six diseases of plants, such as those that affect corn and apple. They reported 99.35% accuracy on a held-out test set. However, the images in the dataset were taken in a lab environment, where variables such as the presence of multiple leaves in an image, orientation, brightness, and soil presenceare not considered. Solving this issue, high-quality datasets have been created by experts. In [9], the authors took images of NLB-infected corn and annotated each lesion individually with line annotations, which is the same dataset used in this paper. Because the images covered a large area, the actual lesions took up a small portion of the overall image. As a result, using a CNN on scaled-down versions of the images resulted in 70% accuracy. It was only after dividing the image into grids and associating a diseased-or-not class to each grid using their line annotations that resulted in 97.8% accuracy. Although successful, the process of dividing the image and then running inferences finally resulting in a high accuracy indicates a potential limitation of CNNs – they are not suited when the characteristics that define a class are small compared to the entire image.

On the other hand, object detection algorithms are suited to looking for specific areas in an image. The single shot detector, i.e., SSD [10] object detection algorithm, was used to detect apple diseases such as Brown Spot and Grey spot [11]. Instead of labeling an entire infected leaf as belonging to a class, the author only labeled the specific symptoms, such as spots, allowing the SSD algorithm to learn that the presence of a particular cluster of pixels determines the final output. The YOLOv3 [12] object detection algorithm was used to locate characteristics of various tomato diseases, such as early blight and mosaic disease [13]. The authors collected their own tomato dataset and annotated them by grouping clusters of diseased symptoms. In [14], the authors used a variant of an SSD and the Faster R-CNN [15] algorithm for grape plant disease object detection. They used existing datasets such as the Plant Village dataset, which mentioned previously is used for CNN training, drew boxes around the grape disease, converting it for object detection use. However, instead of drawing boxes around the grape disease symptoms, they label the entire leaf containing the disease, meaning it could be more specific. In [16], the authors also annotated the Plant Village dataset with bounding boxes for object detection. Instead of limiting to only certain disease classes, the entire dataset was used, resulting in about 54,000 annotated images. In [17], the authors created their object detection dataset called PlantDoc, which includes corn diseases, including both NLB and gray leaf spot. However, the dataset is small, less than 200 images per class, and the annotations are not very specific – clusters of lesions are grouped. This leads to the additional problem of determining whether two lesions belong to the same or different clusters, especially in images where lesions make up most of the corn leaf.

3. Methods

In this section, deep learning methods such as the usage of dense blocks, CBAM, and the proposed optimized algorithm to detect corn disease will be introduced.

3.1. Base Algorithm: YOLOv3

YOLOv3 is an object detection algorithm that boasts high accuracy without sacrificing speed. It is an improvement of YOLO

in 2016 [18] and YOLOv2 in 2017 [19]. Although newer variants exist, such as YOLOv4 [20], for this study, YOLOv3 was chosen as the algorithm to focus on because there is many open source code for the algorithm, meaning it was easier to find a repo with an implementation that could easily be modified. Also, YOLOv3, as shown in Figure 3, uses the darknet-53 backbone to extract features of images. The backbone network utilizes residual blocks containing skip connections, which was introduced in ResNet [21]. These skip connections skip some layers in the backbone, helping to alleviate the vanishing gradient problem and making it easier to tune the earlier layers of a network. In the figure, the residual N blocks consist of a 3×3 convolutional layer and N residual units. The detection stages contain additional convolutional layers for further feature extraction and detect potential objects on three different scales. These scales are used to detect large, medium, and small-sized objects. This improves the performance of varying image sizes. According to the YOLOv3 paper, performance on the Microsoft Common Objects in Context, i.e., MS COCO, dataset, a benchmark used for evaluating object detection algorithms in which the accuracy metric mean average precision, i.e., *mAP*, is commonly used, showed that the three fastest were YOLOv3-320, SSD321, and DSSD321 [22] with 51.5 *mAP*-50/22 *ms*, 45.4 *mAP*-50/61 *ms*, and 46.1 *mAP*-50/85 *ms*, respectively, indicating *mAP* and inference time. Compared to the other algorithms, YOLOv3 is significantly faster while achieving better *mAP*. As a result, YOLOv3 was selected as the base algorithm of our research because efficiency is critical for searching through large cornfields.

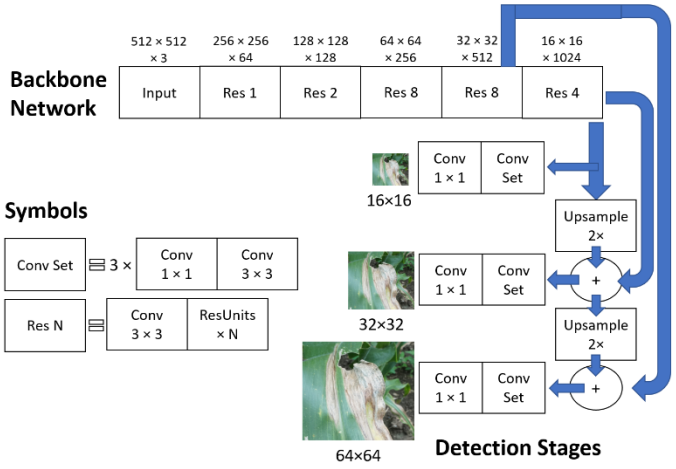


Figure 3: YOLOv3 diagram

3.2. Dense Block

DenseNet [23] is a convolutional neural network architecture that uses dense blocks, in which the output of each convolution layer is connected to the inputs of all subsequent layers. This allows later layers to use information learned in earlier layers and reduces the number of parameters, improving computational efficiency and mitigating the vanishing gradient problem. Transition layers, which consist of a 1×1 convolution and average pooling layer, are placed between groups of dense blocks to reduce the number of parameters and dimensionality. The growth rate, denoted by k, determines the number of new feature maps added for each layer in a dense block.

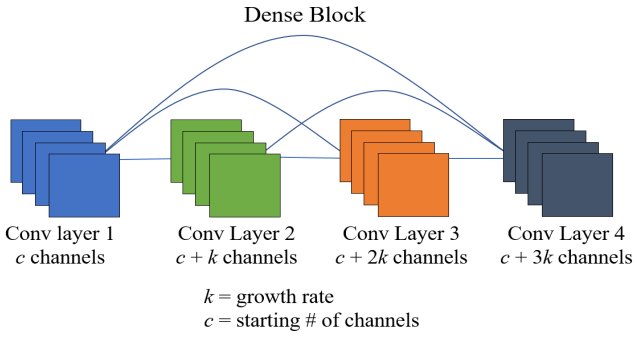


Figure 4: Diagram of 4 layer dense block

3.3. CBAM

CBAM [24] is a method that uses the attention mechanism to replicate how humans pay attention to their environment. It uses both channel and spatial attention to focus on particular objects in a scene and enhance the feature maps of the convolutional layers. Channel attention selects the most important channels and weighs them to improve them, while spatial attention applies max pooling and average pooling to help the model know where to focus in the image. Using this method leads to improved accuracy with only a limited amount of extra computation.

3.4. Proposed Algorithm

The proposed algorithm for an input image of 512x512 pixels is shown in Figure 5. The algorithm includes changes to the base YOLOv3 model, indicated by the yellow and green shaded portions. After the third residual block, a four-layer dense block is inserted, followed by a transition layer which reduces the number of filters and dimensions by half. The fourth and fifth residual blocks of the original backbone network are replaced with two six-layer dense blocks, with another transition layer in between. The growth rate, k, for all dense blocks is set to 128 to create a wide and shallow network for increased speed efficiency. CBAM blocks, which implement the attention mechanism, are placed at the beginning of the detection stages to refine the final feature maps and improve accuracy. The entire proposed model is called the Dense-attention algorithm, while the Dense algorithm is the same but without the CBAM blocks.

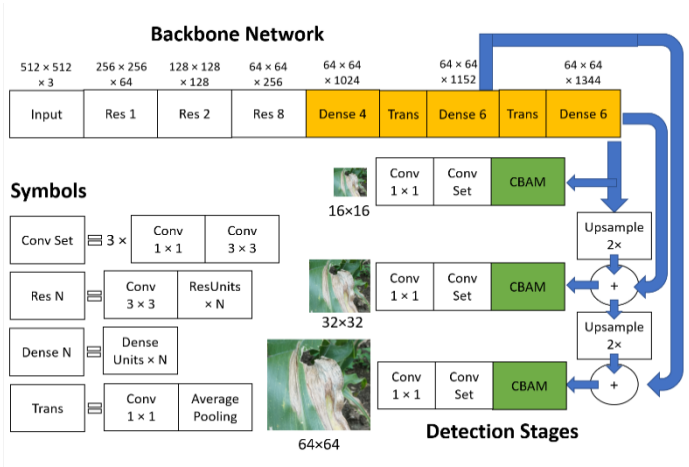


Figure 5: Proposed algorithm diagram

4. Experimental Setup

4.1. Datasets

To evaluate the proposed method in this paper, two datasets, one for NLB and the other for gray leaf spot, are combined. For the first dataset, the *Field images of maize annotated with disease symptoms dataset* [25] was used. The dataset consists of natural environment 4000×6000 pixel images of maize leaves taken by hand during the 2015 growing period in Aurora, New York. The main axis of the NLB lesions in each image is annotated with line annotations. 376 images are extracted. Because of the high image resolutions, each image is split into a 2×3 grid for pixel preservation. For the original line annotations, multiple lines would be used to signify one contingent lesion when only one label should be used. As a result, annotations are reconverted into bounding box format by hand under the supervision of the original line annotations. Figure 6 shows the annotation process. Images without lesions are discarded, resulting in 999 final images. Data augmentation using random rotation and zoom was then used to 4x the dataset size.

Unlike for NLB, high quality datasets for gray leaf spot disease taken from the natural environment and annotated by experts are rare. Although an annotated gray leaf spot dataset from Plant Doc exists, as mentioned in the related works section, few images are provided. Additionally, clusters of spots are grouped as one label, as opposed to the individual disease symptom labeling that the format of the NLB dataset is in. As a result, the Plant Village dataset was chosen to be annotated, starting from scratch. Because the dataset was initially intended for CNN training, annotations are not provided. As a result, like the NLB dataset, new annotations around each spot were created. However, in the gray leaf spot dataset, no expert guidance was provided, meaning best guesses for what constituted a spot were made. Figure 7 shows examples of grey spot annotations. Although fewer images of gray leaf spot were used compared to NLB, more annotations per image for gray were created, balancing the total number of annotations per class. The Plant Village dataset, in addition to the gray leaf spot, contains healthy corn images. 200 healthy images from each dataset were added. Several dataset groupings were used in this study.

Table 1: Dataset description

Dataset	Training images	Validation images	Testing images	Total label count
NLB	3,145	135	135	5,255
Gray	608	24	30	5,677
NLB+Gray	3,753	159	165	10,932
Healthy	4,033	219	225	10,932

Dataset NLB: NLB infected images only, Dataset Gray: Gray leaf spot infected images only, Dataset NLB+Gray: only NLB and gray leaf spot images, and finally, Dataset Healthy: a combination of dataset NLB+Gray with healthy images. Healthy images from the original NLB and Plant Village datasets were also included so that the algorithms could better learn what is not considered diseased through more examples. More info regarding dataset size and label counts is shown in Table 1. Data is available at a project github site[a].

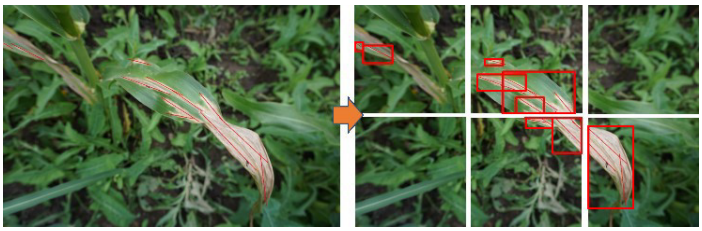


Figure 6: NLB annotation process

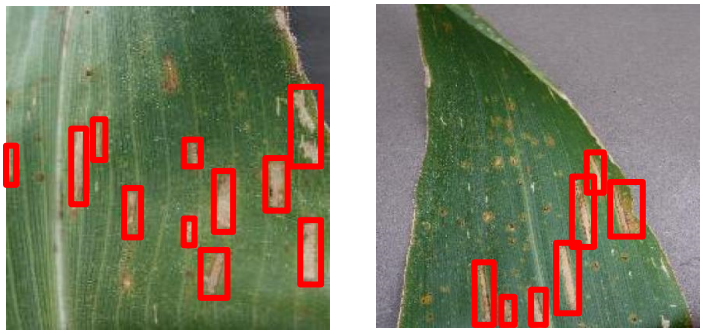


Figure 7: Grey leaf annotation examples

4.2. Evaluation Metrics

Evaluating accuracy for object detection algorithms requires comparing both the label and location of the predicted bounding box to those of the ground truth. *IoU*, i.e., intersection over union, as given in equation (1), is the value of the intersection area of the bounding box, $Area_{Pred}$, and the area of the ground truth box, $Area_{Truth}$, over the union area of the two aforementioned boxes.

$$IoU = \frac{Area_{Pred} \cap Area_{Truth}}{Area_{Pred} \cup Area_{Truth}} \quad (1)$$

If *IoU* is above a certain threshold value, that prediction is counted as a true positive. If not, it is considered a false positive. As shown in equation (2), precision, or PR, is the number of true positives divided by the sum of the number of true positives and false positives. Recall, or RE, shown in equation (3), is the number of true positives divided by the true positives and false negatives.

$$PR = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

$$RE = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

Average precision, i.e., AP, is calculated by finding the area under the precision-recall curve, shown in formula (4). P(r) is the precision value at recall value r. AP subscript k indicates the average precision when *IoU* is at threshold k. $AP_{0.5}$ indicates average precision at the 0.5 *IoU* threshold. If three lesions had *IoU* of 0.2, 0.6, and 0.3, only the lesion with the *IoU* 0.6 would be counted as a true positive.

$$AP = \int_0^1 p(r)dr \quad (4)$$

Inference speed, parameter count, and MFLOPs were also measured. MFLOPs is a unit for how many million floating point operations per second the computer can operate. All results were

[a]https://github.com/beans1321/NLB-Grey-dataset/tree/main

tested on the same GPU and Google colab environment settings for fair comparisons.

5. Results

Model training and evaluation were done on colab with an Nvidia P100 GPU, Intel Xenon 2.2 GHz CPU, and 13 GB of RAM. Code implementations were done using TensorFlow 1.15.0 and Keras 2.1.6 with Python 3.7. The three different algorithms, (i) *Base*, (ii) *Dense*, and (iii) *Dense-attention* were trained and evaluated on images sized 512×512 pixels.

Table 2 shows the $AP_{0.5}$ of the algorithms on the datasets On the NLB dataset, *Base*, *Dense*, and *Dense-attention* had $AP_{0.5}$ of 0.774, 0.806, and 0.821, respectively. In the Gray dataset, *Base*, *Dense*, and *Dense-attention* had $AP_{0.5}$ of 0.484, 0.471, and 0.496, respectively, showing that locating the exact boundaries of the grey spot is much more complicated than finding NLB lesions. One of the possible reasons for this is that NLB may appear more distinctive in the images since the lesions in the images tend to appear brighter than lesions in the Gray dataset. In the NLB+Gray dataset, individual $AP_{0.5}$ was recorded for each class. The $AP_{0.5}$ for each class decreased slightly compared to detecting them in dataset NLB and dataset Gray, in which only one class was present in each, which is expected as multiclass detection is more difficult than single class detection. The slight decrease in $AP_{0.5}$ also indicates that the model has learned to differentiate NLB and Gray leaf spot. In the Healthy dataset, the usage of images of healthy images decreased performance. All results for healthy were worse than the results for the NLB+Gray dataset. *Base* outperformed *Dense* and *Dense-Attention* in terms of finding NLB lesions, with $AP_{0.5}$ of 0.714, 0.675, and 0.702, respectively. However, in terms of finding gray lesions, *Dense-attention*'s $AP_{0.5}$ of 0.473 was still higher than *Base*'s $AP_{0.5}$ of 0.425. One possible reason that adding healthy images did not help performance is that the healthy images may have simply included more objects that looked like lesions, such as a dry or dead leaf, which are common, making training harder for the models.

Table 2: Performance of algorithms measured in $AP_{0.5}$

Dataset	Base	Dense	Dense-attention
NLB	0.774	0.806	0.821
Gray	0.484	0.471	0.496
NLB lesions in NLB+Gray	0.769	0.763	0.785
Gray lesions in NLB+Gray	0.459	0.448	0.483
NLB lesions in Healthy	0.714	0.675	0.702
Gray lesions in Healthy	0.425	0.428	0.473

Examples of detections are shown in Figure 8 and 9. Figure 8 shows detections of NLB by the *Dense-attention* trained from the NLB+Gray dataset. Figure 9 shows detections of Grey spot, also by *Dense-attention* trained from NLB+Gray. In Figure 9, it can be shown that the model has difficulty determining if two lesions close to each other are one or two lesions, as shown in the top row. In the bottom row, it can be shown that the model also has difficulty finding very small lesions.



Figure 8: NLB detection of *Dense-attention* trained on NLB+Gray dataset

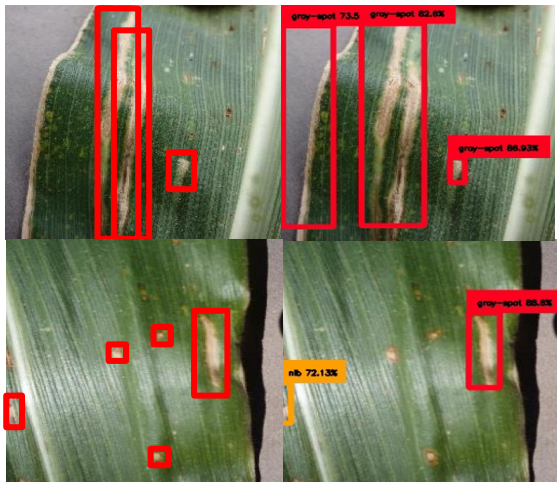


Figure 9: Gray detections of *Dense-attention* (right column) trained on NLB+Gray dataset side by side with ground truth (left column)

Table 3: General algorithm performance

Algorithm	Inference Speed (*ms*)	Parameters (millions)	MFLOPs
Base	36.9	61.5	123.0
Dense	37.4	40.8	81.6
Dense-attention	39.1	40.9	81.7

Table 3 shows the general performance of the algorithms, which shows constant results for the algorithms, no matter what dataset is used. *Base* has the fastest inference speed of 36.9 *ms*, followed by *Dense* with a speed of 37.4 *ms*, and finally *Dense-attention*, with a speed of 39.1 *ms*. There is not much of a major difference in inference speeds among all three algorithms. The usage of dense blocks to replace several layers of the original YOLOv3 backbone has greatly reduced the number of parameters. *Base* has 61.5 million parameters, *Dense* has 40.8 million, and *Dense-attention* has 40.9 million. The usage of CBAM has only increased the number of parameters slightly. The ratios of the parameter count between the algorithms are similar to the ratio of MFLOPs. *Base* has 123.0 MFLOPs, *Dense* 81.6, and *Dense-attention* 81.7. While *Base* is the fastest, *Dense* and *Dense-attention* are similar in speed while having drastically

fewer parameters and MFLOPs. In terms of accuracy, *Dense* seems to have similar results with *Base*, although it was more accurate in the NLB dataset. *Dense-attention,* based on the results shown in Table 2, mostly outperforms both *Base* and *Dense-attention*.

6. Conclusion

The new proposed model, called dense-attention, was built off of YOLOv3 and optimized for both accuracy and speed. New datasets for NLB and gray leaf spot were created and reannotated to be more suitable for object detection tasks. The results showed that dense-attention outperformed the base model in terms of accuracy, parameter count, and computational efficiency, although it was slightly slower. When both NLB and gray leaf spot were combined in the dataset, performance for each class decreased slightly compared to training on just one of the diseases. This suggests that the model was able to distinguish between the two visually similar diseases. In future work, it may be helpful to annotate the gray leaf spot dataset with experts and to include wider views of diseased leaves in the dataset.

References

[1] S. Savary, L. Willocquet, S. J. Pethybridge, P. Esker, N. McRoberts, A. Nelson, "The global burden of pathogens and pests on major food crops," *Nature Ecology & Evolution*, **3**(3), 430–439, 2019, doi:10.1038/s41559-018-0793-y.

[2] T. M. Chaloner, S. J. Gurr, D. P. Bebber, "Plant pathogen infection risk tracks global crop yields under climate change," *Nature Climate Change*, **11**(8), 710–715, 2021, doi:10.1038/s41558-021-01104-8.

[3] M. McConnell, "Feedgrains sector at a glance," USDA Economic Research Service U.S. Department of Agriculture, 27 January 2023, https://www.ers.usda.gov/topics/crops/corn-and-other-feed-grains/feed-grains-sector-at-a-glance/.

[4] D. S. Mueller et al., "Corn yield loss estimates due to diseases in the United States and Ontario, Canada from 2012 to 2015," Plant Health Progress, **17**(3), 211–222, 2016, doi:10.1094/PHP-RS-16-0030.

[5] J. P. Zubrod et al., "Fungicides: An Overlooked Pesticide Class?", Environmental Science & Technology, **53**(7), 3347–3365, 2019, doi:10.1021/acs.est.8b04392.

[6] K. Wise, "Northern corn leaf blight - purdue extension," Purdue University, 2011, https://www.extension.purdue.edu/extmedia/BP/BP-84-W.pdf.

[7] B. Song, J. Lee, "Detection of Northern Corn Leaf Blight Disease in Real Environment Using Optimized YOLOv3," *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, 2022, 0475-0480, doi:10.1109/CCWC54503.2022.9720782.

[8] S. P. Mohanty, D. P. Hughes, M. Salathé, "Using deep learning for image-based plant disease detection," Frontiers in Plant Science, **7**, 2016, doi:10.3389/fpls.2016.01419.

[9] C. DeChant, T. Wiesner-Hanks, S. Chen, E. L. Stewart, J. Yosinski, M. A. Gore, R. J. Nelson, H. Lipson, "Automated identification of northern leaf blight-infected maize plants from field imagery using Deep Learning," *Phytopathology®*, **107**(11), 1426–1432, 2017, doi:10.1094/PHYTO-11-16-0417-R.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, "SSD: Single shot multibox detector," Computer Vision– ECCV 2016, 21–37, 2016, doi:10.1007/978-3-319-46448-0_2.

[11] P. Jiang, Y. Chen, B. Liu, D. He, C. Liang, "Real-Time Detection of Apple Leaf Diseases Using Deep Learning Approach Based on Improved Convolutional Neural Networks", IEEE Access, vol 7, 59069–59080, 2019, doi:10.1109/ACCESS.2019.2914929.

[12] J. Redmon, A. Farhadi, 'YOLOv3: An Incremental Improvement', *arXiv*, 2018, doi:10.48550/ARXIV.1804.02767.

[13] J. Liu, X. Wang, "Tomato diseases and pests detection based on improved Yolo v3 convolutional neural network," Frontiers in Plant Science, **11**, 2020, doi:10.3389/fpls.2020.00898.

[14] S. Ghoury, C. Sungur, A. Durdu, "Real-Time Diseases Detection of Grape and Grape Leaves using Faster R-CNN and SSD MobileNet Architectures," International Conference on Advanced Technologies, Computer Engineering and Science (ICATCES 2019), 39-44, 2019 Alanya, Turkey.

[15] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**(6), 1137–1149, 2017, doi:10.1109/TPAMI.2016.2577031.

[16] M. H. Saleem, S. Khanchi, J. Potgieter, K. M. Arif, "Image-Based Plant Disease Identification by Deep Learning Meta-Architectures," *Plants*, **9**(11), 1451, 2020, doi: 10.3390/plants9111451.

[17] D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, N. Batra, "Plantdoc," *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, 2020, doi:10.1145/3371158.3371196.

[18] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, 'You Only Look Once: Unified, Real-Time Object Detection', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, doi:10.1109/CVPR.2016.91.

[19] J. Redmon, A. Farhadi, 'YOLO9000: Better, Faster, Stronger', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6517-6525, 2017, doi:10.1109/CVPR.2017.690.

[20] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, 'YOLOv4: Optimal Speed and Accuracy of Object Detection', *arXiv*, 2020, doi:10.48550/ARXIV.2004.10934.

[21] K. He, X. Zhang, S. Ren, J. Sun, 'Deep Residual Learning for Image Recognition', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778, 2016, doi: 10.1109/CVPR.2016.90.

[22] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, 'DSSD : Deconvolutional Single Shot Detector', *arXiv*, 2017, doi: 10.48550/ARXIV.1701.06659

[23] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, 'Densely Connected Convolutional Networks', *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269, 2017, doi:10.1109/CVPR.2017.243.

[24] S. Woo, J. Park, J.-Y. Lee, I. S. Kweon, 'CBAM: Convolutional Block Attention Module', *Proceedings of the European Conference on Computer Vision (ECCV)*, 11211, 3-19, 2018, doi: 10.1007/978-3-030-01234-2_1.

[25] T. Wiesner-Hanks, E. L. Stewart, N. Kaczmar, C. DeChant, H. Wu, R. J. Nelson, H. Lipson, M. A. Gore, "Image set for Deep Learning: Field images of maize annotated with disease symptoms," *BMC Research Notes*, **11**(1), 2018, doi:10.1186/s13104-018-3548-6.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 6, 126-138 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Extended Buffer-referred Prefetching to Leverage Prefetch Coverage

Jinhyun So*, Mi Lu

Texas A & M University, Department of Electrical & Computer Engineering, College Station, TX 77843, US

ARTICLE INFO

Article history:
Received: 14 August, 2022
Accepted: 21 November, 2022
Online: 08 December, 2022

Keywords:
Prefetching
Prefetch Coverage
Cache

ABSTRACT

This paper is an extension of the work originally presented in the 26th International Conference on Automation and Computing. This study regarding hardware prefetching aims at concealing cache misses, leading to maximizing the performance of modern processors. This paper leverages prefetch coverage improvement as a way to achieve the goal. Original work proposes two different storage buffers to enhance prefetch coverage; block offset buffer and block address buffer. The block offset buffer updates its contents with the offsets of a cache block accessed, while the block address buffer contains the address of a cache block prefetch-issued. The offset buffer is utilized to speculate a local optimum offset per page. The offset buffer is proposed to adopt multiple lengths of delta history in observing offset patterns from completely trained table. This paper advances to employ incompletely trained table as well, while in other prefetching methods including original work, only completely trained candidates are utilized. Furthermore, we construct the table on the fly. Rather than using only completely built tables, we offer utilizing and updating table concurrently. This paper also proposes a refined metric from existing prefetch accuracy metric, to measure net contribution of a prefetcher. Compared to the original work, we have 2.5% and 3.8% IPC speedup increment with single- and 4-core configuration, respectively, in SPEC CPU 2006. In SPEC CPU 2017, our work achieves 4.5% and 5.5% IPC speedup improvement with single- and 4-core configuration, respectively, over the original work. Our work outperforms the 2nd best prefetcher, PPF, by 2.9% and 2.7% IPC speedup with single- and 4-core configuration, respectively, in SPEC CPU 2006. In SPEC CPU 2017, our work surpasses both Berti by 1% and SPP by 2.1% IPC speedup with 4-core configuration in SPEC CPU 2017.

1 Introduction

IN [1], the author necessitates the implementation of memory hierarchy, which attempts to greatly shorten the average memory access time due to huge performance gap between the processor execution speed and memory latency. The execution speed of the processor has significantly increased while memory has pursued higher densities that causes increased memory latency. The enormous gap has been increasing due to the different objectives in developing the processor and memory. Hierarchy of cache memory has been introduced to reduce the performance gap by improving average memory access time, depending on two kinds of memory reference locality; temporal and spatial locality. However, the cache memory is still limited in reducing the gap with a trade-off between its capacity and the speed of cache hierarchy levels. That is, the cache levels closer to the cores are smaller size but have shorter latency. On the contrary, the cache levels farther apart from cores are of larger size but have longer latency. So, prefetching has been proposed as an effective technique that can bridge the performance gap by proactively fetching data ahead of processor's request into the cache closer to the cores. This paper proposes an effective technique in terms of prefetch coverage.

Prefetching is a mechanism that comprehends the memory access pattern of the program, and speculatively predicts and issues memory addresses ahead of the program's access to them. Hardware prefetching is to employ a standalone hardware that is dedicated to the prefetching. That is, hardware prefetcher predicts the future access to a memory data based on the observed access pattern of memory addresses in the past. Then, the prefetcher requests the data from low level of the memory hierarchy and stores it into a higher level cache close to the processor. Thus, by prefetching data, the prefetcher helps prevent cache misses due to the future access to the data, hiding long latency of low level cache access.

The objective of hardware prefetching is to proactively continue to fetch useful cache blocks from low level cache throughout run time based on memory access patterns. Usually, the algorithm

*Corresponding Author: Jinhyun So, & Email: jxs1525@tamu.edu

of a hardware prefetcher observes the past memory access patterns, predicts a future memory data access, and issues the address of the data. The patterns of a past memory access can be found based on the presumption of the existence of spatial and temporal locality of memory accesses.

In this paper, we propose an extended work of one of state-of-art prefetcher [2] with effective techniques maximizing prefetch coverage[1] at a moderate cost of prefetch accuracy. The contributions of this paper are categorized into the following three parts.

1.1 *Efficient use of hardware storage*

Following two buffers are employed for prefetching; one is block offset buffer and the other one is block address buffer to store the offset of a block and the address of prefetched block, respectively, in a FIFO manner. The offset buffer provides a reduction of hardware overhead by generating virtual tables[2] for prefetching instead of real table. The offsets of cache blocks are stored in the tables virtually generated whenever they are needed; storing the original offsets and generating the virtual tables can save hardware overhead by 45%[3] in terms of total hardware overhead.

1.2 *Maximizing prefetch coverage*

We propose following techniques for maximizing prefetch coverage from the two buffers as an offset prefetching: We use the history of a delta in multiple lengths to find diverse access patterns through access history. We mine access patterns from completely trained (virtual) table. Besides, they are also mined from incompletely trained (virtual) table.[4] Access patterns from the incompletely trained table are utilized until the table's training is completed. After the training is finished, access patterns from the completely trained table are used. Moreover, we also utilize an on-the-fly table that is on-going in building its entries. Taking advantage of the access patterns in the table under construction, prefetching can start as soon as possible even though there are few access patterns when new page is accessed for the first time. Moreover, to expand the opportunity in exploiting access patterns, we advance to make a use of entries of other table built for different pages. Especially, by referencing access patterns from the preceding pages, currently accessed page can have more diversified access patterns.

1.3 *Better accuracy metric*

We propose a modified metric in measuring prefetch accuracy by excluding the undetermined prefetches in existing metric. We name the metric as "accuracy ratio", which is the ratio of number of useful prefetches to the sum of number of useful prefetches and number of useless prefetches. The denominator of the existing metric is total number of prefetches which contains the undetermined prefetches in it. So, we construct denominator only with a determined portion, useful prefetches and useless prefetches, in order to well reflect prefetch accuracy.

2 Background

One of the early proposed hardware prefetching techniques is the simplest sequential prefetching, next line prefetching [3]. The next line prefetching is to prefetch a cache line that follows immediately the miss of the cache line. More advanced prefetching methods employ a prediction table by means of detecting memory access patterns. The table is used to record memory access history and identify its pattern. Also, a prefetching method is proposed for constant stride access pattern that refers to a sequence of memory accesses in which the distance of consecutive accesses is constant [4, 5]. The constant stride pattern also appears in pointer-based data structures [6].

In [7, 8], the author is proposed to find the most likely next address for currently accessed addresses by representing probabilistic correlation between accessed addresses with Markov model. This method needs not only a large storage to store the addresses into a table but it also requires high computational cost to calculate the correlation. Furthermore, Markov prefetching has stale data problem in the table.

Global history buffer (GHB) [9] as a FIFO manner is proposed to reduce a storage overhead as well as to solve a stale data problem. GHB holds recent miss addresses in the FIFO buffer and chains the same miss addresses in the buffer. Following the chain, deltas[5] are computed and adjacent deltas form a pair used as a prefetch key.[6]

Recent prefetching methods adopt a learning system since these mechanisms are possible to give a feedback of varying accuracy and coverage, depending on workloads [10]–[12]. In [12], the author dynamically adjusts prefetching aggressiveness in both positive and negative ways through a feedback system, to achieve better performance and bandwidth-efficiency.

Signature Path Prefetcher (SPP) [13] uses confidence value to adjust the length of a signature path to strike a balance between accuracy and coverage. SPP speculatively predicts memory access patterns, based on a 12-bit signature that represents a sequence of memory access. The 12-bit signature is calculated in the fashion of combining consecutive strides between adjacent accessed cache lines. The signature and a subsequent stride form a pair, used as a prefetching key and a prefetch prediction, respectively.

Perceptron-based Prefetch Filtering (PPF) [14] proposes an additional filter layer enhancing SPP as a way to increase prefetch coverage. PPF is a filter that uses a hased perceptron model to evaluate the usefulness of each prefetch generated by SPP, in order to reach better coverage. The perceptron model uses several features such as physical address, cache line, and page address, etc. to train PPF layer.

Recently, many offset prefetching descendants [11, 15, 16, 19]

[1] Prefetch coverage means the number of memory access patterns that are detected by prefetches. The coverage will be discussed in Sec. 2
[2] Virtual table refers an actually generated table for prefetching but it merely needs temporary hardware storage instead of permanent one
[3] Hardware overhead will be discussed in Sec. 4
[4] Incompletely trained table indicates a table of which the training period is in progress.
[5] Delta is a value of address difference between two adjacent addresses
[6] Each prefetch key has a corresponding prefetch prediction. Different prefetch keys may make different predictions.

have been proposed after Sandbox prefetching (SP)[20] was proposed. SP attempts to find a block offset that gives high performance. The offset is chosen from a set, named sandbox, of pre-selected offset candidates, based on calculated accuracy score of each offset candidate during run-time. Best-offset prefetching (BOP) [11] adds timeliness consideration to SP in order to pursue timely prefetching. BOP tests pre-selected offsets with arrived prefetch-requested block to figure out which offsets fit better in terms of latency of the arrival of prefetched block; BOP checks if the base address of previously prefetched block are equal to currently accessed block's address minus offset. If the equality is satisfied, the score for offset will increase. Otherwise, the score will decrease.

In [15], the author attempts better timely prefetching by calculating best offset page-by-page as compared to BOP's global best offset. The offset is calculated by recording access timing of each cache line in terms of the number of cycles. The measured cycles are referred to decide the proper latency of two arbitrary accessed cache lines in the same page. If there are cache lines accessed within the chosen latency, Berti uses burst mode for them.

Instruction Pointer Classifier based Prefetching (IPCP) [17] proposes two cache level, L1 and L2, prefetching at the same time with multiple instruction pointers to speculate different access patterns and cover a wide spectrum of access patterns. IPCP classifies instruction pointers into constant stride, complex stride, and stream. IPCP suggests different prefetching methods based on the type of instruction pointer.

In [18], the author introduces a reinforcement learning algorithm of prefetching to evaluate prefetch quality, pursuing system-aware prefetching. Observing the current memory bandwidth usage, Pythia intends to obtain highly accurate, timely prefetching by correlating program context information such as cache line address, program counter value, etc. to prefetch decision.

2.1 Useful and Useless prefetch

Usefulness of a prefetch is determined by whether a prefetched block is accessed by a program or not. In implementation, a prefetch bit of a block is employed to evaluate the usefulness of the prefetch. The prefetch bit is set when a prefetched block is inserted into cache memory. Then, it is unset when the block is accessed by the program. If the prefetch bit of a prefetched block is unset within the time that the block is evicted, corresponding prefetch will be regarded as a useful prefetch. In other words, the useful prefetch fetches a block that is accessed in the near future so the access occurs before the block is evicted. With the access, the useful prefetch results in eliminating a cache miss. On the contrary, if the prefetch bit remains set, corresponding prefetch will be considered as a useless prefetch. That is, useless prefetch fetches a block that is evicted with no access to it. So, the prefetch wastes cache space and memory bandwidth.

2.2 Temporal and Spatial Locality

Locality is that a program exhibits a tendency to reference the same data accessed recently or a data located closely to the recently accessed data. The principle of temporal locality is that recently accessed memory addresses by a program are likely to be accessed again in the near future.[21] The principle of spatial locality is that other nearby memory addresses have a likelihood of being referenced if a memory address is referenced. For example, sequential prefetching takes advantage of spatial locality; the simplest sequential prefetching scheme is to prefetch next cache block, one block lookahead of current access block[3].

2.3 Physical page contiguity

An address translation is an essential mechanism that maps a virtual address into a physical address to support virtual memory for modern processors. The translation is page-based operation so each virtual page corresponds to a physical page. The transfer between the two different address space can be a challenge to a hardware prefetcher since the translation can separate two contiguous virtual addresses into two distant addresses in physical address space. Hardware prefetcher has no knowledge of the translation since it is located at the side of a cache [4, 11, 22, 23]. So, prefetching should stop if prefetched data crosses over a page boundary of reference address.

Sequence: A, A+1, A+3, A+4, A+6, A+7, A+9, A+10, A+12, A+13, A+15, A+16, A+18, ...
Delta: +1 +2 +1 +2 +1 +2 +1 +2 +1 +2 +1 +2
(a) Original sequence and delta pattern

Sequence: A, A+1, A+3, A+4, (A+6), A+7, A+9, (A+10), A+12, A+13, (A+15), A+16, A+18, ...
Delta: +1 +2 +1 +3 +2 +3 +1 +3 +2
(b) Changed sequence and delta pattern

Delta Pattern Table	
Delta(prefetch key)	Delta Prediction
1	2
2	1

Figure 1: Effect of Prefetching on Miss Address Stream

2.4 Miss addresses and Prefetch hit

Miss addresses have been used for prefetching as a reference of memory access history. The miss addresses can be affected by prefetching; some of cache misses can be removed by useful prefetches. That is, correct prefetching based on miss addresses in the past changes the subsequent miss addresses, resulting in incorrect prefetching unless the prefetching reflects the change in the miss addresses.

Figure 1 is an example that shows the change in a miss address stream by prefetching. We ignore access latency for simplicity. We start with how to construct the Delta Pattern Table from the stream, and then explain about how a prediction from the table changes the miss address stream.

The original sequence shown in Figure 1a starts with a base address of A and shows a Delta pattern of 1 and 2, alternatively. The Delta Pattern Table shown in Figure 1c is constructed based on the Deltas observed in the sequence shown in Figure 1a. It is indexed by Delta (prefetch key) 1 and 2 accordingly. For each Delta, the Delta Prediction is entered. For example, from A to A+1, the table stores Delta 1. The next address is from A+1 to A+3 between

which the Delta is 2, so the table stores in entry Delta Prediction 2. In the sequence, Delta 1 is followed by Delta 2. For the table lookup, the entry of Delta with value 1 will find in the corresponding entry, Delta Prediction, a value 2. For example, address A+6 can be predicted from address A+4 by the Delta Pattern Table since the Delta of 1 observed from A+3 to A+4 is matched with the Delta (prefetch key) in the first entry in the table. So, its corresponding Delta Prediction is 2 so the table predicts address A+6 by adding 2 to address A+4. Then, the address A+6 is changed to be a hit address (the parenthesis with A+6 shown in Figure 1b indicates a hit address); the A+6 is no longer used as a reference. Therefore, the Delta, +2 (circled in Figure 1a), is not observed in Figure 1b. Due to the missing Delta, next incoming address A+7 cannot be predicted by the table. Furthermore, next incoming address A+9 cannot be predicted since the observed delta between A+4 and A+7 is 3 that does not match with any Delta (prefetch key) in the table. Likewise, correct prediction of address A+10 disturbs next two incoming addresses, A+12 and A+13, to be predicted by the Delta Pattern Table. In Figure 1b, underlined addresses indicates such addresses.

Reconstruction of the pattern table could be proposed as a solution to above issue due to the changed miss address stream. However, the update of the table cannot be effective in reflecting the miss address patterns since the original sequence of the memory access does not hold the new delta value, 3, in this example sequence. Therefore, it is important to keep prefetch hit addresses to identify patterns from miss address stream. Also, it is necessary to record which address has been prefetched by a prefetcher.

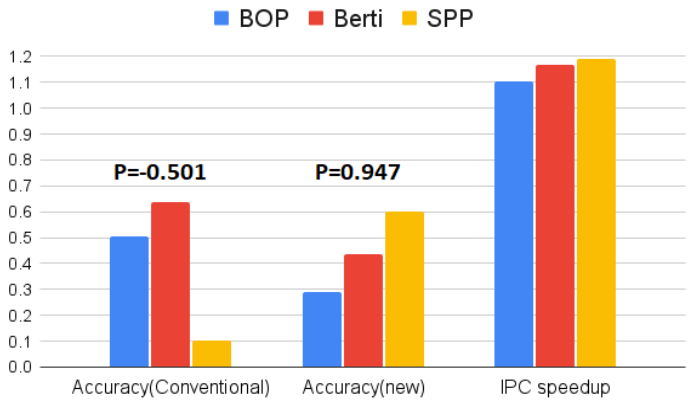


Figure 2: Pearson's Coefficient

2.5 Evaluation Methodology

There are two metrics to evaluate the performance of prefetching; prefetch accuracy and prefetch coverage. Prefetch accuracy measures a prefetcher in that how accurately the prefetcher predicts which memory addresses will be accessed in the future; on the other hand, prefetch coverage measures a prefetcher in that how diverse access patterns the prefetcher is capable of detecting against variations of the access patterns.

The accuracy and coverage has a relationship of inherent trade-off between them. For example, next line prefetcher can make a prediction for simple access pattern having a stride of 1 with a high accuracy but has a limited scope in coverage, meaning that the prefetcher cannot generate diverse predictions other than the stride of 1. On the other hand, if a prefetcher wants to achieve a wide scope in coverage, it should sacrifice its accuracy as a cost. This is, for a wide scope in coverage, a prefetcher should consider a wide change in simple to complex access patterns. So, regarding the broad variation on all the patterns would be more likely to lead a prefetcher to an inaccurate prediction. The imprecise prediction results in useless cache blocks to be fetched so that cache space can be wasted as well as cache pollution and memory bandwidth consumption can increase. Consequently, the inaccuracy of a prefetcher degrades the system performance.

We propose a refined metric for prefetch accuracy in order to make the accuracy metric more relevant to IPC speedup performance. Existing accuracy metric is the ratio of the number of useful prefetches to the total number of issued prefetches[12]. The numerator, the number of useful prefetches, in the metric is directly proportional to the IPC speedup performance since useful prefetches hides cache miss latency. However, the denominator, total number of issued prefetches, shows low correlation to IPC speedup performance. The total number of issued prefetches counts all the prefetches that are issued during only the "current" IPC speedup measure period. Actually, the IPC speedup measured in the current measure period can be affected by the prefetches issued in the "past"(before current) measure period. During the current measure period, the hit(or miss) of the prefetch issued in the past period increases(or decreases) IPC speedup. In other words, the total number of issued prefetches does not count the prefetches that were issued in the past period which, however, impacts the currently measured IPC speedup

Another reason of the low correlation between the total number of issued prefetches and IPC speedup is that the total number of issued prefetches contains the number of prefetches that turn out to be neither useful nor useless by the end of current IPC speedup measure period. They can be determined as either hit or miss "after" the current measure period. Thereby, in the current measure period, we do not have enough information about how those prefetches affects IPC speedup. However, in the conventional metric, such prefetches constitutes a portion of the total number of issued prefetches which is inversely proportional to the IPC speedup.

The low correlation between existing prefetch accuracy metric and IPC speedup is shown in Figure 2 with Pearson's correlation coefficient of -0.501. The value close to 0 means a low correlation while the value close to +1 or -1 indicates a linear correlation. One can see that in Figure 2 the prefetcher SPP has the lowest accuracy in the conventional metric but the IPC speedup of this prefetcher shows the highest IPC speedup, compared to other two prefetchers, BOP and Berti. We run the three prefetchers, BOP, Berti, and SPP, since they show almost the same value in the other metric, called prefetch coverage, in SPEC CPU 2006 benchmark to exclude the effect of prefetch coverage on the IPC speedup. We propose the modified metric, named prefetch accuracy ratio in this paper, as shown in Equation 1.

$$\text{Prefetch Accuracy Ratio} = \frac{\text{Number of Useful Prefetches}}{\text{Number of Useful Prefetches} + \text{Number of Useless Prefetches}} \quad (1)$$

Equation 1 evaluates the accuracy of a prefetcher based on the prefetches evaluated as useful and useless during "current" measure period only. The equation directly compares the useful prefetches and useless prefetches that are detected during the "current" measure time. Therefore, the metric, prefetch accuracy ratio, can show how much the useful prefetches contribute to the performance improvement over a performance degradation from useless prefetches. For example, a prefetch accuracy ratio of 1 means that a prefetcher gives only a positive effect to the system performance by generating only useful prefetches with no useless prefetches. As shown in Figure 2, Pearson's coefficient value of 0.947 shows a high correlation between the prefetch accuracy ratio and IPC speedup, so SPP represents the highest accuracy ratio, leading to the highest IPC speedup.

Another metric, prefetch coverage, is defined as the number of useful prefetches over the number of cache misses with no prefetching; prefetch coverage is proportional to the number of useful prefetches as shown in Equation 2:

$$\text{Prefetch coverage} = \frac{\text{Number of Useful Prefetches}}{\text{Total Number of Cache Misses}} \quad (2)$$

From Equation 1 and 2, we can conclude that the number of useful prefetch can be a key factor to achieve both high accuracy and wide coverage at the same time. The number of useful prefetches would not be proportional to the prefetch accuracy since increase in the number of useful prefetches involves increase in the number of useless prefetches in general. On the other hand, the number of useful prefetches is proportional to prefetch coverage since the change in the number of useful prefetches does not affect total number of cache misses.

Based on aforementioned metrics, we evaluate existing prefetchers and our proposed prefetcher in terms of IPC[7] speedup in section 3 and 4.

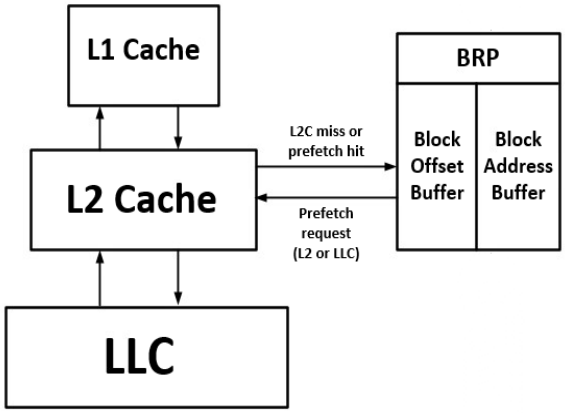


Figure 3: Overall BRP Structure[2]

The BRP module consists of two buffers as a storage; block offset buffer and block address buffer. The block offset buffer stores an offset of each block page-by-page. On the other hand, the block address buffer stores the address of blocks that are prefetched. The two buffers updates their data, according to FIFO manner; the oldest one is evicted first and the latest one is stored as a last item in the buffer, as avoiding stale data problem. In this section, we discuss how the two buffers are employed in reaching moderate accuracy ratio while maximizing prefetch coverage. Another feature of our BRP is that it does not request a prefetch in the case that it traverses a boundary of the page to which referred data belongs as discussed in Section 2.3.

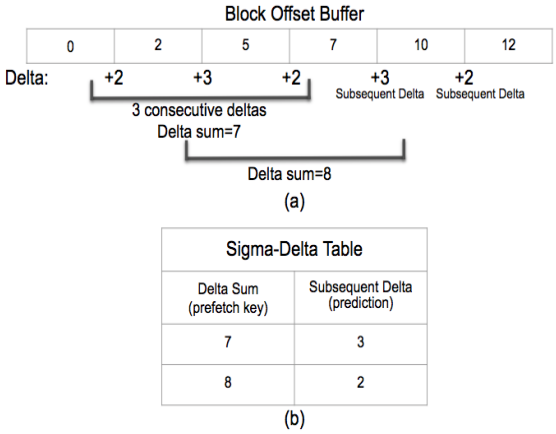


Figure 4: Example of Sigma-Delta Table generation (b) from block offset buffer (a) with delta history length of 3.

3 Design

Figure 3 illustrates the overall structure of our prefetcher, BRP. Both L2 cache miss and prefetch hit addresses trains our prefetcher. BRP sends a request to fill a prefetch into either L2 cache or last-level cache (LLC); LLC is requested to be filled by the prefetch when miss status holding register (MSHR)[8] has no opening because it is full of its items.

3.1 Offset Buffer

Offset buffer stores a block offset of pages that have been accessed recently. Based on the tag attached to the buffer, block offsets can be stored to a offset buffer with the same tag. That is, One of the offset buffers, of which the tag is the same as it of the page of the block accessed, collects the latest offset of the block. If a new page

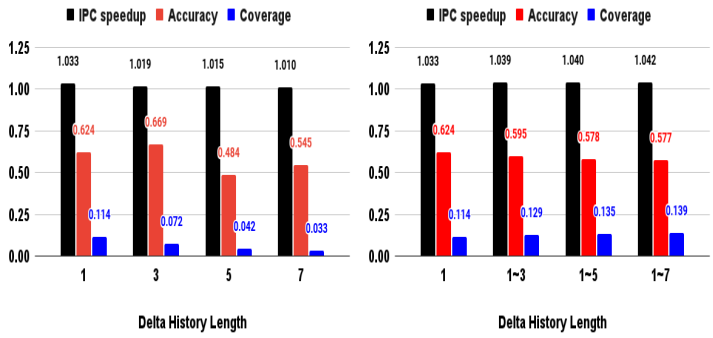
[7] All IPCs, instruction per cycle, are computed as arithmetic average of the IPC across the benchmarks
[8] MSHR holds pending load/store accesses that refer to the missing cache

is accessed for the first time, a new tag is generated based on a base address of the page, and a new offset buffer attached to the new tag is also built, loading the offset on it. The storage capacity of the buffer is set by users. If all the buffer is fully occupied by its offsets, the buffer evicts its oldest one and stores the latest offset as a last item in it.

The purpose of collecting block offsets is to observe the pattern of deltas among offsets; the delta is the arithmetic difference between the values of two block offsets that are successively accessed in the same page. [2] suggests a table, named Sigma-Delta Table, to identify the delta pattern from delta sequence as shown in Figure 4. Delta Sum, the sum of number of consecutive deltas, is used as an index; corresponding to it is the subsequent delta, the delta coming next.

Figure 4 represents an example of how a Sigma-Delta Table is generated from an offset buffer containing block offsets of 0, 2, 5, 7, 10 and 12. The example shows a case that three consecutive deltas are added up. The generated delta sequence is 2, 3, and 2 from the offset sequence of 0, 2, 5, and 7 as shown in Figure 4a, of which the sum is 7. The next Delta is 3 (to move from offset 7 to 10). So, the sum of 7 and the subsequent delta of 3 constructs a pair as an entry of the table shown in Figure 4b. In the table, in the case that Delta Sum 7 is given, from the same row and next column under Subsequent Delta, a value of 3 is read out.

In implementation, the number of consecutive deltas is set by users with regard to hardware overhead. In fact, the Sigma-Delta Table is built as a virtual table that does not require a permanent hardware overhead. That is, all the entries in the table is derived from offsets in the buffer only when they are referenced for prefetching. Hardware overhead of the table will be discussed in Sec. 4.



(a) Single Delta History Length (b) Multiple Delta History Length

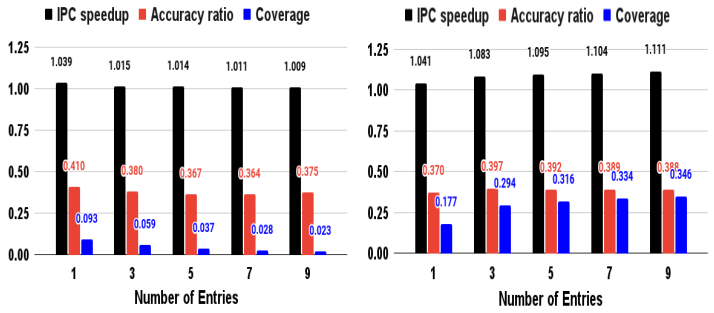
Figure 5: Comparison of delta history length (a) single (b) multiple

3.1.1 *Multiple delta history length*

As aforementioned, the Sigma-Delta Table is built by pairs of delta sum and subsequent delta which is calculated by deltas. The number of deltas used is equal to the delta history length. [2] suggests a method of using multiple delta history lengths rather than a single delta history length. So, the multiple Sigma-Delta Tables are built according to multiple delta history lengths. For example, if the maximum length of the delta history is set to be 4, four of the Sigma-Delta Tables are built; That is, each table belongs to different delta history length. Different number of consecutive deltas are used to construct the Sigma-Delta Table. All the generated sigma-delta pattern tables are subject to being used for prefetching.

The higher prefetch coverage is achieved by the proposed method as shown in Figure 5b, compared to the method of using single delta history length shown in Figure 5a. IPC speedup shown in Figure 5a gets decreased by the coverage reduction as a single delta history length increases. Compared to the decrease, in Figure 5b, IPC speedup is sustained as more multiple lengths of delta history gets involved. This is because the multiple delta history lengths cause an increase in prefetch coverage that cancels out the IPC speedup decrement caused by accuracy ratio reduction.

Briefly, multiple delta history lengths improves prefetch coverage by generating more useful prefetches, resulting in IPC speedup improvement. Several methods will be discussed to reach further increase in prefetch coverage in the section onward.



(a) Complete trained table only (b) (Complete+Incomplete) trained table

Figure 6: Completely and Incompletely Trained Table (a) complete only (b) complete+incomplete

3.1.2 *Complete/Incomplete training*

After a Sigma-Delta Table fills all of its entries, training session is started to evaluate the validity of the patterns that the table stores. Surely, we use a Sigma-Delta Table of which the training session is completely finished with its validity for prefetching. In addition, we propose to utilize the Sigma-Delta Table for prefetching while its training session is in progress in order to avoid no prefetch issued during the training session. Adopting the table with on-going training helps increase the scope of prefetch coverage. Figure 6b reflects the effect of the incompletely trained table additionally used, as compared to Figure 6a. Figure 6b exhibits a coverage increase by 2 to 15 times, compared to the coverage shown in Figure 6a, a result from using the completely trained table only. The increased coverage is due to the increase in the number of useful prefetches that a table under its training generates.

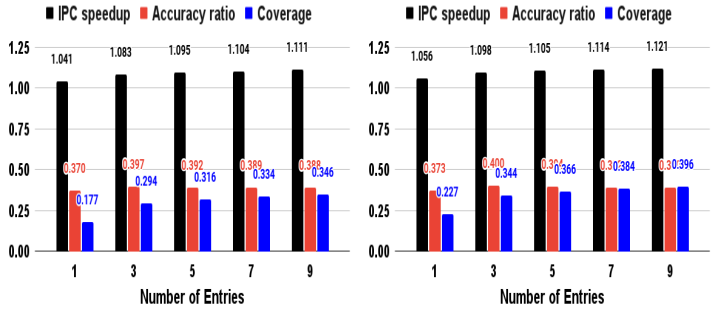
Furthermore, in Figure 6a, coverage decreases as the number of entries increases since the increase in the number of entries in the table requires longer training session. As a result, no prefetching duration increases, resulting in decrease in useful prefetches. However, Figure 6b shows the effectiveness of our proposed method

especially when more entries are employed. This is because the increased duration of training session for the case of more entries makes incompletely trained tables having longer time to generate more useful prefetches.

According to our design intention, the IPC speedup increases by 1 to 11% due to the increased coverage; IPC speedup is 1.04 to 1.11 in Figure 6b versus 1.03 to 1.01 in Figure 6a. The proposed method is more effective in the case of greater number of entries in a Sigma-Delta Table. One interesting observation in Figure 6b is that the additional use of the table continuing its training provides slight increase in accuracy ratio, as compared to Figure 6a, except the case of single entry of Sigma-Delta Table; 3 to 7% increase across 3 to 9 entries and 10% decrease in single entry. Another interesting observation from Figure 6b is that the case of 3 entries exhibits better results in both accuracy ratio and coverage than the results of single entry.

Training mechanism adopted in this work is to check the validity of collected delta patterns stored in the Sigma-Delta Table. The training is to compare values of entries in the Sigma-Delta Table with subsequently accessed block offsets that update a block offset buffer; to check if the pair of delta sum and delta prediction in the entry is equal to a pair of them calculated from the updated offsets in the offset buffer. If the equality happens, the table will get a hit, so called table hit. Otherwise, the table will get a table miss.

In implementation, we employ a flag to denote whether each Sigma-Delta Table obtains a table hit or not. On the other hand, a table miss score is counted with 2-bit saturating counter. If a table reaches a maximum table miss score, 3 with the 2-bit counter, all the entries in the table will be removed and new entries will, then, be generated by more recent offsets stored in the offset buffer. After training is finished, the table attaining a table hit and having lowest table miss score is used as top priority for prefetching.



(a) Sigma-Delta Table with no on-the-fly (b) Sigma-Delta Table with on-the-fly

Figure 7: Sigma-Delta Table usage (a) no on-the-fly (b) on-the-fly

3.1.3 *On-the-fly table*

We propose to use the Sigma-Delta Table on the fly to avoid no prefetching issued until the table is full of its entries. That is, we employ the table even under construction to generate a candidate of a prefetch to reach further increase in coverage. The on-the-fly table is effective in increasing coverage since it can generate a prefetch while the table either accumulates its entries or replaces its existing entries with new entries due to reaching a maximum table miss score. Also, the table can generate prefetches sooner based on the few patterns that belong to the new page accessed for the first time. According to our design intention, the IPC speedup increases by 1.3% due to the increased coverage; IPC speedup is 1.055 to 1.126 in Figure 7b versus 1.04 to 1.11 in Figure 7a.

3.1.4 *Referring to other pages*

Multiple Sigma-Delta Tables are generated to figure out the offset patterns for individual page. In [2], the author proposes referring to those tables so other near pages can reference the offset patterns to issue their prefetches. In other words, each page refers to the offset pattern, the pairs of Delta Sum and a Delta, in the table that has been accessed little earlier. So, the patterns detected from both currently accessed page and earlier accessed pages are utilized for current prefetch issue. Especially, this referring method should be effective in generating diverse prefetches when the page does not have enough identified patterns due to first time access to the page, early stage of building the Sigma-Delta Table, etc.

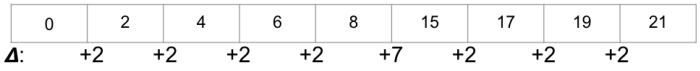


Figure 8: *libq*: Example of Spike Pattern

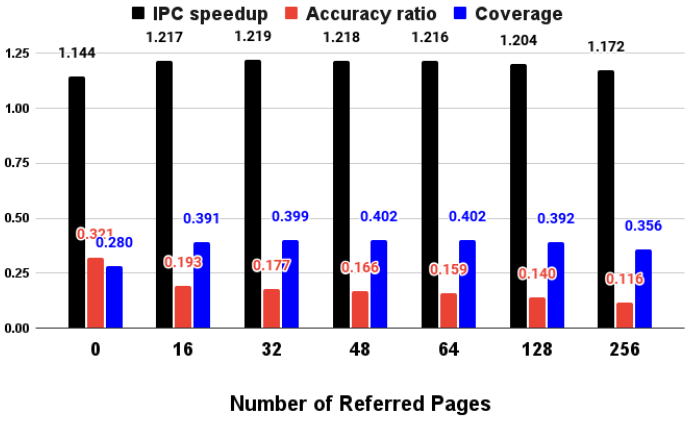


Figure 9: Number of Referred Pages Sensitivity

Furthermore, the referring mechanism gives a chance to mine performance improvement from irregular pattern. For example, spike pattern, the delta sequence with the single delta of +7, shown in Figure 8 from *leslie* benchmark is hard to be predicted due to its minority in patterns. Also, the delta spike can appear irregularly and its magnitude can be variable. Based on locality property among near pages, it is possible that preceding pages go through both the same timing and magnitude of spike pattern. So, we leverage access patterns of the preceding page in generating prefetches for subsequently accessed pages.

Obviously, there should be a cost of an initial miss in capturing the irregular pattern when the pattern shows up for the first time. With the sacrifice of the initial miss, the the pattern can be identified and referred by other near pages that may go through the same access pattern. This method is very effective to performance improvement for some benchmarks such as *calculix*, *lbm*, *leslie*, *libq*, *soplex*, etc. through simulation. In those benchmarks, some of the pages that are accessed sequentially tends to have the same offset patterns.

Compared to other prefetchers, BOP[11] is not successful in identifying the delta spike pattern since the prefetcher focuses only on globally optimal offset, using only delta value of +2 as a majority. GHB[9] could succeed in predicting the pattern as it captures two consecutive deltas as a pair, preserving the sequence of individual delta. By maintaining the delta sequence, GHB constructs pairs of two adjacent deltas such as (+2,+2), (+2,+7), and (+7,+2) that are used as a prefetch key. So, the pair, (+2,+2), is used to predict the spike delta, +7. However, in the case that the magnitude of next delta spike varies, GHB is hard to predict correct spike delta. Furthermore, GHB needs hardware storage in order to store the pairs of two deltas.

The performance improvement from referring to other pages is shown in Figure 9 with different number of referred pages. There is a 39% increase in prefetch coverage between no page referred and 16 pages referred. On the other hand, there is a 40% decrease in accuracy ratio between them. As a result, IPC speedup increases by 6%. On the other hand, coverage, accuracy ratio, and IPC speedup are almost the same among 16 to 128 referred pages. Also, referring to 256 pages degrades IPC speedup by 3.9% due to the decrease in prefetch coverage by 9.3%. Performance degradation occurs in the case that we refer many pages, which contains pages accessed in the distant past. With the observation, 32 pages are set to be referred.

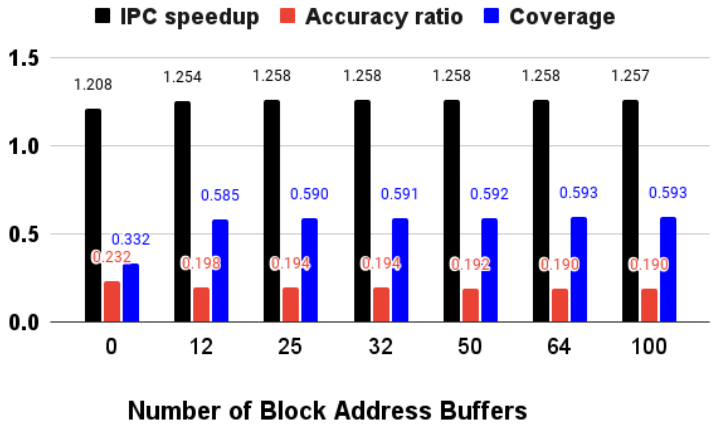


Figure 10: Block Address Buffer Sensitivity

3.2 Block Address Buffer

As discussed in Section 2.4, [2] proposes recording memory addresses that have been prefetch requested, including prefetch hit addresses. So, the access pattern can be preserved as the original pattern of the miss addresses. The paper offers a block address buffer to store the history of recent prefetch addresses. The record is used for multi-degree prefetching that indicates issuing multiple prefetchings per prefetch prediction. Additionally, the history helps filter out duplicate prefetches that are issued in the past.

As shown in Figure 10, there is an IPC speedup increase of 3.7% between no prefetch buffer and 12 prefetch buffers, because prefetch coverage increases by 76% while accuracy ratio decreases by 14%. On the other hand, compared to 12 buffers, 32 buffers shows 0.27% increase in IPC speedup. Also, there is no significant IPC speedup improvement with more than 32 buffers. The result represents that an optimal number of buffers is 32 since more than 32 buffers requires more hardware overhead with little return.

4 Performance Evaluation

We evaluate the performance of the prefetcher, BRP, with added techniques mentioned in the previous section. We compare it with other prefetchers' performance in terms of IPC speedup, accuracy ratio, and coverage, in this section.

4.1 Simulation Methodology

To evaluate BRP and other prefetchers, a subset of both the SPEC CPU 2006 benchmark[24] and 2017 benchmark[25] is used since the SPEC CPU 2006 benchmark is known as memory-intensive characteristics and SPEC CPU 2017 is developed recently. We use individual thread for a single core simulation and 4-thread mixes randomly selected from the benchmark for 4-core simulation. Simpoints [26], of which the interval is 10M instructions, are used to measure the IPC speedup. We simulate 200M instructions as a warmup and then measure the performance with the following 200M instructions. ChampSim simulator[27] is used for the evaluation. The framework of the Champsim simulator is the 3rd Data Prefetch Championship.

Table 1: Processor Configuration

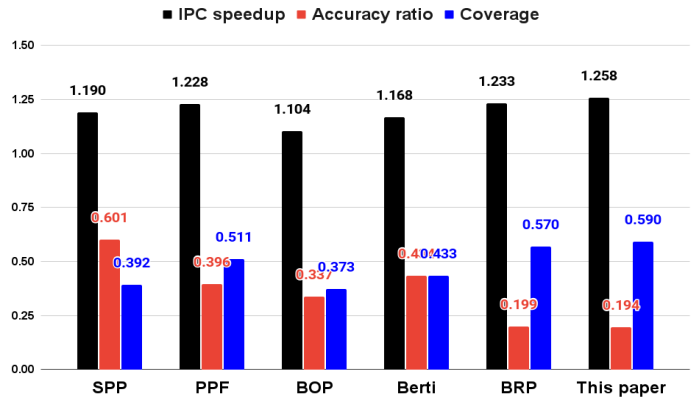
Core Parameters	1-4 Cores, OoO, 4GHz, 256 entry ROB, 4-wide
Branch Predictor	16K entry bimodal, 20 cycle misprediction penalty
Private L1 Dcache	32KB, 8-way, 8 MSHRs, LRU, 4 cycles
Private L2 Cache	256KB 8-way, 16 MSHRs, LRU 8 cycles, Non-inclusive
Shared L3 (LLC)	2MB/core, 16-way, LRU 12 cycles, Non-inclusive
Main Memory	4GB, 1-2 64 bits channels, 8 ranks/channel, 8 bank/rank, 1600MT/s

Table 1 summarizes the simulation configuration for the testing systems. The CPU is clocking at 4 GHz clock rates per core with an out-of-order scheduler. Each CPU core has its private cache; L1 and L2 cache. All the four cores shares a single L3 cache. The L1 cache is divided into instruction and data cache, of which size is 32 KB. L2 cache size is 256KB. The block size and page size of the cache is 64B and 4KB, respectively. In the single core simulation, single DRAM channel is adapted. In the 4-core simulation, two

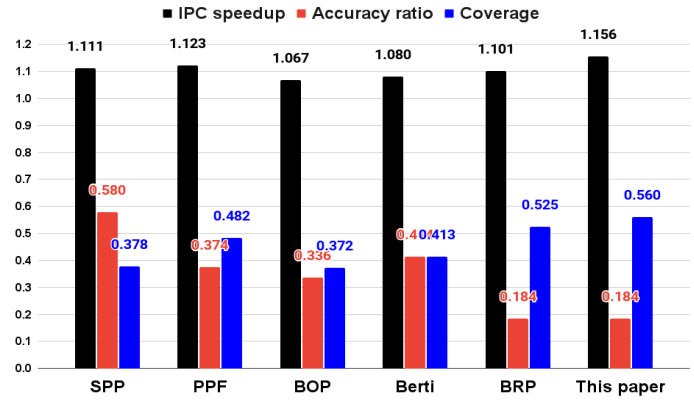
DRAM channels are adapted. Also, L1 and L2 cache contain 8 and 16 prefetch queues, respectively.

L1 and L2 cache have 8 and 16 miss MSHRs, respectively. The prefetch queue temporarily holds the block address which is requested for prefetching until it is issued. The MSHR and prefetch queue helps CPU continue to execute following instructions of a program since they store pending cache misses and prefetches. Only L2 cache access initiates the prefetcher and no prefetcher exists in other cache levels. All the prefetched data are stored into either L2 cache or L3 cache.

The performance of this work is compared with the origin work, BRP, as well as prefetchers such as SPP, PPF, BOP, and Berti, which are discussed in section 2. Those prefetchers are compared in terms of following three metrics; IPC speedup, prefetch accuracy ratio, and prefetch coverage. We use the original code of these prefetchers submitted to DPC-2 and DPC-3. We follows BRP's configuration which is set as follows: 1) Delta history length up to 3; length of 1, 2, and 3. 2) Maximum entries of 3 in sigma-delta table. 3) Block offset buffers for 512 pages. 4) Referring 32 pages. 5) 25 block address buffers.



(a) SPEC 2006: The impact of prefetch accuracy ratio and coverage on IPC speedup

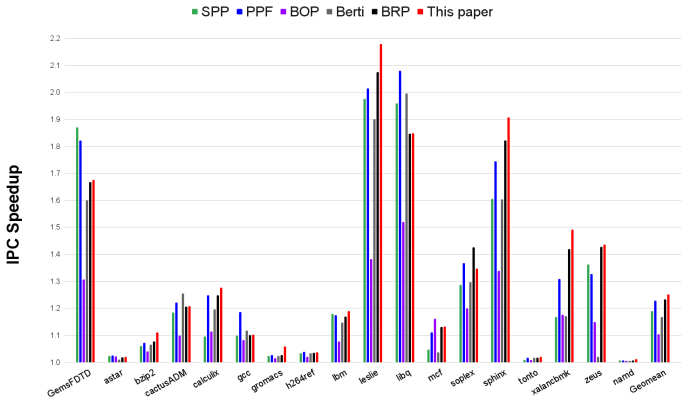


(b) SPEC 2017: The impact of prefetch accuracy ratio and coverage on IPC speedup

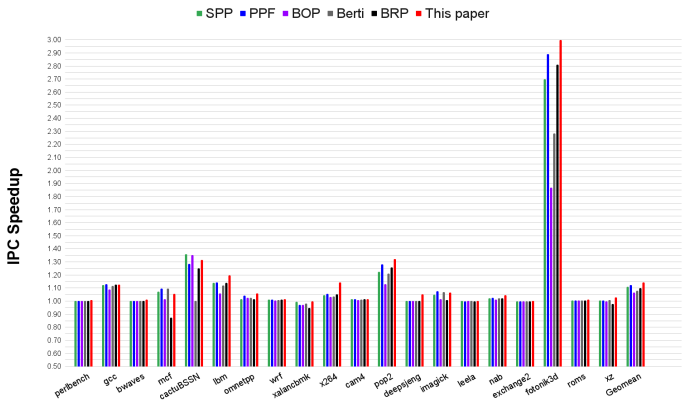
Figure 11: Single-core IPC speedup

4.2 Single Core Performance

Figure 11 shows the average IPC speedup of all the prefetchers with both SPEC 2006 and 2017 benchmark. The IPC speedup is a normalized IPC value to the IPC value of no prefetching baseline. Our work outperforms all the other prefetchers with the best average IPC speedup of 1.257 and 1.143 on SPEC CPU 2006 and 2017 benchmark, respectively, due to the widest prefetch coverage. As shown in Figure 11a, in SPEC CPU 2006 benchmark, our work exhibits 58.2%, 50.3%, 36.3%, 15.4%, and 3.5% improvement in coverage over BOP, SPP, Berti, PPF, and BRP, respectively, as our work shows the lowest accuracy ratio according to the design intention. So, our work achieves 25.7% IPC speedup improvement, which is 15.3%, 8.9%, 6.7%, 2.9%, and 2.5% more than the BOP, Berti, SPP, PPF, and BPR, respectively. Also, in SPEC CPU 2017 benchmark, as shown in Figure 11b, our work exhibits 50.5%, 47.9%, 35.6%, 16%, and 6.5% improvement in coverage over BOP, SPP, Berti, PPF, and BRP, respectively. So, our work achieves 14.3% IPC speedup improvement, which is 7.7%, 6.4%, 3.2%, 2.7%, and 4.5% more than the BOP, Berti, SPP, PPF, and BPR, respectively.



(a) SPEC 2006: IPC speedup by individual benchmark



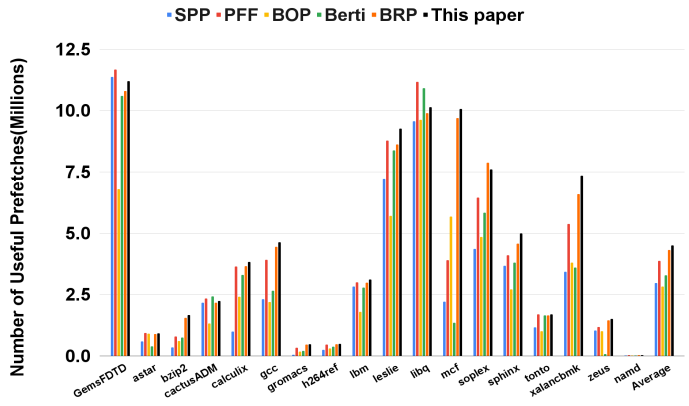
(b) SPEC 2017: IPC speedup by individual benchmark

Figure 12: IPC speedup by Benchmark

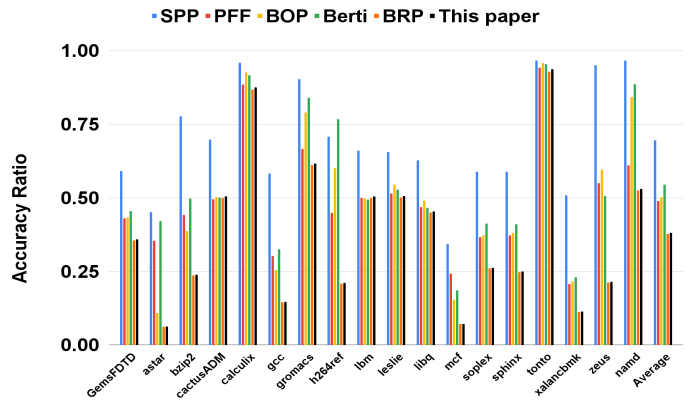
Figure 12 represents IPC speedup by individual benchmark. In SPEC CPU 2006(Fig. 12a), our work shows significant IPC improvement from benchmarks such as *GemsFDTD*, *leslie*, *libq*, *so-*

plex, *sphinx*, *xalancbmk*, and *zeus*, as compared to no prefetching baseline. This is because our work achieves high coverage from them by generating large number of useful prefetches as shown in Figure 13a. Especially, our work achieves the best IPC speedup in 10 applications such as *calculix*, *gromacs*, *leslie*, *soplex*, etc., compared to other prefetchers. As compared to the original work, BRP, our work shows better IPC speedup across nearly all the benchmarks. For example, our work exhibits 5%, 8%, and 7% IPC speedup improvement in *leslie*, *sphinx*, and *xalancbmk*, respectively.

In SPEC CPU 2017(Fig. 12b), our work shows outstanding IPC improvement from applications such as *cactuBSSN*, *pop2*, and *fotonik3d*, as compared to no prefetching baseline. This is because our work achieves high coverage from them by generating large number of useful prefetches as shown in Figure 14a. Especially, our work achieves the highest IPC speedup in 16 out of 20 benchmarks. As compared to BRP, our work also shows better IPC speedup across nearly all the benchmarks; especially, our work exhibits 20%, 8%, and 6% IPC speedup increase in *mcf*, *x264*, and *fotonik3d*, respectively.



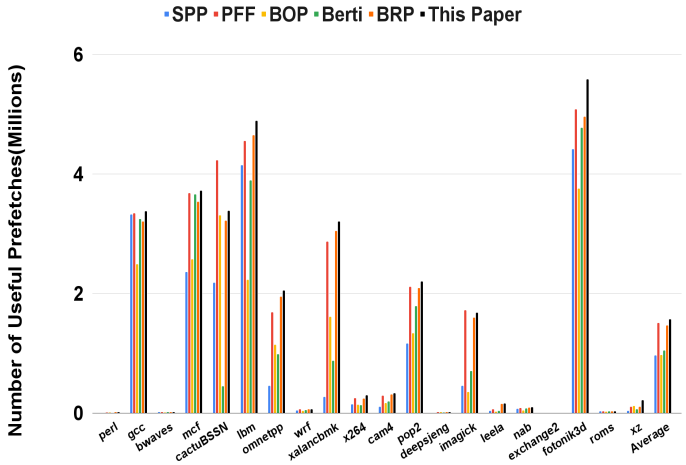
(a) SPEC CPU 2006: The number of useful prefetches by Individual Benchmark



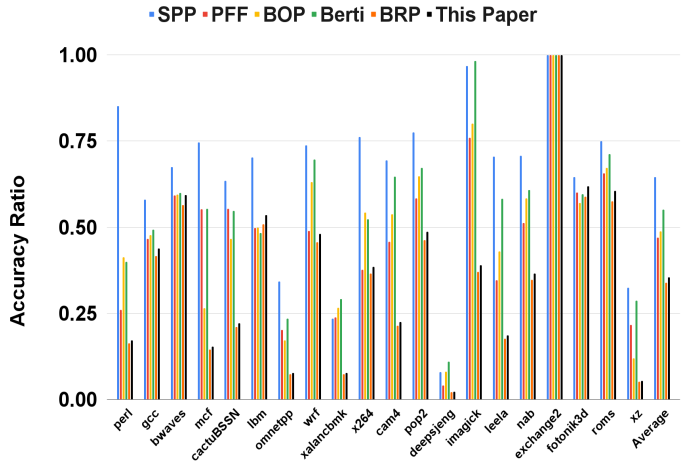
(b) SPEC CPU 2006: Accuracy Ratio by Individual Benchmark

Figure 13: Coverage and Accuracy Ratio with SPEC CPU 2006

As shown in SPEC CPU 2006(Fig. 13a), our work generates the largest amount of useful prefetches in the applications such as *bzip2*, *calculix*, *gromacs*, *lbm*, *leslie*, *sphinx*, *xalancbmk*, and *zeus* so that we have the highest IPC speedup. For those applications, high coverage is a key to predict their patterns, which would be diverse and irregular. On the other hand, in those applications such as *gcc* and *mcf*, our work generates the largest number of useful prefetches, which brings the highest coverage on them, but our work does not reach the highest IPC speedup. This is because that the comparable accuracy ratio would be also required to achieve better IPC speedup.



(a) SPEC CPU 2017: The number of useful prefetches by Individual Benchmark



(b) SPEC CPU 2017: Accuracy Ratio by Individual Benchmark

Figure 14: Coverage and Accuracy Ratio with SPEC CPU 2017

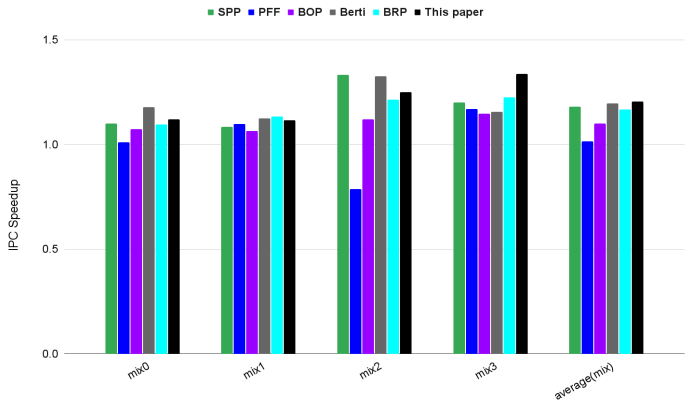
As shown in SPEC CPU 2017(Fig. 14a), our work generates the largest amount of useful prefetches in the applications such as *lbm*, *omnetpp*, *xalancbmk*, *x264*, *pop2*, *deepsjeng*, *nab*, *fotonik3d*, and *xz* so that we have the highest IPC speedup on them, accordingly. For those applications, high coverage is a key factor to achieve the high IPC speedup because they would have diverse and irregular patterns. On the contrary, in those benchmarks such as *mcf* and *camp4*, our work does not achieve the highest IPC speedup even though the highest coverage is reached with the largest number of useful prefetches. This is because the enough accuracy ratio would be also involved.

Table 2: SPEC CPU 2006: Multi Programmed Workloads

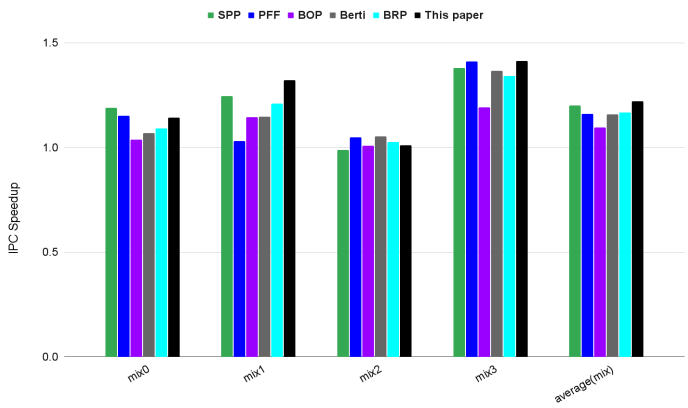
mix0	GemsFDTD, astar, bzip2, cactusADM
mix1	calculix, gcc, gromacs, h264ref
mix2	lbm, leslie3d, mcf, soplex
mix3	sphinx, tonto, xalancbmk, zeus

Table 3: SPEC CPU 2017: Multi Programmed Workloads

mix0	perlbench, gcc, bwaves, mcf
mix1	cactuBSSN, lbm, omnetpp, wrf
mix2	xalancbmk, x264, cam4, pop2
mix3	exchange2, fontonik3d, roms, xz



(a) SPEC CPU 2006



(b) SPEC CPU 2017

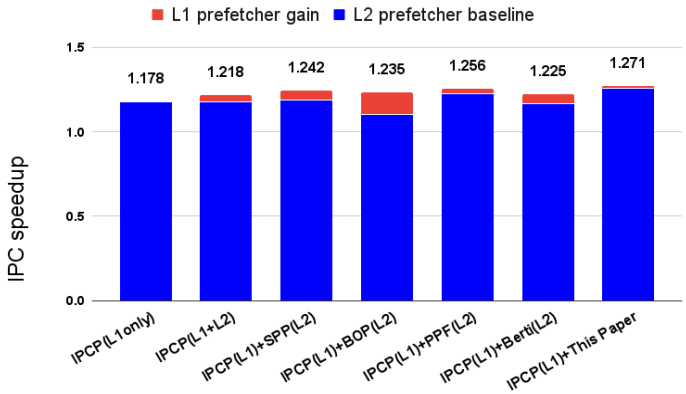
Figure 15: 4-core IPC speedup

4.3 Multi-Core Performance

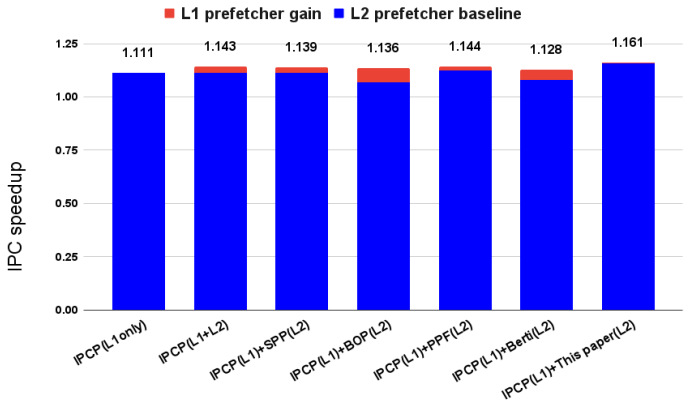
We generate 4 multi-programmed mixes, each consists of 4 traces as shown in the two tables; table 2 and table 3. Each trace in the mix is assigned to a different core for the multi-core simulation.

In SPEC CPU 2006(Fig. 15a), our work accomplishes 20.6% geometric mean IPC speedup increment across 4 mix workloads, compared to no prefetching baseline. Especially, our work exhibits 33.8% IPC speedup improvement in mix3 over the baseline, since our work is effective to appllications such as *sphinx*, *xalancbmk* and *zeus* in 4-core configuration as well. The performance improvement of our work is the best among all the prefetchers. our work surpasses the second highest one, Berti, by a 1.0% IPC speedup. Compared to single core test, in multi-core test, the IPC speedup improvement of our work over Berti reduces by 7.9% since the our work issues more prefetches than berti, throttling LLC and DRAM bandwidth.

In SPEC CPU 2017(Fig. 15b), our work accomplishes 22% geometric mean IPC speedup increase over no prefetching baseline across 4 mix workloads, which is the best performance improvement among all the prefetchers. Especially, our work shows 41.2% IPC speedup improvement in mix3 over the baseline, since our work is effective to appllications such as *fontonik3d* and *xxz* in 4-core configuration as well. Our work outperforms the second highest one, SPP, by a 2.1% IPC speedup. Compared to single core test, in multi-core test, the IPC speedup improvement of our work over SPP reduces by 1.2% since the aggressive prefetching of our work throttles LLC and DRAM bandwidth more in the multi-core configuration by generating both high number of useful and useless prefetches.



(a) SPEC CPU 2006



(b) SPEC CPU 2017

Figure 16: Single-Core: L1$ and L2$ prefetching IPC speedup

4.4 Multi-level prefetching performance

We added an L1 cache prefetcher to the test configuration to compare the performance of the prefetchers with IPCP which proposes multi-level prefetching, by an L1 prefetcher and an L2 prefetcher at the same time. We combine the IPCP L1 prefetcher with other proposed prefetchers acting as L2 prefetchers to conduct the test.

The L1 prefetcher of IPCP provides 17.8% and 11.1% IPC speedup in SPEC CPU 2006 and 2017, respectively, as compared to no prefetching. As shown in Figure 16a, our work shows the highest IPC speedup with an additional of 0.093% IPC speedup made by the L1 prefetcher in SPEC CPU 2006 benchmark; Our work outperforms the second highest, PPF, by 1.2% IPC speedup. In this test, BOP receives the maximum benefit of 0.131% IPC speedup, an increase from 1.104% to 1.235%, due to the IPCP L1 prefetcher. As shown in Figure 16b, in conjunction with the IPCP L1 prefetcher, our work also achieves the highest IPC speedup, surpassing the second highest, IPCP(L1+L2), by 0.018% in the IPC speedup. Please note, the L1 prefetcher gain here is only 0.004%, of the lowest amount.

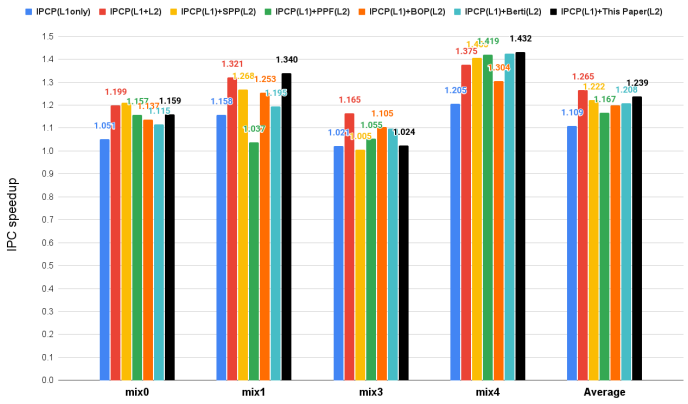


Figure 17: Multi-Core: L1$ and L2$ prefetching IPC speedup

In 4-core simulation, the two level IPCP prefetcher reached the hightest IPC speedup in SPEC CPU 2017, with 1.265 in average. As indicated in Figure 17, our work achieved the second highest performance in average, with an IPC speedup of 1.239, surpassed by the IPCP(L1+L2) only by 0.026%.

Table 4: Prefetcher Storage Overheads

Buffer-referred Prefetching(BRP)	5.512KB
Signature Path Prefetching(SPP)	5.507KB
Perceptron-Based Prefetch Filtering(PPF)	39.34KB
Best-Offset Prefetching(BOP)	1.85KB
Best-Request-Time Prefetcher(Berti)	22.1KB

4.5 Storage Overhead And Performance Contribution

The total hardware storage of our work is equal to the hardware overhead of the original work, BRP, of which the overhead is 5.512KB, with each individual component shown in Table 5. Our work does not increase the hardware overhead from the original work. There is no extra hardware overhead required with the introduced mechanism of employing incompletely trained Sigma-Delta Tables for generating prefetch candidates. Furthermore, we utilize the original Sigma-Delta Table under construction so no additional hardware is needed. The component with the largest overhead is block offset buffer (3.32KB). The buffer stores multiple 6-bit offsets per page. Berti holds heavy hardware overhead of 22.1KB, because it pursues an increase in both prefetch coverage and accuracy at the same time. Especially, the heavy overhead of Berti is caused by adapting diverse features such as instruction pointer, recorded page table, etc. in order to ensure prefetch accuracy. As shown in Table 5, BOP has the lowest overhead since its storage contains a few cache lines that has been requested recently. However, BOP derives low performance since the storage is employed to find out global best offset value as a simple manner.

Table 5: BRP Storage Overhead

Structure	Quantity	Component	Storage
Block Offset buffer	512	Offsets(6bit), Tag(16bit), lru(9bit), burst mode(2bit), aggressiveness(2bit),	33280 bits
Sigma-delta table accessories	1536	history length(2bit), valid(1bit), miss count(2bit), hit flag(1bit)	9216 bits
Block Address buffer	25	address(64bit)	1600 bits
Total = 33280 + 9216 + 1600 = 44096 bits = 5.512KB			

With the virtual Sigma-Delta Tables, nearly 45% of hardware storage is saved against using real Sigma-Delta Tables in total storage overhead. Hardware storage is allocated for storing block offsets in the buffers but then the table is virtually generated only when they are used, instead of giving permanent hardware storage for the real tables. The block offset buffers only except its accessories occupy 2.304KB; 512 buffers exist(each for a single page), each buffer has 6 block offsets, and each offset is 6-bit long. On the other hand, when the real Sigma-Delta Table is adapted with assigned constant storage, the total hardware overhead occupies 6.912KB; 3 tables exists per page (total 512 pages), there are 3 entries per table and each entry is of 12 bits (6 bits for delta sum and the other 6 bits for delta prediction). About 67% storage is reduced in terms of constructing pattern tables for prefetching, and about 45% hardware overhead decreases in terms of total hardware overhead.

5 Conclusion

In this paper, we have shown an extended work of Buffer-referred Prefetching, achieving high performance gain with low hardware overhead. First of all, it provides techniques that increase the prefetch coverage with no additional hardware overhead, compared to the original work. With such techniques, our work helps bring

out lots of useful prefetches from Sigma-Delta Tables under construction even though the table is not complete in its training. The technique is effective to the cases that few access patterns generated due to new page access, pattern transfer within the same page, and irregular patterns. This paper also offers a refined metric, called accuracy ratio, for measuring prefetch accuracy in order to directly take into account both performance improvement and degradation from a prefetcher. Future work would be to add a perceptive filter as a replacement of the block address buffer to achieve better prefetch filtering or well-tuned aggressive prefetching. Our work accomplishes 25.7% and 20.6% IPC speedup improvement over no prefetching baseline with single- and 4-core configuration, respectively, in SPEC CPU 2006 benchmark. In SPEC CPU 2017 benchmark, our work reaches 14.3% and 22.1% IPC speedup increase over no prefetching baseline with single- and 4-core configuration, respectively. Compared to the original work, we have 2.5% and 3.8% IPC speedup increment with single- and 4-core configuration, respectively, in SPEC CPU 2006. In SPEC CPU 2017, our work achieves 4.5% and 5.5% IPC speedup improvement with single- and 4-core configuration, respectively, over the original work. Our work outperforms the 2nd best prefetcher, PPF, by 2.9% and 2.7% IPC speedup with single- and 4-core configuration, respectively, in SPEC CPU 2006. In SPEC CPU 2017, our work surpasses both Berti by 1% and SPP by 2.1% IPC speedup with 4-core configuration in SPEC CPU 2017.

References

[1] W. A. Wulf and S. A. McKee. Hitting the memory wall: implications of the obvious. SIGARCH Comput. Archit. News, **23**(1),20–24, 1995.

[2] J. So and M. Lu, Buffer-referred Data Prefetching: An Effective Approach to Coverage-Driven Prefetching, 2021 26th International Conference on Automation and Computing (ICAC), 2021, 1-6, doi: 10.23919/ICAC50006.2021.9594139.

[3] A. J. Smith. Sequential program prefetching in memory hierarchies. In IEEE Transactions on Computers, **11**, 7–21, 1978.

[4] J. W. C. Fu, J. H. Patel,and B. L. Janssens. Stride directed prefetching in scalar processors. In Wen-mei W. Hwu, editor, Proceedings of the 25th Annual International Symposium on Microarchitecture, Portland, Oregon, USA, November 1992, pages 102–110. ACM / IEEE Computer Society, 1992.

[5] J.Baer and T. Chen. An effective on-chip preloading scheme to reduce data access penalty. In Joanne L. Martin, editor, Proceedings Supercomputing '91, Albuquerque, NM, USA, November 18-22, 1991, pages 176–186. ACM, 1991.

[6] F. Dahlgren and Per Stenstroom. Effectiveness of hardware-based stride and sequential prefetching in shared-memory multiprocessors. In Proceedings of the 1st IEEE Symposium on High-Performance Computer Architecture (HPCA 1995), Raleigh, North Carolina, USA, January 22-25, 1995, 68–77. IEEE Computer Society, 1995.

[7] D. Joseph and D. Grunwald. Prefetching using markov predictors. In Proceedings of the 24th Annual International Symposium on Computer Architecture, ISCA '97, pages 252–263, New York, NY, USA, 1997. ACM.

[8] P. Pathak, M. Sarwar, and S. Sohoni. Memory prefetching using adaptive stream detection. In Proceeding of 2nd Annual Conference on Theoretical and Applied Computer Science 5 November 2010.

[9] K. J. Nesbit and J. E. Smith.Data cache prefetching using a global history buffer. In In Proceedings of the International Symposium on High-Performance Computer Architecture, 2004.

[10] I. Hur and C.Lin. Memory prefetching using adaptive stream detection. In 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-39 2006), 9-13 December 2006, Orlando, Florida, USA, 397–408. IEEE Computer Society, 2006.

[11] P. Michaud. Best-offset hardware prefetching. In 2016 IEEE International Symposium on High Performance Computer Architecture (HPCA), pages 469–480, 2016.

[12] S. Srinath, O. Mutlu, H. Kim, and Y. N. Patt. Feedback directed prefetching: Improving the performance and bandwidth-efficiency of hardware prefetchers. In 13st International Conference on High-Performance Computer Architecture (HPCA-13 2007), 10-14 February 2007, Phoenix, Arizona, USA, pages 63–74. IEEE Computer Society, 2007.

[13] J. Kim, S. H. Pugsley, P. Gratz, A. L. N. Reddy, C. Wilkerson, and Z. Chishti. Path confidence based lookahead prefetching. In The 49th Annual IEEE/ACM International Symposium on Microarchitecture, 1-6, 2016.

[14] E. Bhatia, G. Chacon, S. H. Pugsley, E. Teran, P. Gratz, and D. A. Jimenez. Perceptron-based prefetch filtering. In in Proceedings of the 46th International Symposium on Computer Architecture, 2019.

[15] A. Ros. Berti: A per-page best-request-time delta prefetcher. In 3rd Data Prefetching Championship, 2019.

[16] M. Shakerinava, M. Bakhshalipour, P. Lotfi-Kamran,and H. Sarbazi-Azad. Multi-lookahead offset prefetching. The Third Data Prefetching Championship, 2019.

[17] S. Pakalapati and B. Panda, Bouquet of instruction pointers: Instruction pointer classifier-based spatial hardware prefetching, 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), 118–131, 2020

[18] R. Bera, K. Kanellopoulos, A. Nori, T. Shahroodi, S. Subramoney, and O. Mutlu, Pythia: A customizable hardware prefetching framework using online reinforcement learning, 54th Annual IEEE/ACM International Symposium on Microarchitecture, 1121–1137, 2021

[19] D.Yadav and C. Paikara. Arsenal of hardware prefetchers. Computing Research Repository, abs/1911.10349, 2019.

[20] S. H. Pugsley, Z. Chishti, C. Wilkerson, P. F. Chuang, R. L. Scott, A. Jaleel, S.-L. Lu, K. Chow, and R. Balasubramonian. Sandbox prefetching: safe run-time evaluation of aggressive prefetchers. In 2014 IEEE International Symposium on High Performance Computer Architecture (HPCA), 2014.

[21] S. Mittal. A survey of recent prefetching techniques for processor caches. ACM Comput. Surv., 49(2):35:1–35:35, 2016.

[22] S. Somogyi, T. F. Wenisch, M. Ferdman, and B. Falsafi. Spatial memory streaming. J. Instr. Level Parallelism, 13, 2011.

[23] M. Shevgoor, S. Koladiya, R. Balasubramonian, C. Wilkerson, S. H. Pugsley, and Z. Chishti. Efficiently prefetching complex address patterns. In Milos Prvulovic, editor, Proceedings of the 48th International Symposium on Microarchitecture, MICRO 2015, Waikiki, HI, USA, December 5-9, 2015, pages 141–152. ACM, 2015.

[24] SPEC CPU 2006.https://www.spec.org/cpu2006.

[25] SPEC CPU 2017.https://www.spec.org/cpu2017.

[26] T. Sherwood, E. Perelman, G. Hamerly, and B. Calder. Automatically characterizing large scale program behavior. In Kourosh Gharachorloo and David A. Wood, editors, Proceedings of the 10th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-X), San Jose, California, USA, October 5-9, 2002, pages 45–57, 2002.

[27] DPC-3.The champsim simulator. https://github.com/ChampSim/ChampSim.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 3, 129-138 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Indoor Position and Movement Direction Estimation System Using DNN on BLE Beacon RSSI Fingerprints

Kaito Echizenya, Kazuhiro Kondo*

Graduate School of Science and Engineering, Yamagata University, Yonezawa, Yamagata 9928510, Japan

ARTICLE INFO

Article history:
Received: 28 March, 2022
Accepted: 05 June, 2022
Online: 24 June, 2022

Keywords:
BLE beacons
RSSI fingerprints
Indoor positioning
Deep learning

ABSTRACT

In this paper, we propose a highly accurate indoor position and direction estimation system using a simple fully connected deep neural network (DNN) model on Bluetooth Low Energy (BLE) Received Signal Strength Indicators (RSSIs). Since the system's ultimate goal is to function as an indoor navigation system, the system estimates the indoor position simultaneously as the direction of movement using BLE RSSI fingerprints recorded indoors. To identify the direction of movement along with the position, we decided to use multiple time instances of RSSI measurements and fed them to a fully-connected DNN. The DNN is configured to output the direction with the location simultaneously. RSSIs are known to be affected by various fluctuating factors in the environment and thus tend to vary widely. To achieve stable positioning, we examine and compare the effects of temporal interpolation and extrapolation as preprocessing of multiple RSSI sequences on the accuracy of the estimated coordinates and direction. We will also examine the number of beacons and their placement patterns required for satisfactory estimation accuracy. These experiments show that the RSSI preprocessing method optimum for practical use is interpolation and that the number and placement of beacons to be installed does affect the estimation accuracy significantly. We showed that there is a minimum number of beacons required to cover the room in which to detect the location if the estimation error is to be minimized, in terms of both location and direction of movement. We were able to achieve location estimation with an estimation error of about 0.33 m, and a movement estimation error of about 10 degrees in our experimental setup, which proves the feasibility of our proposed system. We believe this level of accuracy is one of the highest, even among methods that use RSSI fingerprints.

1 Introduction

In recent years, almost everyone owns a smartphone. Some of the most often used applications on smartphones are those with location and navigation capabilities, such as Google maps. Global Positioning System (GPS) signals received from satellites are used for location detection in many mapping and navigation applications. However, these applications cannot be used indoors since the reception of GPS signals is generally poor due to the building structures. As an alternative to GPS, indoor positioning systems have been developed in recent years using various methods such as WiFi, Indoor Messaging System (IMES), and other wireless LAN standards, as well as dedicated devices such as Light Detection and Ranging (LiDAR) [1–6]. However, there are many issues to be solved for the construction of the system, such as the availability of terminals supporting these standards, and the installation cost.

Many of the indoor localization systems were based on WiFi. For example, [7, 8] use Deep Networks with WiFi RSSI fingerprints. Both use Autoencoders as the DNN. It is shown in [8] that an average error of 1.21 m is possible in an apartment of 14.5 m by 4.5 m, but the number of Access Points (APs) is 59, which is considerably large for a room this size. [9] uses Channel State Information (CSI) fingerprints instead of RSSI fingerprints with an Autoencoder. CSIs provide receive levels of multiple subcarrier signals which can be collectively used to detect the receive level and direction more accurately than a simple RSSI. They were able to achieve an average error of 0.9425 m inside a living room of size 4 m by 7 m with just one AP. However, CSI is not available on all network interface cards.

In this research, we focused on BLE beacons [10, 11]. BLE is one of the short-range wireless communication standards, and it is possible to conduct wireless communication at a lower cost and lower power consumption (often operating on battery for years without the need for replacement), and are easier to install and

*Corresponding Author: Kazuhiro Kondo, 4-3-16 Jonan, Yonezawa, Yamagata 9928510, Japan, kkondo@yz.yamagata-u.ac.jp

maintain compared to other methods, such as WiFi, which generally costs much more and requires a power line. In the real world, BLE technology is quickly spreading in various forms, such as devices for cash-less payment at cash registers and beacons that provide detailed product promotion information, including discount coupons, to nearby mobile devices. In [12], localization performance using BLE beacons is compared to WiFi, and was shown that localization performance using BLE potentially outperforms WiFi. Research using the BLE has been conducted using a variety of approaches [13–19]. Application of such systems to a university campus navigation system has been explored in [20].

The theoretical values of RSSIs received from beacons are found to decay inversely proportional to the square of the distance from the beacons. However, RSSI is known to fluctuate widely in real environments [21]. This is due to the effect of walls, obstacles, and radio interferences from other radio communication devices, among others. Therefore, the Euclidean distance calculation method, which calculates the current position directly from the magnitude of RSSI, generally gives an error of several meters or more. In this research, we propose a method that uses DNNs on the collected RSSI fingerprints [22–24]. This method requires the collection of RSSI data in advance in the environment where we want to perform positioning. It is known that RSSI variations caused by people and obstacles in the environment affect the estimation significantly. In this study, we use DNNs with RSSI fingerprints as training data, and we expect that DNNs can flexibly respond to RSSI variations in position estimation.

One of our target applications for this research is a navigation system that combines an indoor positioning system with augmented reality (AR). One of the intended applications of this system is an auditory navigation system for the visually impaired [25]. In this system, audio sign signals are localized towards the target for navigation. This signal is used as a cue to indicate the direction the user should move next. Localization of the auditory sign signal is achieved by convolving the head-related transfer function (HRTF) corresponding to the direction of the sign signal with the monaural signal. Since this direction is relative to the user's current orientation (the front), it is necessary to detect this orientation at the same time as the position. Therefore, the DNN used in this research estimates the direction of movement (the orientation) as well as the position simultaneously.

There has been some research on using DNNs with BLE RSSI fingerprints (*e.g.*, [26]). However, since our intended application also requires the estimation of the movement of the user, we decided to use multiple time instances of RSSI measurements and fed this to a DNN model to estimate the current position as well as the direction of the movement of the user simultaneously [22–24]. We believe the novelty of our method is this simultaneous estimation, where we employ multiple time instances of RSSI measurements to model the dynamically changing position due to users moving in a path. By explicitly training our DNN model with changing RSSI measurements due to motion, we can detect the direction of the movement from the temporal changes in the RSSI and also estimate the position more accurately when the RSSI measurements are changing due to motion, potentially degrading accuracy with conventional stationary modeling.

There have been attempts to emulate the WiFi RSSI movement data to be fed to an RNN model [1] to estimate the movement. However, their RSSIs are measured in a static state, which does not capture the dynamic change in the RSSIs due to motion. There also have been attempts to track the user position by using both BLE beacons and Pedestrian Dead Reckoning (PDR) [27]. However, the BLE beacons seem to be only used to correct the drift caused by PDR in this work.

This paper is organized as follows. Section 2 describes the proposed method for estimating indoor position and direction of movement using DNN. Section 3 describes the conditions of RSSI preprocessing and the experiments on the placement and number of beacons to be placed. In Section 4, the experimental results are described and discussed. In Section 5, we conclude our work and discuss future challenges.

2 Proposed Method

We propose an indoor position and movement direction estimation system using a DNN that has been previously trained with RSSI fingerprints of a given room. The RSSI fingerprints, *i.e.*, RSSI measurements from all available BLE beacons, are manually collected in the actual environment (test room) to train the DNN. As stated in the introduction, we plan to eventually integrate the proposed method into a navigation system that generates audio sign signals toward the target position. To do so, the direction in which the user is moving (and also facing towards) needs to be detected so that the sign signal can be localized relative to the detected direction. Thus, we decided to use multiple time instances of the RSSI measurements as RSSI fingerprints to detect the direction from the difference between the multiple temporal RSSI measurements. In this paper, we used two time instances, *i.e.*, one set of RSSI measurements at the current time interval, and one measurement from the prior time interval. The explicit integration of motion estimation with position estimation may be what is unique with this method.

We used the RSSI measurements reported in the advertisement packets from the beacons. These advertisement packets are designed to be sent periodically to all nodes in the proximity. However, we found that the time intervals of these packets are often not periodical. Some packets were also lost. Thus, it was necessary to preprocess the RSSI measurements to obtain data in periodic intervals at the exact time instances. In cases where the packets were lost, the RSSI data was estimated from data that was received intact before or after the lost packet. The preprocessing of the RSSI measurements will be described in detail in the following section.

The training configuration of the DNN is shown in Figure 1. The RSSI set for the current time and the prior time interval is fed to the DNN. The DNN estimates the current position and the direction of movement from the given RSSI sets. This is compared to the labeled (true) position and movement direction, and the mean square error (MSE) is calculated, which is fed to the DNN to update the internal weights. This is repeated to minimize the MSE.

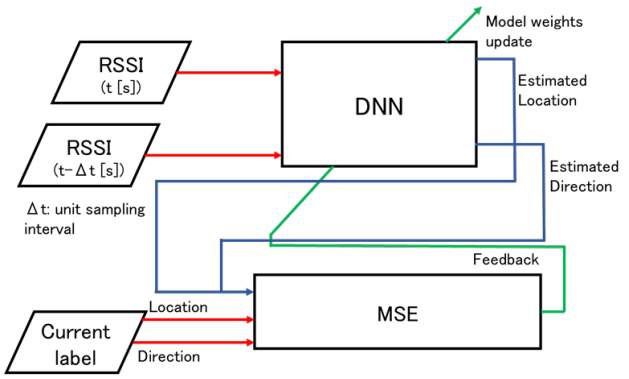


Figure 1: Block diagram of the proposed method. MSE is the Mean Square Error calculation between the estimated and the labeled position and movement direction values.

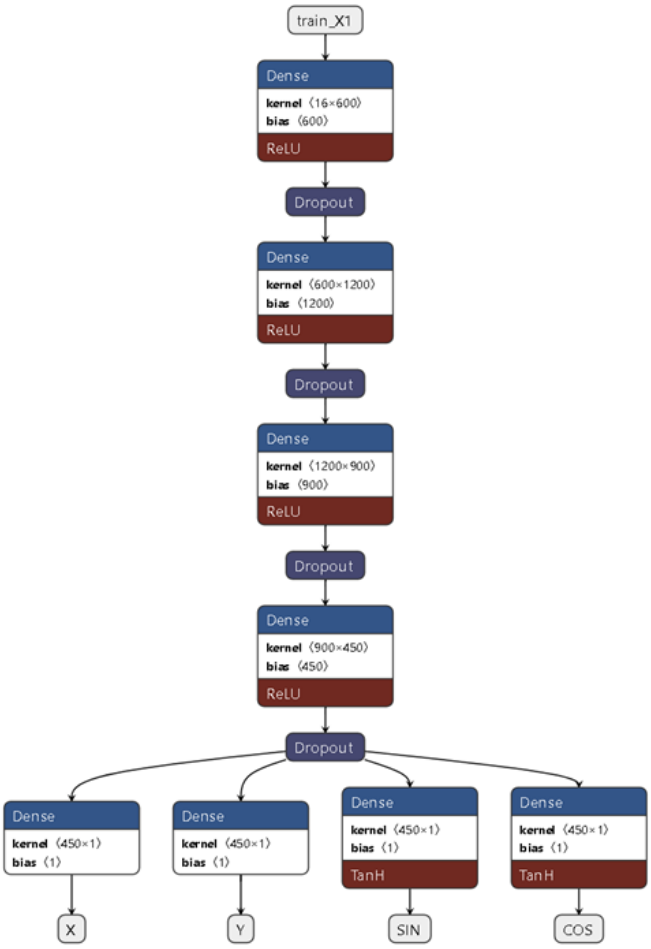


Figure 2: Configuration of the DNN for continuous values of x-y positions and movement directions.

To estimate the user's position, we feed the RSSI from all available BLE beacons for two consecutive time intervals to the trained DNN. The trained DNN will output the estimated position and direction of movement simultaneously from these inputs.

Table 1: Parameters of the DNN model.

Item		Conditions	
Units in each layer		16-600-1200-900-450-4	
Epochs		1000	
Batch size		256	
Dropout		0.5	
Activation	Hidden layers	ReLU	
	Output layers	X, Y	linear
		Dir. (cat.)	softmax
		Dir. (cont.) ($sin\theta$, $cos\theta$)	tanh

In this experiment, the RSSI was measured with no obstacles in the environment. However, since RSSI measurements collected from multiple data sequences are used as the training data of the DNN model, it is expected that estimation with some number of obstacles can be accomplished with a relatively small increase in error. The parameters of the DNN are shown in Table 1. The structure of DNN when the direction of movement is treated as a continuous value is shown in Figure 2.

As stated before, DNNs are used to simultaneously estimate the position and the direction of movement. The units in the input layer are given the RSSI of multiple beacons for two consecutive point-in-time instances in the training data. In the output layer, two units are used for the x- and y-coordinates of the estimated position and additional units are used for the estimation of the movement direction.

We modeled the motion direction as either categorical or continuous. The categorical model outputs one of eight directions (0, 45, 90, 135, 180, 225 270, or 315 °) as the estimate as one-hot output. Note that the categorical estimation values result in an estimation error, if any, of at least 45° [22], however.

On the other hand, the continuous output processes the movement directions as continuous values. The advantage of this is that the numerical difference (error) between the estimated and true values can also be expressed as continuous values. For example, in the experiment in [22], when the direction of movement was treated as a continuous value instead of a categorical value, the results showed that the frequency of fatal errors in the estimation of the direction of movement (*e.g.*, estimates that are the exact opposite of the correct value) was greatly reduced. The estimation of the angle θ, the direction of movement, is calculated according to (1) using $\sin\theta_{estimate}$ and $\cos\theta_{estimate}$ estimated as continuous values, and where $\theta_{estimate}$ is the estimated angle of movement.

$$\theta = \tan^{-1}\frac{\sin\theta_{estimate}}{\cos\theta_{estimate}} \quad (1)$$

We decided to use $\sin\theta$ and $\cos\theta$ to express the movement direction as output from the DNN. This is because these values are continuous, ranging from -1 to $+1$. If they were expressed in raw degrees, we would have to deal with discontinuities between 0 and 360° (if expressed as positive values) or $-180°$ and $+180°$ (if expressed using positive values for the right half, and negative values for the left half). The output values $\sin\theta$ and $\cos\theta$ are converted to movement direction by applying $\tan^{-1}$ to the ratio between $\sin\theta$ and $\cos\theta$ ratios (*i.e.*, $\tan\theta$). We would have to deal with discontinuities at $+90°$ and $-90°$ for this conversion, but this can be done relatively

easily by observing the positive/negative combinations of $\sin\theta$ and $\cos\theta$ values.

For the training data of the DNN, RSSI measurements were conducted when the receiver moved at a constant walking speed in one of eight predetermined directions (described in Section 3). These data were labeled with the correct position and direction. For training, 11,336 sets were used, of which 20% were used as validation data. For the test data, we used 2,434 sets.

The configuration of our proposed indoor position and movement direction estimation system is shown in Figure 3. This system is a client-server type system. First, the user's smartphone collects the RSSIs transmitted from multiple beacons at regular intervals. Next, the smartphone sends the collected RSSIs to the PC server via UDP packets. The server processes the received RSSIs, estimates the position and direction of movement using a DNN model that has been trained in advance, and sends the estimated position and direction back to the smartphone using UDP packets. By repeating the above operation, continuous estimation becomes possible. By providing DNN models that match the environment on the server-side, the computational load on the client-side can be reduced.

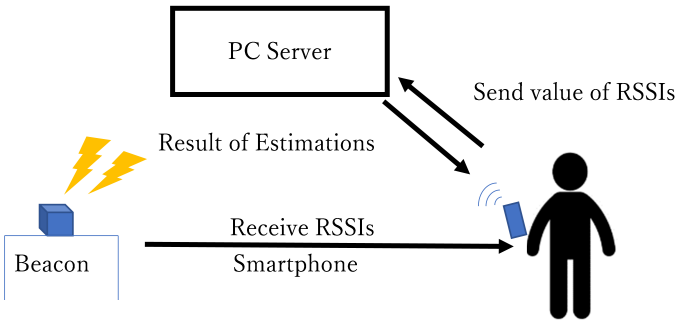


Figure 3: Proposed client-server system configuration.

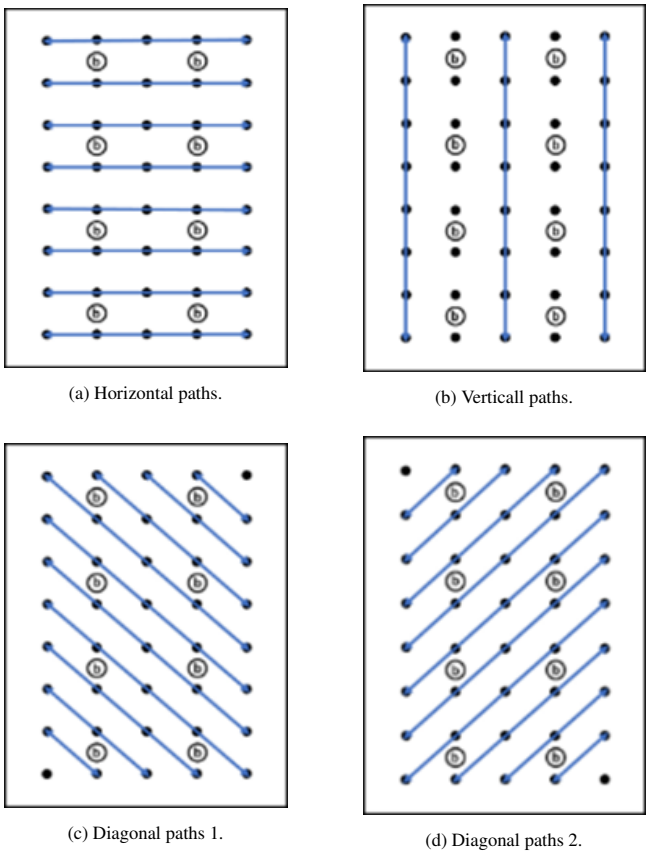
3 Experimental Conditions

We conducted two evaluation experiments explained here. The first experiment (A) was designed to evaluate the accuracy achievable using a fixed beacon position configuration and a fixed number of beacons (eight, which is the maximum number to be considered in this paper). The goal of this experiment is to compare the effect of preprocessing on the RSSIs of the estimation accuracy. The second experiment (B) was designed to evaluate the effect of the number of beacons and their positions on the estimation accuracy. The goal was to find the minimum number of beacons and their positions required for acceptable estimation accuracy.

3.1 *Experiment A: Comparison of estimation accuracy among RSSI processing methods*

The experimental environment is shown in Figure 4. As shown in this figure, eight beacons, shown here as ⓑ, were placed on chairs at regular intervals in a rectangular room (5 m x 8 m). A total of 40 grid points (5 x 8) were set at 1 m intervals. For this environment, we assumed a total of 31 vertical, horizontal, and diagonal linear paths in total. We collected data for each path by moving at a constant speed of 0.5 m/s repeatedly from the edge to the other edge of each path. We used a smartphone with BLE reception capability, SONY Xperia5 (SO-01M), as the receiving device, and Mybeacon-Pro (MB004) from Aplix Ltd. as the BLE beacons. We measured the RSSI data every 0.25 seconds while moving on each path. To acquire the RSSI data at the exact instance, temporal interpolation and extrapolation of the RSSI sequence data are considered in this experiment.

Interpolation is a method to calculate the RSSI on the estimated coordinates using the RSSI measured before and after passing through the estimated grid points, while extrapolation is a method to calculate the RSSI on the estimated grid points using the RSSI acquired before passing through the estimated grid points. When collecting RSSI data in a real environment for experiments, we want to acquire RSSI precisely on the grid point positions we want to estimate (the grid points shown in Figure 4), but this may be difficult due to discrepancies in the RSSI measurement timing. In addition, in rare cases, RSSI may not be collected from some beacons (due to internal processing delays or packet losses), and the collected RSSI sequence data may be collected at unequal time intervals. Therefore, we use interpolation and extrapolation to create RSSI data that are equally spaced in time.



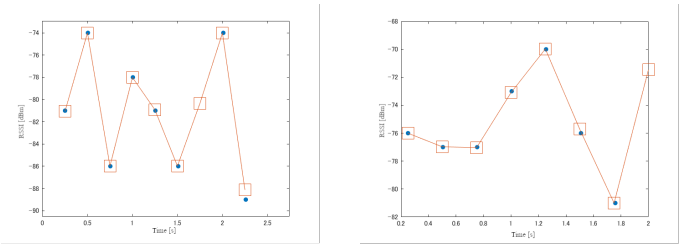
(a) Horizontal paths.

(b) Verticall paths.

(c) Diagonal paths 1.

(d) Diagonal paths 2.

Figure 4: Beacon placement and path configuration for experiment A. Beacons are shown as ⓑ, and the dots indicate the grid points at which the position and the motion direction is evaluated.



(a) An example of interpolation. (b) An example of extrapolation.

Figure 5: Examples of interpolation and extrapolation of RSSI data. The blue dots indicate the measured RSSIs, and the squares indicate the smoothed data used in the later estimation stages using DNNs.

Figure 5 shows an example of interpolation and extrapolation of RSSI data. In this figure, the collected RSSI data is represented by dots and the interpolated and extrapolated data by squares. In this experiment, RSSI data was collected while walking at a constant speed of 0.5 m/s. The subjects are expected to pass over the grid points in Figure 4 every 2 seconds. The RSSI to be acquired at exact timing is acquired from a data sequence that is supplemented at intervals of 0.25 s using interpolation or extrapolation. The interpolation is based on the piecewise cubic Hermitian interpolation polynomial (pchip) [28, 29], and the extrapolation is based on cubic spline interpolation [30, 31].

In the example in Figure 5 (a), the sample at 1.75 s was lost and was interpolated from samples before and after this interval. In Figure 5 (b), the sample at 2 s was not available and was extrapolated from preceding samples. Up to seven preceding samples and one succeeding sample in time are used to interpolate one coordinate in the collected data. Extrapolation uses up to 8 past samples. Since RSSI values in the real environment vary widely, it is not possible to process accurate RSSI data if the number of data used for interpolation and extrapolation is small. Thus, based on preliminary experiments, we decided to use the above-mentioned preceding and succeeding samples for smoothing as a compromise between accuracy and real-time constraints. We will also compare the accuracy without RSSI processing (no interpolation or extrapolation).

We will also examine whether categorization or continuous angular values are appropriate for estimating the direction of movement. The output layer of DNN is different depending on how to handle the movement direction. If the movement direction is output as a categorical value, the output layer will be 10 (x-, y-coordinates, and 8 additional outputs for one of 8 directional categories, *i.e.* $8 \times 45°$, as one-hot output). If the movement direction is output as a continuous value, the output layer of DNN will be 4 (x-, y-coordinates, and $\sin\theta$ and $\cos\theta$ values).

To compare the accuracy of the position estimation, we created a heat map showing the average estimation error for all coordinates and the average estimation error for each coordinate with each RSSI processing method. The average position estimation error, d, is calculated as in (2).

$$d = \sum_{i=1}^{N} \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2}}{N} [m] \tag{2}$$

Here, N is the number of estimations, and x and y are the true coordinates, and x_i and y_i are estimated x and y positions. As stated before, to treat the movement direction as a continuous value in the DNN, the output of the DNN is set to be $\sin\theta$ and $\cos\theta$. Therefore, the error in estimating the direction of movement, θ_{error}, is calculated as shown in (3) and (4).

$$\theta_{estimate} = \tan^{-1} \frac{\sin\theta_{estimate}}{\cos\theta_{estimate}} [°] \tag{3}$$

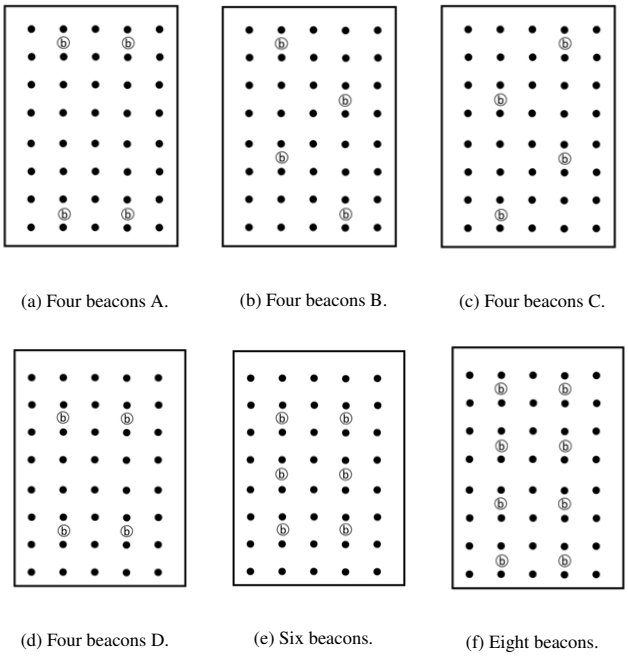
$$\theta_{error} = \frac{1}{N} \sum_{i=1}^{N} |\theta_{true} - \theta_{estimate}| [°] \tag{4}$$

Here, θ_{true} is the true direction of movement, and $\sin\theta_{estimate}$ and $\cos\theta_{estimate}$ are estimated sin and cos values output from the DNN.

3.2 *Experiment B: Comparison of estimation accuracy by placement and number of installed beacons*

The number and placement pattern of the installed beacons were set as shown in Figure 6 in this experiment to investigate the effect of beacon placement and the number of beacons.

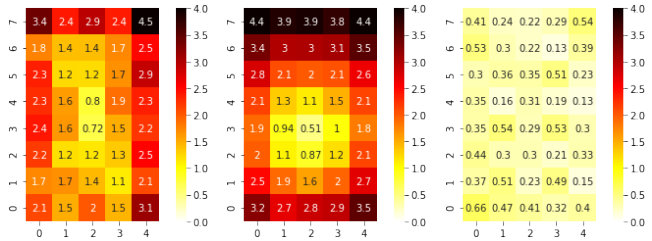
Beacons were placed on the chair at regular intervals as shown in Figure 6, and a total of 40 grid points (5×8) were set at 1 m intervals. As in Experiment A, a total of 31 vertical, horizontal and diagonal linear paths were assumed, and data were collected by moving at a constant speed of 0.5 m/s repeatedly from the edge of each path. The collected RSSI data were interpolated to obtain the exact RSSI measured at the estimated grid points. All movement directions were treated as continuous values. Figures 6 (a) to (d) show the configuration of the beacon placement to compare estimation accuracy for four patterns using four beacons. In Figures 6 (d) to (f), the number of beacons to be placed is varied to compare the estimation accuracy. The calculation of the estimation accuracy is done in the same way as in Experiment A.



(a) Four beacons A. (b) Four beacons B. (c) Four beacons C.

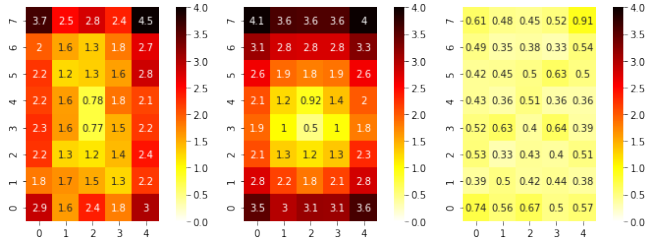
(d) Four beacons D. (e) Six beacons. (f) Eight beacons.

Figure 6: Beacon placement for experiment B. Beacons are shown as ⓑ, and the dots indicate the grid points at which the position and the motion direction is evaluated.



(a) No processing. (b) Extrapolation. (c) Interpolation.

Figure 7: Position estimation error using movement direction with **continuous** values.



(a) No processing. (b) Extrapolation. (c) Interpolation.

Figure 8: Position estimation error using movement direction with **categorical** values.

Table 3: Average direction of movement estimation error with **categorical** direction values.

Direction	Average estimation error [°]		
	No processing	Extrapolation	Interpolation
All	112.60	85.24	**10.61**
L	57.09	38.44	**10.99**
R	45.15	100.83	**12.78**
U	164.79	90.36	**10.52**
D	136.71	80.18	**11.10**
UL	135.48	65.16	**9.34**
LL	119.57	53.65	**12.96**
UR	153.00	98.91	**3.13**
LR	125.68	96.35	**15.28**

Table 4: Direction estimation error distribution by angle error range.

Error range	Occurrences	
	Continuous	Categorical
$157.5 \leq \theta \leq 180.0$	13	9
$112.5 \leq \theta \leq 157.5$	18	25
$67.5 \leq \theta \leq 112.5$	78	102
$22.5 \leq \theta \leq 67.5$	314	269
$0.0 \leq \theta \leq 22.5$	2011	2029

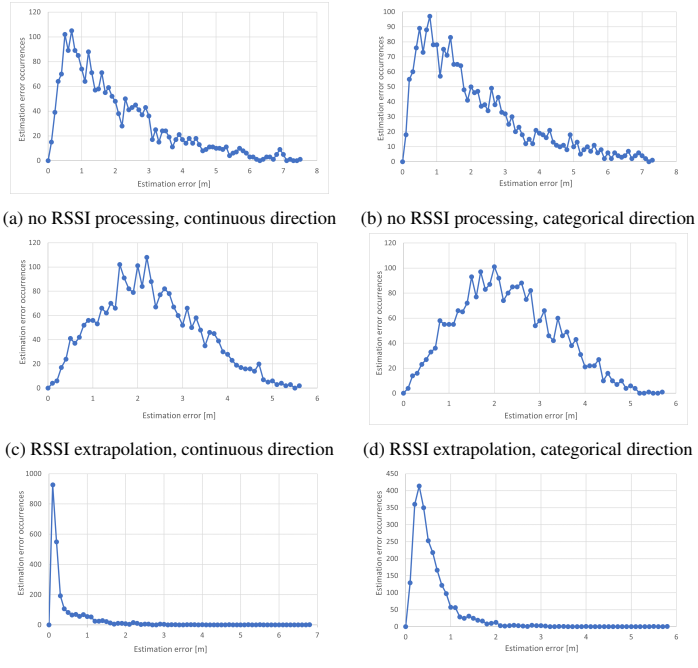
4 Results and Discussions

4.1 Experiment A

The experimental results are shown in Figure 7 through 10 and Tables 2, 3, and 4, respectively.

Table 2: Average direction of movement estimation error with **continuous** direction values. The path direction notations are L:horizontal right to left, R: horizontal left to right, U: vertical bottom to far top, D: vertical top to near bottom, UL: diagonal lower right to upper left, LL: diagonal upper right to lower left, UR: diagonal lower left to upper right, LR: diagonal upper left to lower right, and All: average over all directions.

Direction	Average estimation error [°]		
	No processing	Extrapolation	Interpolation
All	115.96	78.28	**12.96**
L	59.02	73.76	**10.75**
R	57.48	76.02	**14.72**
U	159.49	109.44	**14.38**
D	143.86	73.51	**11.12**
UL	131.55	89.99	**10.27**
LL	130.26	53.65	**15.44**
UR	156.59	90.53	**7.14**
LR	123.78	67.81	**17.68**

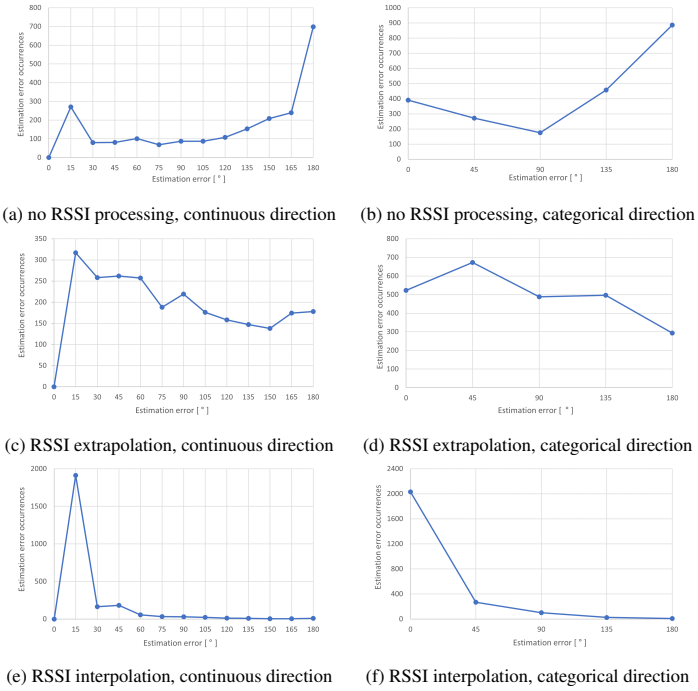


(a) no RSSI processing, continuous direction (b) no RSSI processing, categorical direction

(c) RSSI extrapolation, continuous direction (d) RSSI extrapolation, categorical direction

(e) RSSI interpolation, continuous direction (f) RSSI interpolation, categorical direction

Figure 9: Position estimation error distribution by RSSI processing method and angular expression.



(a) no RSSI processing, continuous direction
(b) no RSSI processing, categorical direction
(c) RSSI extrapolation, continuous direction
(d) RSSI extrapolation, categorical direction
(e) RSSI interpolation, continuous direction
(f) RSSI interpolation, categorical direction

Figure 10: Direction estimation error distribution by RSSI processing method and angular expression.

Figures 7 and 8 show the position estimation results when the direction of movement is treated as a continuous value and when it is treated as a categorical value, respectively. The average error at each grid position is color-coded as shown in the color bars. Figs. (a), (b), and (c) in both figures show the average estimation error at each grid position when RSSI is not processed, extrapolated, and interpolated, respectively. Note that we trained and tested the DNN with the same RSSI sequence handling, *i.e.*, applying no processing on the RSSI data in both training and testing to get (a), applying extrapolation to both the training and test data for (b), and interpolation to both the training and test data for (c), respectively.

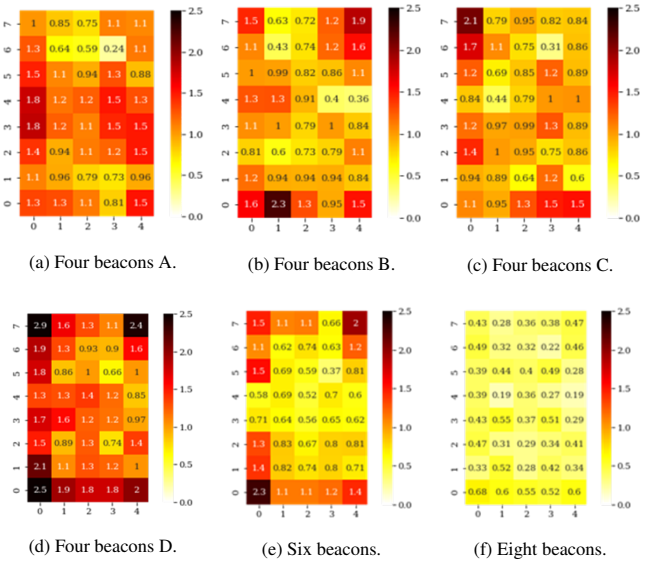
The average position estimation errors for the three methods in Figure 7 were 1.777 m, 2.069 m, and 0.330 m, respectively. The average estimation errors of the three methods in Figure 8 were 1.812 m, 2.061 m, and 0.469 m, respectively. These results show that interpolation can provide much higher position estimation accuracy. On the other hand, the difference in the treatment of the direction of movement (categorical vs. continuous) does not seem to have a significant impact on the accuracy of position estimation.

Tables 2 and 3 show the direction estimation errors when the direction of movement is treated as a continuous value and when it is treated as a categorical value, respectively. Table 4 also compares the error range distribution of both continuous and categorical direction. From Table 2 and 3, we can see that interpolation is still effective in estimating the direction of movement. However, depending on how the direction of movement is treated, the magnitude of the error remains the same for both categorical and continuous values. This is also evident in Table 4, where we can see that both methods generate approximately the same number of instances with errors in each range. It should be added, however, that the advantage of using continuous values is that if the number of categories of directions to be estimated needs to be changed, a new labeling and learning process is required, whereas this is not necessary when using continuous values.

Figures 9 and 10 show the distribution of the position estimation error and the direction estimation error for each combination of RSSI preprocessing and moving direction handling, respectively. These correspond to conditions included in Figures 7 and 8 and also Tables 2 and 3. In these figures, the occurrences of the position estimation errors are plotted for every 0.1 m interval. The error in the direction of movement is plotted in 15° increments for continuous values, and in 45° increments for categorical values. The rather large increment for categorization was because we designed the categorization in these increments, while the continuous value treatment allows arbitrary intervals, for which we arbitrarily used bins in 15° increments for error analysis. From the results, we can see that the interpolation for both position and direction of movement estimation is stable with little error overall. In particular, when the direction of movement is treated as a continuous value and the RSSI values are interpolated, the position estimation error peaks in the range of 0.0 to 0.2 m, as shown in Fig. 9e. For the direction of movement, a peak can be seen in the range of 0 to 15°, as shown in Fig. 10e . This tendency is not seen in the extrapolation or unprocessed cases, indicating the effectiveness of interpolation. The large average position estimation error with extrapolation compared to unprocessed RSSI is probably due to the inaccuracy of RSSI calculation based on past data and the inability to remove RSSI variations by extrapolation alone. However, the extrapolation succeeded in slightly reducing the estimation error of the direction of movement. From these results, we conclude that interpolation is the best method out of the three tested for processing RSSI for both indoor position and direction of movement estimation.

4.2 *Experiment B*

The experimental results of Experiment B are shown in Figures 11 through 14, and Tables 5 and 6.



(a) Four beacons A.
(b) Four beacons B.
(c) Four beacons C.
(d) Four beacons D.
(e) Six beacons.
(f) Eight beacons.

Figure 11: Position estimation error for experiment B.

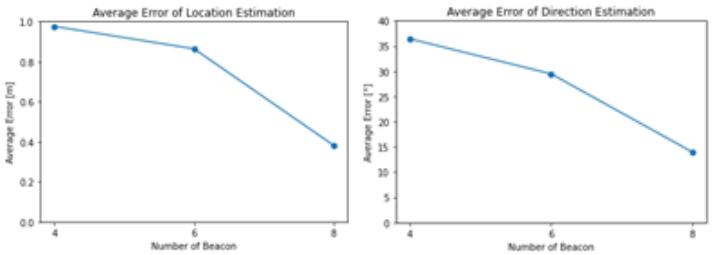
Table 5: Average direction of movement estimation error with various beacon placement pattern.

Direction	Average estimation error [°]			
	(a)	(b)	(c)	(d)
All	38.93	**36.18**	36.56	47.94
L	38.06	35.31	**32.21**	39.65
R	36.82	44.22	43.03	**35.17**
U	46.98	**36.56**	37.11	83.21
D	42.01	**37.12**	47.06	48.26
UL	44.88	39.67	**34.91**	45.81
LL	35.49	37.44	**34.85**	40.98
UR	31.83	**25.40**	29.72	35.20
LR	38.70	**33.24**	34.82	63.49

Table 6: Average direction of movement estimation error vs. the number of beacons.

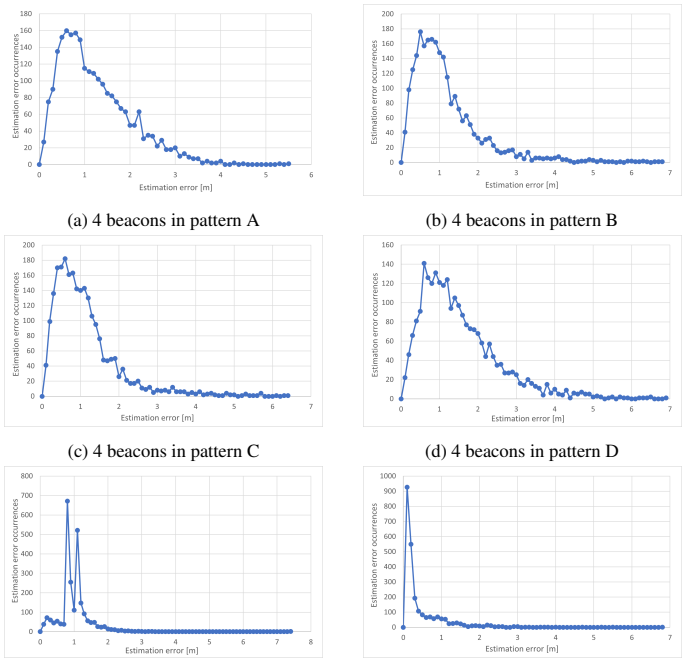
Direction	Average estimation error [°]		
	(d)	(e)	(f)
All	47.94	33.45	**10.61**
L	39.65	26.64	**10.99**
R	35.17	18.63	**12.78**
U	83.21	62.89	**10.52**
D	48.26	25.46	**11.10**
UL	45.81	32.36	**3.13**
LL	40.98	31.60	**15.28**
UR	35.20	29.99	**9.34**
LR	63.49	48.15	**12.96**

Figures 13a through 13d and 14a through 14d show the relationship between the position and direction estimation error and corresponding number of occurrences for the four beacon installation patterns (A, B, C, and D). These correspond to conditions in (a), (b), (c), and (d) in Figure 11 and Table 5. In these Figures, the position estimation error is plotted every 0.1 m, and the error in the direction of movement is plotted in 15° increments. In the case of four beacons used, we can see a large peak in the position estimation error distribution at less than 1 m in all patterns. However, the wide range and the multiple peaks of the estimation error distribution indicate that stable estimation is not possible. As for the moving direction estimation error, a peak can be seen around 15°, again in all patterns. We point out that beacon pattern D contains notable error occurrence distribution up to 180° compared to other patterns, in which the occurrences generally converge to a small number of instances at larger error values.



(a) position estimation error. (b) Direction estimation error.

Figure 12: Position and direction estimation error vs. number of beacons

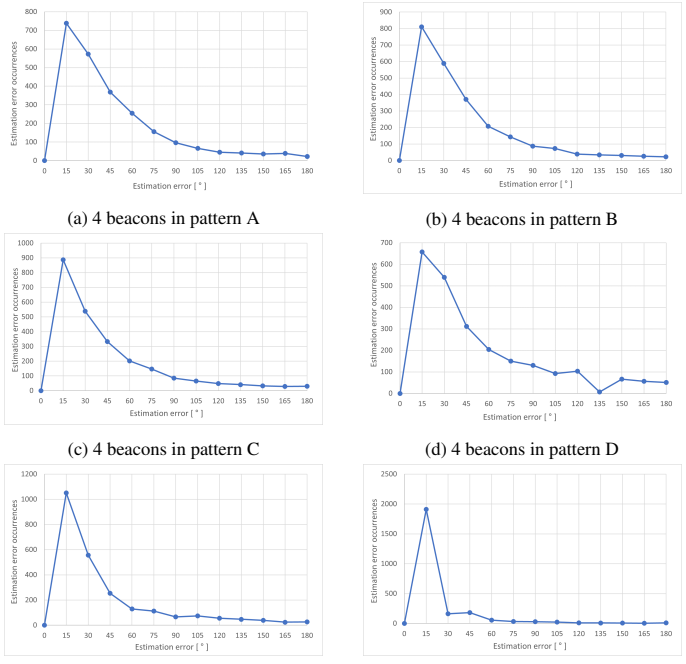


(a) 4 beacons in pattern A (b) 4 beacons in pattern B

(c) 4 beacons in pattern C (d) 4 beacons in pattern D

(e) 6 beacons (f) 8 beacons

Figure 13: Position estimation error distribution by number of beacons and its placement.



(a) 4 beacons in pattern A (b) 4 beacons in pattern B

(c) 4 beacons in pattern C (d) 4 beacons in pattern D

(e) 6 beacons (f) 8 beacons

Figure 14: Direction estimation error distribution by number of beacons and its placement. Directions were treated as continuous values.

Figures 13d through 13f and 14d through 14f show the relationship between the estimation error and the corresponding number of occurrences with 4, 6, and 8 installed beacons. They correspond to conditions in (d), (e), and (f) in Figure 11 and Table 6. In terms of position estimation error, the distribution of estimation error is shown in a rather wide range, often showing a notable number of

Table 7: Comparison of average position estimation errors.

System	Ave. loc. est. error [m]	Room size [m]	No. of APs/beacons
Proposed	0.33	5 x 8	4 to 8
WiFi CSI+Autoencoder [9]	0.9425	4 x 7	1
WiFi fing. + Autoenc. (WiDeep) [8]	1.21	14.5 x 4.5	59
BLE DOD [17]	0.54	16 x 10	2
BLE fing. + 3D-CNN [26]	0.72	3 x 6	8

occurrences up to about 3 m with four beacons. In the case of six installed beacons, there are two large peaks close to 1 m. In the case of eight installed beacons, there is only one large peak close to 0 m. These seem to indicate that increasing the number of beacons contributes to stabilizing the position estimation. As for the error in the estimation of the direction of movement, the peak of the error generated is in the range of 0 to 15° for any number of beacons, but the large estimation error is reduced when the number of beacons is increased. In particular, when there are 8 beacons installed, the percentage of estimation errors of 45° or more for all directions of movement in the test data is 7.3%. We believe this is sufficient accuracy for the intended application of our system.

Table 6 also compares the estimation error of the direction of movement when the number of beacons is 4, 6, and 8. From this Table, we can see that the estimation error becomes smaller as the number of beacons increases, similar to the position estimation results. The estimation accuracy percentage can be calculated using (5), where $\theta_{estimation_error}$ [°] ($0 \leq \theta_{estimation_error} \leq 180$) is the angle estimation error.

$$1 - \frac{\theta_{estimation_error}}{180}[\%] \tag{5}$$

From (5), the estimation accuracy of 4 beacons is 73.4%, that of 6 beacons is 81.4%, and that of 8 beacons is 94.1%. Therefore, the estimation accuracy of 8 beacons is 20.7% higher than that of 4 beacons. From the above results, we can say that increasing the number of beacons contributes significantly to the improvement of stable estimation accuracy, both in terms of position and motion direction. Thus, for a room with a size similar to the one in this study, the average value of the position estimation error can be expected to be well below 1 m, and the average movement direction estimation error can be expected to be around 10° if 8 beacons are used, which can both be considered to be sufficient accuracy for practical indoor navigation.

Although it is quite difficult to compare estimation results with other schemes since they will differ in terms of conditions (room size, number of APs/beacons and their placement, etc.), we have summarized the position estimation error comparison in Table 7. From this table, it seems safe to say that the proposed system achieves comparable or better estimation accuracy compared to the other similar methods. There is no data on movement direction estimation accuracy to compare with the proposed method, however.

5 Conclusion

In this study, we investigated an indoor position and direction estimation system consisting of a smartphone and BLE beacons. The RSSI measurements of multiple BLE beacons were fed into a DNN model that estimated both position and direction of movement simultaneously. Multiple time instances of the RSSI of the BLE beacons were used for the estimation. The post-processing of the RSSI of the BLE beacons and the number of beacons required for accurate estimation were considered. The results showed that temporal interpolation is the most effective post-processing of RSSIs before using them as input data for DNN models. In particular, interpolation contributes to the stabilization of the accuracy of the estimation of the direction of movement. Eight beacons in a space of 5 m × 8 m resulted in an average position estimation error of 0.330 m for all grid positions in this room, and a total error of 10.61° in the estimation of the direction of movement. We believe this is one of the most accurate estimations even among methods that rely on fingerprints. We also showed that the estimation of the direction of movement should be treated as a continuous value to increase the accuracy of the direction of movement estimation. Regarding the placement pattern and the number of beacons required, we confirmed that sparsely and evenly distributed beacons and a sufficient number of beacons contribute to stable estimation. In the measurement tested room, when 8 beacons were installed, the percentage of estimation errors of 45° or more for all directions of movement in the test data was 7.3%, significantly less than with a smaller number of beacons. With this level of accuracy, we believe that the system can be applied to actual indoor navigation systems.

In this experiment, we measured RSSIs with no obstacles in the surroundings. However, when we consider actual usage scenarios, objects and people in the room often become obstacles. Since we used multiple RSSI sequences to train the DNN model, the models may be able to handle some deviation in the RSSI measurements due to object placement owing to the generalization capabilities of the DNN. However, some form of reinforced learning will become necessary to handle an extensive amount of deviations due to the introduction of obstacles. This is currently out of the scope of this paper but is planned to be investigated in the future. We also assume the system to be used on the same floor and thus assume that the z-axis values are fixed. Multiple floor navigation systems will be investigated in the future as well.

Additionally, all user paths modeled in this work were simple straight paths at constant speeds. We would like to further expand our model to handle more complex paths, consisting of different direction and speed combinations.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The authors would like to thank the members of the Kondo Laboratory for their useful discussions.

References

[1] H.-G. Shin, Y.-H. Choi, C.-P. Yoon, "Movement Path Data Generation from Wi-Fi Fingerprints for Recurrent Neural Networks," Sensors, **21**(8), 2021, doi:https://doi.org/10.3390/s21082823.

[2] D. Manandhar, K. Okano, M. Ishii, H. Torimoto, S. Kogure, Maeda, Hiroaki, "Development of Ultimate Seamless Positioning System Based on QZSS IMES," in 21st International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2008), 1698–1705., Savannah, GA, 2008.

[3] Y.-T. Wang, C.-C. Peng, A. Ravankar, A. Ravankar, "A Single LiDAR-Based Feature Fusion Indoor Localization Algorithm," Sensors, **18**(4), 2018, doi: https://doi.org/10.3390/s18041294.

[4] A. Yassin, Y. Nasser, M. Awad, A. Al-Dubai, R. Liu, C. Yuen, R. Raulefs, E. Aboutanios, "Recent Advances in Indoor Localization: A Survey on Theoretical Approaches and Applications," IEEE Communications Survey & Tutorials, **19**(2), 1327–1346, 2017, doi:https://doi.org/10.1109/COMST.2016.2632427.

[5] A. Correa, M. Barcelo, A. Morell, J. L. Vicario, "A Review of Pedestrian Indoor Positioning Systems for Mass Market Applications," Sensors, **17**(8), 2017, doi:https://doi.org/10.3390/s17081927.

[6] S. H. Marakkalage, B. P. L. Lau, Y. Zhou, L. Ran, C. Yuen, W. Q. Yow, K. H. Chong, "WiFi Fingerprint Clustering for Urban Mobility Analysis," IEEE Access, **9**, 69527–69538, 2021, doi:https://doi.org/10.1109/ACCESS.2021.3077583.

[7] Z. E. Khatab, A. Hajihoseini, S. A. Ghorashi, "A Fingerprint Method for Indoor Localization Using Autoencoder Based Deep Extreme Learning Machine," Sensors, **2**(1), 2018, doi:https://doi.org/1109/LSENS.2017.2787651.

[8] M. Abbas, M. Elhamshary, H. Rizk, "WiDeep: WiFi-based Accurate and Robust Indoor Localization System using Deep Learning," in IEEE International Conference on Pervasive Computing and Communication, Kyoto, Japan, 2019, doi:https://doi.org/10.1109/PERCOM.2019.8767421.

[9] X. Wang, L. Gao, S. Mao, S. Pandey, "CSI-based Fingerprinting for Indoor Localization : A Deep Learning Approach," IEEE Transactions on Vehicular Technology, **66**(1), 763–776, 2017, doi:https://doi.org/10.1109/TVT.2016.2545523.

[10] C. Takahashi, K. Kondo, "Indoor Positioning Method for Augmented Audio Reality Using iBeacons," in 2015 IEEE Global Conference on Consumer Electronics (GCCE 2015), 468–469, Osaka, Japan, 2015, doi:https://doi.org/10.1109/GCCE.2015.7398636.

[11] C. Takahashi, K. Kondo, "Accuracy Evaluations of an Indoor Positioning Method Using iBeacons," in 2016 IEEE Global Confer-ence on Consumer Electronics (GCCE 2016), 425–426, Kyoto, Japan, 2016, doi:https://doi.org/10.1109/GCCE.2016.7800465.

[12] X. Zhao, Z. Xiao, A. Markham, N. Trigoni, Y. Ren, "Does BTLE measure up againstWiFi? A comparison of indoor location performance," in European Wireless 2014, 263–268, Barcelona, Spain, 2014.

[13] Z. Chen, H. Zou, H. Jiang, Q. Zhu, Y. Soh, L. Xie, "Fusion of WiFi, Smartphone Sensors and Landmarks Using the Kalman Filter for Indoor Localization," Sensors, **15**(1), 715–732, 2015, doi:https://doi.org/10.3390/s150100715.

[14] M. Nakano, N. Komuro, K. K., "Indoor Positioning Method based on BLE Location Fingerprint with Statistics Ap-proach," in 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE 2019), 1160–1163, IEEE, Osaka, Japan, 2019, doi:https://doi.org/10.1109/GCCE46687.2019.9015459.

[15] P. Kriz, F. Maly, T. Kozel, "Improving Indoor Localization Using Bluetooth Low Energy Beacons," Mobile Information Systems, **2016**, 1–11, 2016, doi: https://doi.org/10.1155/2016/2083094.

[16] S. Sadowski, P. Spachos, "Optimization of BLE Beacon Density for RSSI-Based Indoor Localization," in 2019 IEEE International Conference on Communications Workshops (ICC Workshops), Shanghai, China, 2019, doi: https://doi.org/10.1109/ICCW.2019.8756989.

[17] K. Kikuchi, R. Tazawa, N. Honma, "DOD-Based Indoor Localization Using BLE Beacons," in 2017 Microsoft Indoor Localization Competition IPSN 2017, Pittsburgh, PA, USA, 2017.

[18] Y. R. Kumar, J.-Y. Pyun, "BLE beacon based Indoor Positioning using Variance-Weighted KNN," in Symposium of the Korean Institute of Communications and Information Sciences, 128–129, Seoul, Korea, 2018.

[19] H. Yanagimoto, K. Hashimoto, T. Matsuo, "Indoor Positioning Estimatio Using BLE Beacons," in The 13th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2018), 124–128, Pattaya, Thailand, 2018.

[20] M. Hadwan, R. U. Khan, K. I. M. Abuzanouneh, "Towards a Smart Campus for Qassim University: An Investigation of Indoor Navigation System," Advances in Science, Technology and Engineering Systems Journal, **5**(6), 831–837, 2020, doi:https://doi.org/10.25046/aj050699.

[21] J. Paek, J. Ko, H. Shin, "A Measurement Study of BLE iBeacon and Geometric Adjustment Scheme for Indoor Location-Based Mobile Applications," Mobile Information Systems, **2016**, 2016, doi:https://doi.org/10.1155/2016/8367638.

[22] K. Echizenya, K. Kondo, "Estimation of indoor position and motion direction for smartphones using DNN to BLE beacon signal strength," in 2020 IEEE International Conference on Consumer Electronics Taiwan, (ICCE-TW), Taoyuan, Taiwan, 2020, doi:https://doi.org/10.1109/ICCE-Taiwan49838.2020.9258282.

[23] K. Echizenya, K. Kondo, "Comparison of RSSI Processing Methods for Improved Estimation Accuracy in BLE Indoor Position and Movement Estimation System Using DNN," in 2021 IEEE Global Conference on Consumer Electronics (GCCE 2021), 171–172, IEEE, Kyoto, Japan, 2021, doi: https://doi.org/10.1109/GCCE53005.2021.9622021.

[24] K. Echizenya, K. Kondo, "The Effect of Density and Placement of BLE Beacons on Indoor Location and Motion Direction Estimation Accuracy," in 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), APSIPA, Tokyo, Japan, 2021.

[25] T. Kitagawa, K. Kondo, "Evaluation of a Visually Handicapped Navigation System Using 3D Audio on a Smartphone under Realistic Conditions," in 2018 IEEE Global Conference on Consumer Electronics (GCCE 2018), 467–468, Nara, Japan, 2018, doi:https://doi.org/10.1109/GCCE.2018.8574718.

[26] K. Tasaki, T. Takahashi, S. Ibi, S. Sampei, "3D Convolutional Neural Network-Aided Indoor Positioning Based on Fingerprints of BLE RSSI." in 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA-ASC), Auckland, New Zealand, 2020.

[27] Z. Chen, Q. Zhu, H. Jiang, H. Zou, Y. Soh, L. Xie, R. Jia, C. Spanos, "An iBeacon Assisted Indoor Localization and Tracking System," in The 14th International Conference on Information Processing in Sensor Networks (IPSN '15), Seattle, WA, USA, 2015.

[28] Mathworks, "Piecewise Cubic Hermite Interpolating Polynomial (PCHIP)," https://www.mathworks.com/help/matlab/ref/pchip.html.

[29] F. N. Fritsch, R. E. Carlson, "Monotone Peicewise Cubic Interpolation," SIAM Journal of Numerical Analysis, **17**, 238–246, 1980, doi:https://doi.org/10.1137/0717021.

[30] Mathworks, "Cubic spline data interpolation (spline)," https://www.mathworks.com/help/matlab/ref/spline.html.

[31] C. de Boor, A Practical Guide to Splines, Applied Mathematical Sciences, Springer-Verlag, 1978.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 5, 129-145 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Long-term Bottom-up Modeling of Renewable Energy Development in Morocco

Jabrane Slimani[*,1], Abdeslam Kadrani[1], Imad EL Harraki[2], El hadj Ezzahid[3]

[1]*Research Laboratory in Information Systems, Intelligent Systems and Mathematical Modeling, National Institute of Statistics and Applied Economics, Rabat, Morocco*

[2]*Research Laboratory in Information Systems, Intelligent Systems and Mathematical Modeling, National School of Mines of Rabat,Rabat, Morocco*

[3]*Faculty of Legal, Economic and Social Sciences of Rabat–Agdal, Mohammed V University of Rabat, Rabat, Morocco*

ARTICLE INFO

Article history:
Received: 04 September, 2022
Accepted: 04 October, 2022
Online: 25 October, 2022

Keywords:
Energy system modelling
Power system
Renewable energies
OSeMOSYS
Optimization model
Morocco

ABSTRACT

Renewable energy is an essential source of green growth for countries facing a shortage of fossil fuels. They offer a sustainable, inexhaustible, carbon-free solution to the future energy dependency of nations. Morocco, which has no traditional energy resources, depends entirely on the international primary energy market to meet its growing demand. For this reason, Morocco launched the National Energy Strategy in 2009 to reach 42% renewable production by 2020. This strategy has been renewed to 52% by 2050. Thanks to this policy, the country has been able to address most of its energy challenges. This study analyzes the energy mix of Morocco from 2010 to 2050. The methodology adopted is to simulate Morocco's electricity mix for this period. We assumed we were at the beginning of deploying the country's energy policy to assess the adopted strategic decisions. The analysis shows that the different technological solutions for electricity production chosen at the beginning of Morocco's energy transition could be better. Indeed, the decision to develop concentrated solar power as the leading renewable source and coal as a backup option, for example, appears to be contested. However, according to the third scenario of our study, renewables have the potential to become the main source of energy for the Moroccan power grid.

1 Introduction

This paper is an extension of work initially reported in the "2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)" [1].

The Kingdom of Morocco relied primarily on foreign supplies to meet its growing energy needs during the first decade of the 21st century [2]. Forecasts for 2008 suggested that the country's high rates of population growth and urbanization, as well as its economic prosperity, would contribute to higher energy consumption and a greater mismatch between supply and demand [3]. This dependence on imported fossil fuels has widened the country's trade and financial deficits [4]. On the other hand, Morocco has significant potential to meet its energy needs through renewable sources such as hydroelectricity, solar energy, and wind power [5].

Considering the country's greatest renewable energy potential and the significant scarcity of fossil-fuel reserves, the kingdom initiated a new national energy policy in 2009 [6]. This strategy prioritized supply security, energy mix diversity, low cost, safety, efficiency, and environmental cleanliness [7]. This strategy intended to raise renewable energy's installed capacity share to 42% in 2020 and 52% in 2030 [8].

Although it has a percentage of renewable energy (RE) that is around 36.8%, Morocco has not yet accomplished the goal it set for itself in the energy transformation field. The target was to achieve a share of renewable energy in the power sector equal to 42 percent by the end of 2020 [9]. Despite this, the renewable energy targets specified have been increased to exceed the current goal of 52% of the national power mix by 2030 [10, 11].

The most significant delay was in the implementation of the Solar Plan [12]. This enormous project, with a significant capacity for electric energy production, has provided Morocco with a valid

[*]Corresponding Author: Jabrane SLIMANI, Research Laboratory in Information Systems, Intelligent Systems and Mathematical Modeling, National Institute of Statistics and Applied Economics, Rabat, Morocco, Email: jslimani@insea.ac.ma

energy alternative and a tool for socioeconomic growth, allowing it to become a "reference" in Africa for renewable energy production! In addition, the accomplishments in the hydroelectricity field follow the goals. Towards the wind side, we expect to surpass the goals barely. Despite this, the cumulative impact of the two, whose targets will be surpassed in 2021, will not be sufficient to compensate for the delay in the solar plan [13]. Because of this circumstance, we began to wonder whether the decisions and approaches suggested in the Moroccan solar plan, developed after the energy strategy implementation, were especially relevant. In addition, we questioned the reliability of the guidance of the subject-matter experts who supported these choices.

In light of this, this article provides a historical and projected assessment of Morocco's electricity system. The plan is to simulate the power mix in Morocco from 2010 to 2050 as though we were at the beginning of the country's energy policy rollout. The analysis is based on OSeMOSYS (Open-Source Energy Modeling System), a long-term planning model that allows us to compute the energy supply mix that best fulfills the energy service demands in each year and a time step of the investigated scenario while reducing total discounted costs.

The remainder of this work will be organized as follows: First, we will go through the OSeMOSYS model and the applied analytical approach. In this section, we will go over the changes made to the model code's initial version. These changes provided a more accurate picture of Moroccan energy policy. Section 03 then presents our model's various input data and assumptions. We also recall the scenarios' assumptions and attributes. Then, in Section 04, we will present and debate the key findings of the simulated scenarios, and ultimately, we will give the study's primary conclusions. The final purpose of our analysis is to evaluate Morocco's judgments made between 2010 and 2020, as well as the best solutions that should be chosen. Furthermore, would it be possible to speed the shift away from fossil fuels between 2020 and 2050, and what would the costs be?

2 Material and Methods

2.1 Tool: Open-Source energy Modelling System

Bottom-up modeling approaches are applied to optimize power systems in developing nations with suitable technology and supply resources [14]. One example of a model that fits this description is the bottom-up, dynamic, linear optimization model known as the Open-Source Energy Modelling System (OSeMOSYS), which is used for integrated evaluation and energy planning [15]. OSeMOSYS seeks to accomplish this by considering various technological, economic, and environmental factors, all while striving to achieve the lowest possible total discounted cost [16]. This model's designers constructed it to divide its functionality into several "blocks." These features are connected to the following aspects: prices, capacity adequacy, energy balance, emissions, and provisions for renewable energies [17]. What distinguishes one block from another are the equations, the formulae and the restrictions, the intermediate variables, and the parameters introduced by the analyst. [15, 16, 18].

Initially, the code for OSeMOSYS was developed in GNU MathProg. More recently, it has been rewritten in the GAMS (General algebraic modeling system) and Python programming languages [19]. Within the context of the conference publication [1], we carried out the simulation utilizing the GAMS implementation of the model. However, in this version, with the additional information we were able to obtain throughout our research project, we could carry out the simulation using the Python version that was running under pyomo.

2.2 Basic Code Adjustments

Our modeling of the Moroccan power system aims to determine the most efficient way to meet the country's predicted long-term electricity demand, while also considering Morocco's energy strategy targets in terms of the amount of renewable capacity that may be built. In the original version of OSeMOSYS, constraints relating to the availability of resources, the characteristics of renewable technologies, and the development of demand were already incorporated. Although, the requirements of EnR integration into the power system are stated in terms of yearly renewable energy production considerations.

The modeler may introduce the integrative constraint for RES in the energy system by using Equation 1. The different terms of the equation are detailed in Table 1. "r" and "y" represent the data sets for the region and the modeling year, respectively. This equation is encoded using the Pyomo coding scheme, as in box 01. Nevertheless, applying this equation to the specific situation of Morocco's energy strategy is not an ideal solution.

$$\forall_{r,y} \quad REmPT_{r,y} \times RETPA_{r,y} \leq TREPA_{r,y} \tag{1}$$

$$\forall_{r,y} \quad REmCT_{r,y} \times TPCA_{r,y} \leq TRECA_{r,y} \tag{2}$$

Therefore, to accurately represent the constraints imposed by the renewable installed capacity, we had to convert the previous mathematical "equation (1)" into the new "equation (2)". The different terms of the "equation (2)" are detailed in Table 2. Additionally, constraints were introduced to the original code, as demonstrated in "box 02". The technique of obtaining the TotalRECapacityAnnual variable is determined by "equation (3)", "t" symbolizes technologies' power plants. The different terms of "equation 3" are detailed in Table 3. The second term in "equation (2)" is the variable TotalPowerCapacityAnnual derived from the TotalCapacityAnnual variable and the PowerTagTechnology parameter as shown in "equation (4)". The additional parameter PowerTagTechnology was included to allow the model to distinguish between electricity-generating technologies and other technologies defined in the model. The different terms of the "equation (4)" are detailed in Table 4. The REMinCapacityTarget(r, y) parameter is the third component of the "equation (2)". It was created in order to set a minimum target for renewable capacity.

$$\forall_{r,t,y} \quad \sum_{t} TCA_{r,t,y} \times REtgT_{r,t,y} = TRECA_{r,y} \tag{3}$$

$$\forall_{r,t,y} \quad \sum_{t} TCA_{r,t,y} \times PtgT_{r,t,y} = TPCAnnual_{r,y} \tag{4}$$

Table 1: Summary of datasets used

Abbreviation	OSeMOSYS designation	Description
REmPT	REMinProductionTarget	The minimum renewable production target desired by the analyst (parameter)
RETPA	RETotalProductionOfTargetFuelAnnual	The Annual Production of the fuels marked as renewable in the model (variable)
TREPA	TotalREProductionAnnual	The annual production of all technologies marked as renewable in the model
REmCT	REMinCapacityTarget	The minimum renewable capacity target desired by the analyst (parameter)
TPCA	TotalPowerCapacityAnnual	Annual capacities of technologies that convert primary energy into power (new variable)
TRECA	TotalRECapacityAnnual	A new variable introduced to the system to identifie the total annual renewable capacity
REtgT	RETagTechnology	A binary parameter indicating renewable technologies, with a value of 1 indicating renewable technologies and 0 otherwise (parameter)
TCA	TotalCapacityAnnual	The existing total capacity of technology “t” for year “y” (variable).
PtgT	PowerTagTechnology	A binary parameter indicating power technologies, with a value of 1 indicating electricity-generating technologies and 0 otherwise (parameter)

3 Data Collection and Analysis Approach

3.1 Base year data

The year 2010 will serve as the foundation for building our analysis, as was said earlier. As a result, when the model was created, the current properties of the power system were used as starting parameters for the optimization process. Consequently, the treatment got off to the most effective beginning point imaginable. As a result, our model considers the electricity generation industry in Morocco at the end of 2009 [20]. There is information supplied on the capacity that was built before the year 2010. (Figure 1).

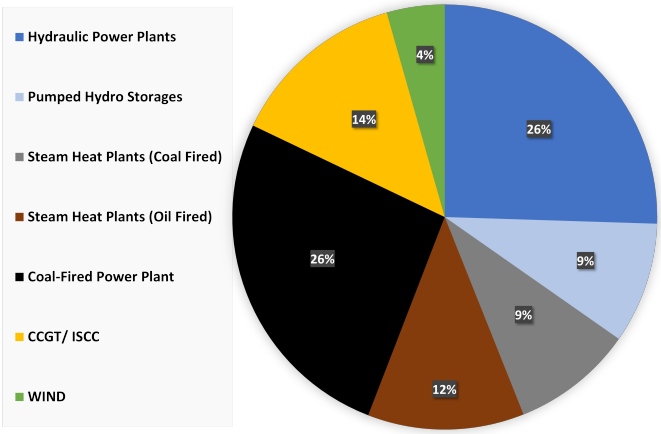


Figure 1: Installed capacity at the end of 2009 in Morocco [20]

It is important to note that the electricity system in Morocco does not cover the entire kingdom territory. This is mainly some generation capacity installed on the demand side in remote areas, such as the kingdom’s south. These regions are supplied with power by gas turbines and diesel generators connected to a local grid. In our research, we did not take into consideration these capacities. Instead, we have considered the fact that the electricity fleet must connect to the national grid to meet the demand associated with this region.

3.2 Reference Energy System

When modeling energy systems, it is common practice to employ a representation in the form of a network for all the practical tasks required to provide end-users with a variety of energy sources. The term “Reference Energy System” (RES) refers to this particular representation [21]. In the context of our study, the selected RES is shown in Figure 2. The block denotes the various technologies in the RES, while the lines stand in for the various energy carriers. Technologies can refer to all kinds of processes performed on certain energy carriers, such as extraction, refinement, conversion, transportation, distribution, and usage of that energy carrier.

In the case of Morocco, where energy is primarily obtained through imports, we have not considered the relatively insignificant sector of primary energy extraction. On the other hand, the refinery has not been considered because its activity level dropped significantly after 2010 and stopped in 2015. As we had stated previously, when it came to converting primary energy into electricity, we depended on the technologies already in place at the beginning of 2010. These include coal-fired power plants, natural gas-fired power plants, oil-fired power plants, grid-connected diesel generators, wind farms, hydroelectric power plants, and a pumped storage power station. We have included in this mixture two technologies that are candidates for the development of renewable capacity. These technologies are photovoltaic and solar concentration.

The capacity of these power stations has been aggregated in a

modeled setting. In other words, we do not simulate each power plant isolated. Instead, a block of power plants that convert one type of primary energy into power will be depicted together. Decision-makers can then pick the size of the new capacity required after obtaining the total ideal capacity required. In the case of transmission and distribution networks, precisely the same method was taken. These grids were modeled as a single technology that transforms the electricity from these plants into transmission electricity for the transmission grid, and then another technology that transforms this electricity into electricity for end users for the distribution grid. The transmission grid and the distribution grid were both modeled as separate technologies. On the demand side, each type of electricity consumption was combined into a single estimate for the national electricity demand.

3.3 Inputs and assumptions

OSeMOSYS's simulation of the Moroccan power system relied on several assumptions to generate accurate results.

3.3.1 Temporal precision

In the conference paper [1], we decided to sequence the modeling horizon over five years. In this extended essay paper, however, we have chosen annual modeling. In addition, to simulate the seasonal and daily variation of electricity demand and the intermittent nature of renewable energy sources, we divided each year of the model into a total of six-time steps. There are two periods each day: day and night. There are also three seasons: winter, intermediate, and summer. This gives us a total of six periods: WD and WN for the daytime and the night of a day during the winter; ID and IN for the daytime and the night of a day during the intermediate season; and SD and SN for the daytime and the night of a day during the summer.

3.3.2 Electricity Demand

According to the data provided by ONEE in its numerous activity reports [22], the yearly increase rate of energy consumption was around 6% from 2010 to 2016. This growth rate was observed throughout the entire period. Thereafter, this rate continued to fall until it reached 5% in 2018. Furthermore, the Moroccan Ministry of Energy analyzed energy demand until 2050 [23], which established three different demand evolution scenarios for the time spanning from 2020 to 2050. A trend macroeconomic scenario, in which the Moroccan economy maintains the same trend (3.5%) recorded from 2009 to 2018; a high macroeconomic scenario, in which a gradual economic recovery is expected to reach 5% in 2050; and a pessimistic macroeconomic scenario, in which the downward trend recorded from 2009 to 2018 is expected to continue to 2% in 2050. ONEE statistics have been used as the basis for our examination of demand from 2010 through 2018. On the other hand, for 2019-2050, we have decided to go with the findings of the trend scenario developed through the research carried out by the Ministry of Energy (Figure 3).

3.3.3 Costs

Any such study relies highly on the investment prices selected for the technologies being considered, in addition to the costs of fuel, operation and maintenance (O&M), and non-operational expenditures. The investment costs and fixed operation and maintenance (O&M) expenses were expressed in USD per kW of installed capacity. with variable O&M costs stated in US dollars per gigajoule and fuel costs expressed in US dollars per unit. These costs were gathered from various sources [24–30], and we have displayed them in Figures 4,5, and 6, respectively.

3.3.4 Common data and hypothesis

Several other statistics and assumptions are shared in all scenarios. These figures were obtained from a variety of sources [24, 31–33]. Table 2 summarizes the various values estimated during the modeling period.

The capacity factor is the most crucial criterion to consider in this category. It represents the available capacity for each period, usually given as a proportion of the total installed capacity. This value remains constant throughout the year for conventional technologies, making it possible to consider unplanned disruptions. On the other hand, the value of renewable energy sources fluctuates depending on the period in question. Thus, the intermittent nature of these technologies is also represented here. Figure 7 provides a visual representation of the various capacity factors of renewable energy that were taken into account in our research.

The second parameter is the Availability Factor. It represents the most extended period that a particular technology can function throughout an entire year, represented as a percentage of the year ranging from 0 to 1. It enables breaks to be scheduled in advance. The third one is the operational Life Cycle, which is the usable life of a technology measured in years. The other metrics include each technology's efficiency and CO2 emission rate.

To establish certain constraints, we had to make some additional assumptions. Among these assumptions is the discount rate, set at 5%, based on the average rate applied to loans by Moroccan banks [34]. Furthermore, the reserve margin was set at 20% of the installed capacity, and no emission limits were imposed.

3.3.5 Assumptions and Scenario characteristics

Under the Moroccan energy plan approved in 2009, three scenarios have been selected for implementation. The main difference between these scenarios is the share of renewable energy in the installed electricity capacity. The first scenario is a Business As Usual (BAU) scenario that does not impose a lower limit on the share of renewables. This scenario reflects a trend in the electricity mix. At the same time, the optimization model is proposed to incorporate renewable energies. The second scenario is the one that matches the target of the Moroccan energy policy, which is to reach a rate of 42% of installed renewable power in 2020 and 52% in 2030. The third scenario is the scenario that estimates the feasibility of increasing the rate of integration of renewable energy in the electricity mix. For this purpose, the Moroccan energy strategy's objective has been revised to 60% of installed renewable power in 2030 and 80% in 2050.

4 Results and Discussion

To fully comprehend the results of the Moroccan energy policy, the findings are contrasted with the real situation that occurred on the reality in Morocco from 2010 to 2018. Recommendations will be based on the research done for the years 2019–2050.

4.1 Power Generation Capacity

If we look at the three scenarios together, we see that, in the first scenario, the proportion of renewable energy sources is relatively low (Figures 8;9;10). Indeed, renewable energy is scarce from 2010 to 2035; This results from the very high cost of these technologies at the onset of the modeling period and the unavailability of a minimum renewable capacity target. In this scenario, solar energy has not been chosen to be included, but the incorporation of wind energy into the mix will begin in earnest in the year 2035. This can be explained by the fact that technology relating to wind energy has become more economically mature than those relating to solar energy.

On the other hand, one can make the observation that the total capacity required to fulfill the demand in the third scenario is more significant than that required in the first two scenarios. The high integration rate of RE, which reached 60% in 2030 and 80% in 2050, is the reason for this gap in capacity. In addition, it appears that the retirement of coal-fired power plants is happening faster as a result of investments in RE. In fact, coal-fired generation capacities are still present until 2040 in the first scenario; however, in the second and third scenarios, these capacities are almost nonexistent beyond 2030. Regarding the utilization of natural gas, we can observe that this production method is present in each scenario. Furthermore, the supremacy of this conventional technology over its alternatives can be attributed to the reliability of the technology and the low cost of the commodity, in this case, natural gas, compared to the cost of fuel oil.

Table 2: Data and hypothesis common to all scenarios [24, 31–33]

		Technology							
	Units	Coal_PP	Oil_PP	GAS_PP	WIND	PV	CSP	HYD	PHS
Avarage capacity factor	%	85	90	87	35	27	70	51	100
Availability factor	%	84	89	91	100	100	100	95	100
Life cycle	Years	40	25	30	25	25	25	80	50
Efficiency	%	39	35	54	100	100	100	32	100
CO2 emissions	ton/PJ	0.0905	0.0589	0.0503	0	0	0	0	0

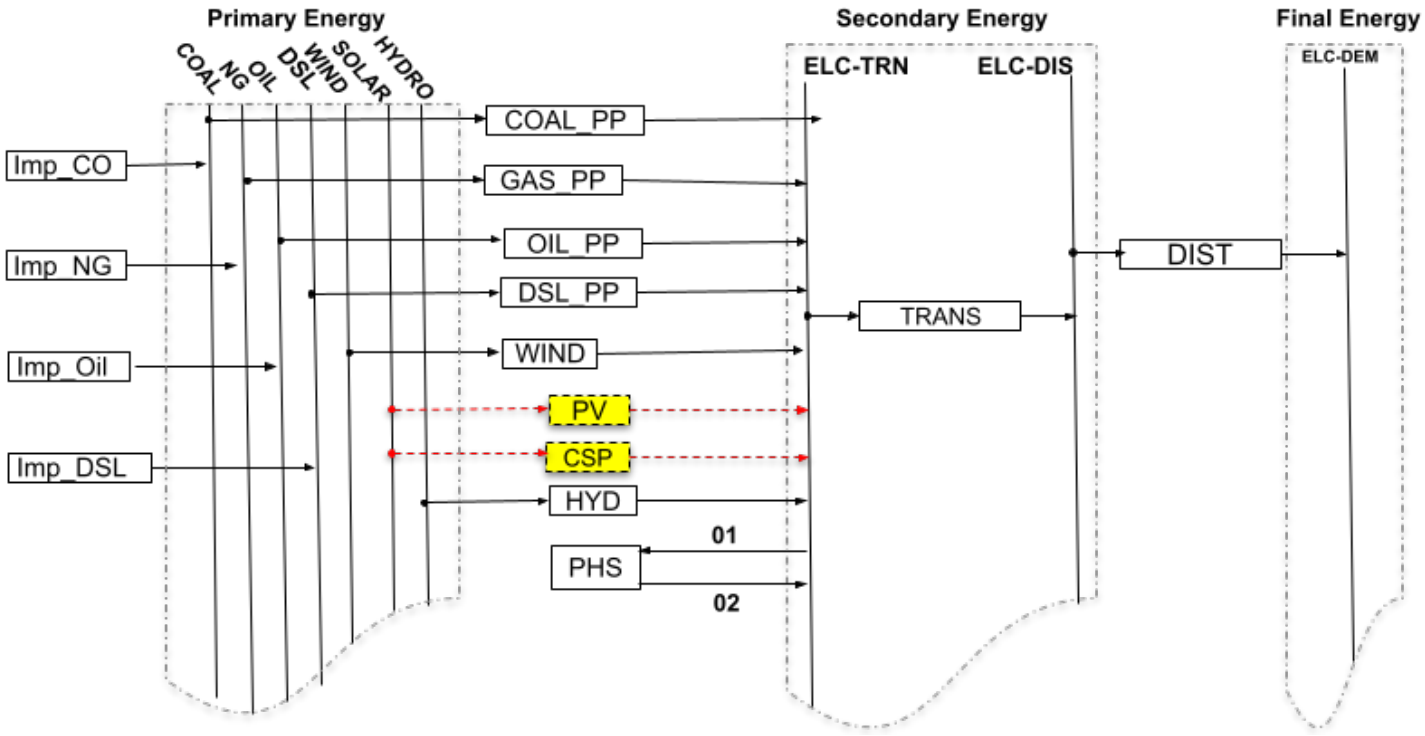


Figure 2: Electricity supply model of Morocco's reference energy system in OSeMOSYS/ source : illustration, made by the authors

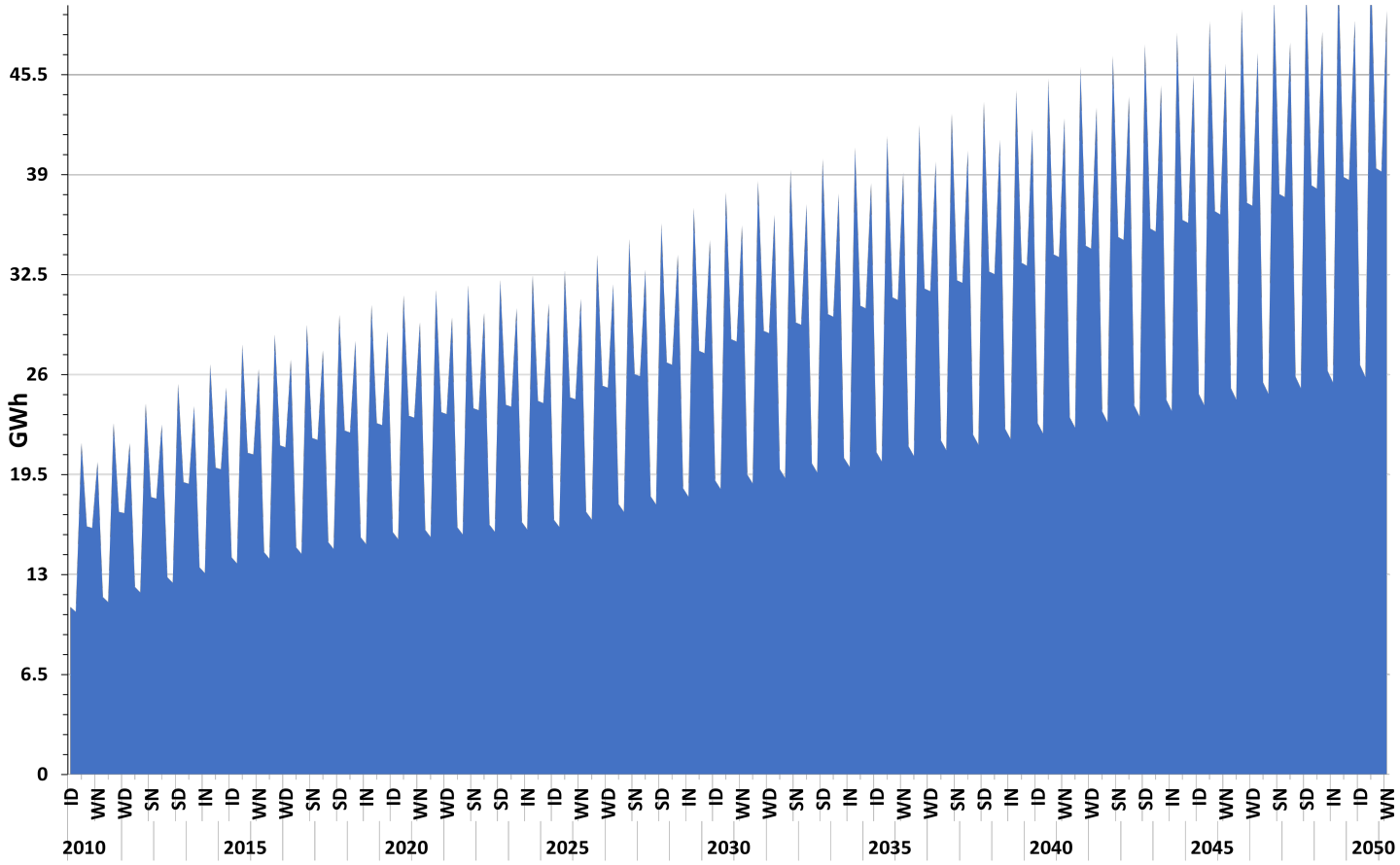


Figure 3: Electricity demand 2010-2050 in GWh

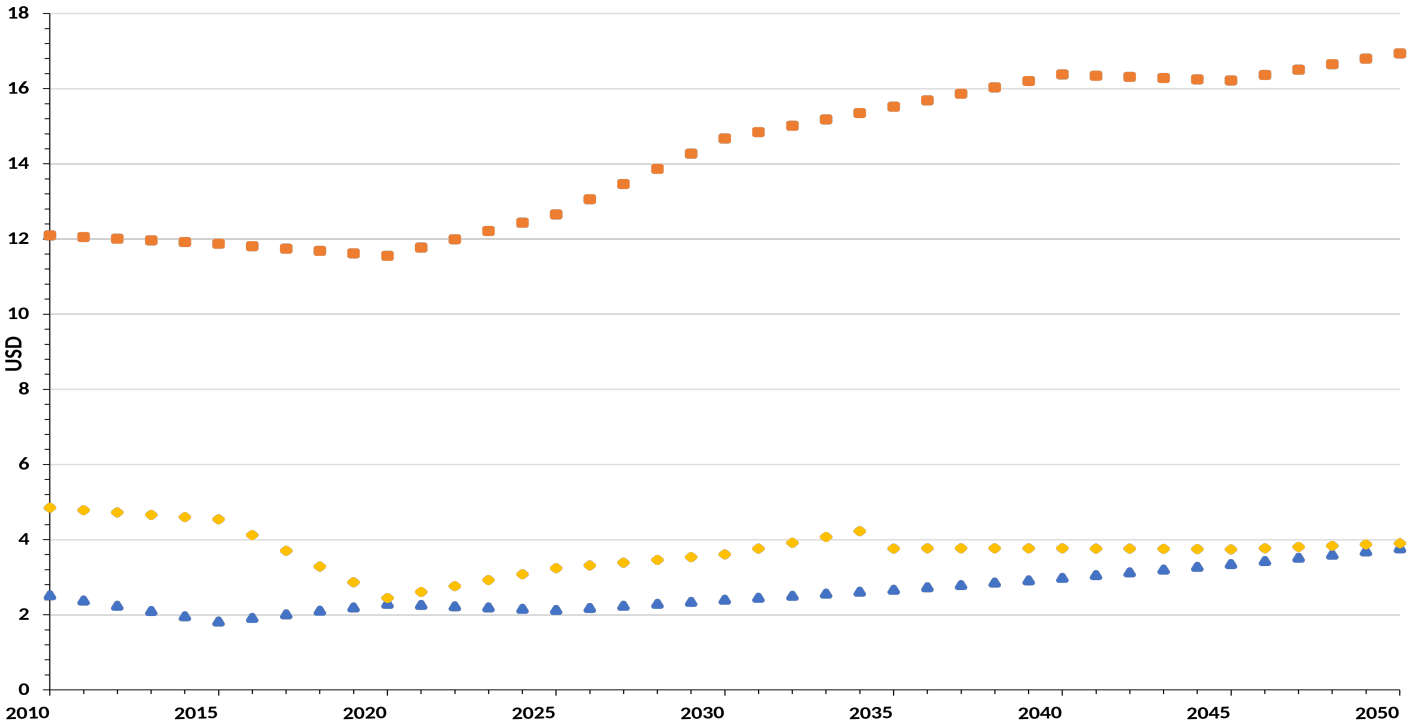


Figure 4: Energy Prices by Source, 2010-2050

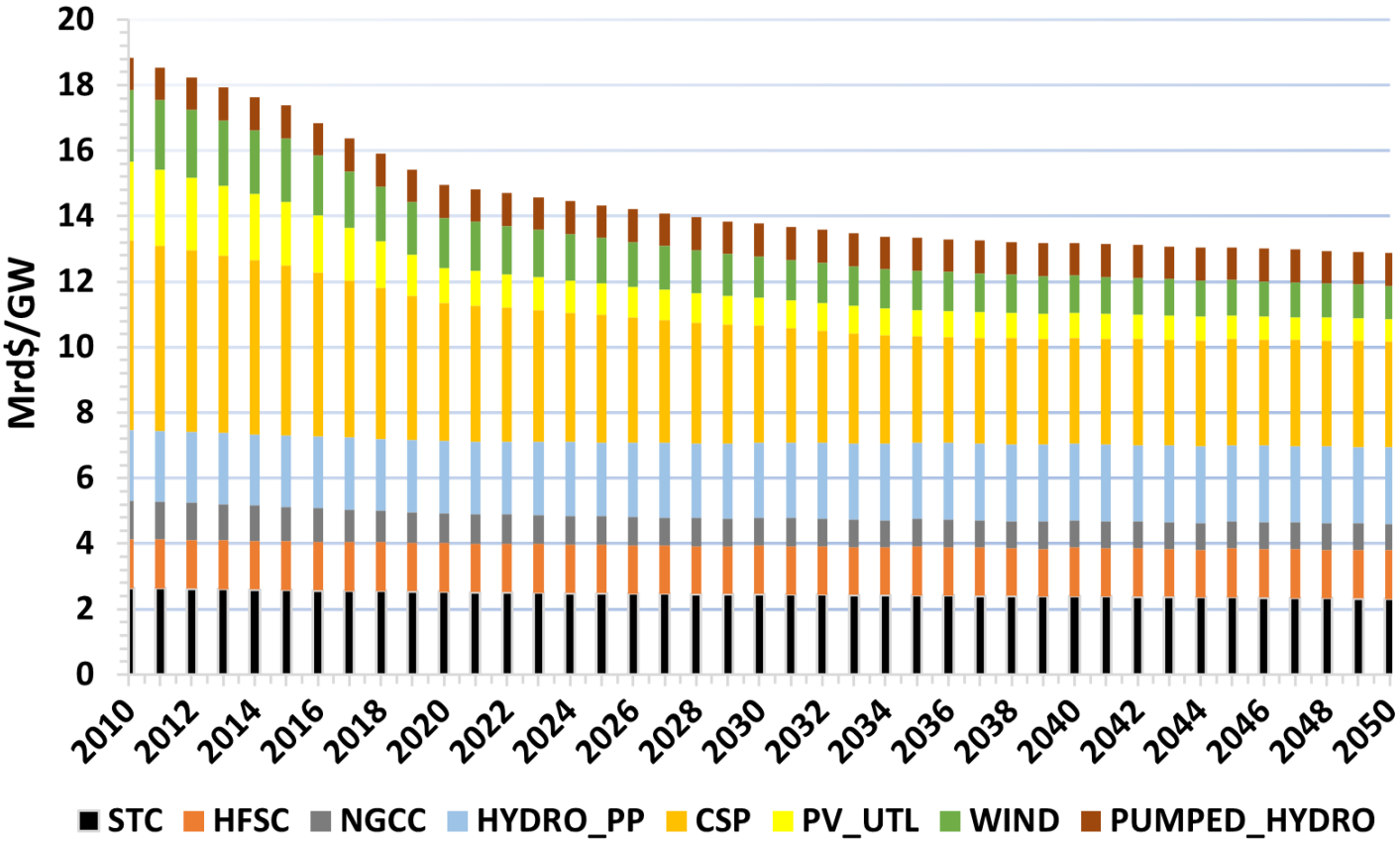


Figure 5: Capital cost for power plants 2010-2050

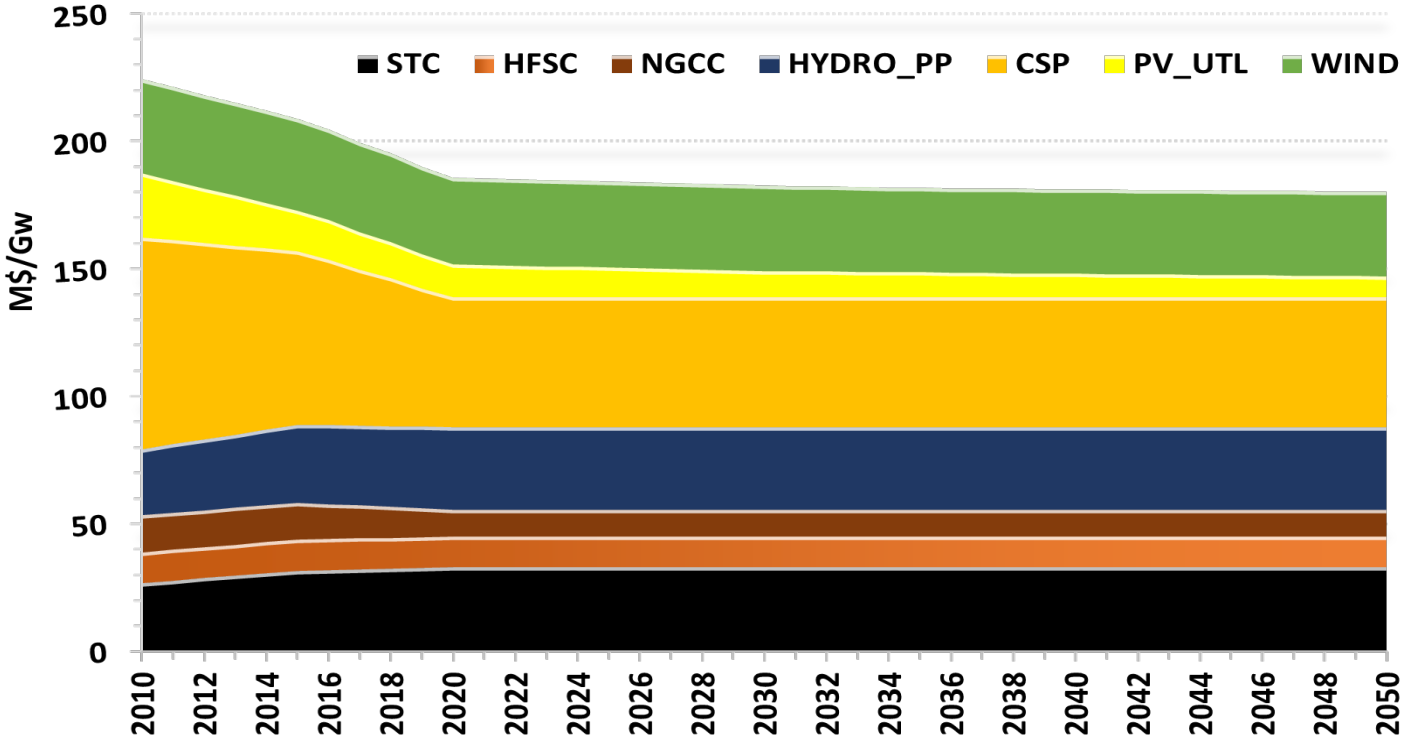


Figure 6: Fixed O&M costs for power plants,2010-2050

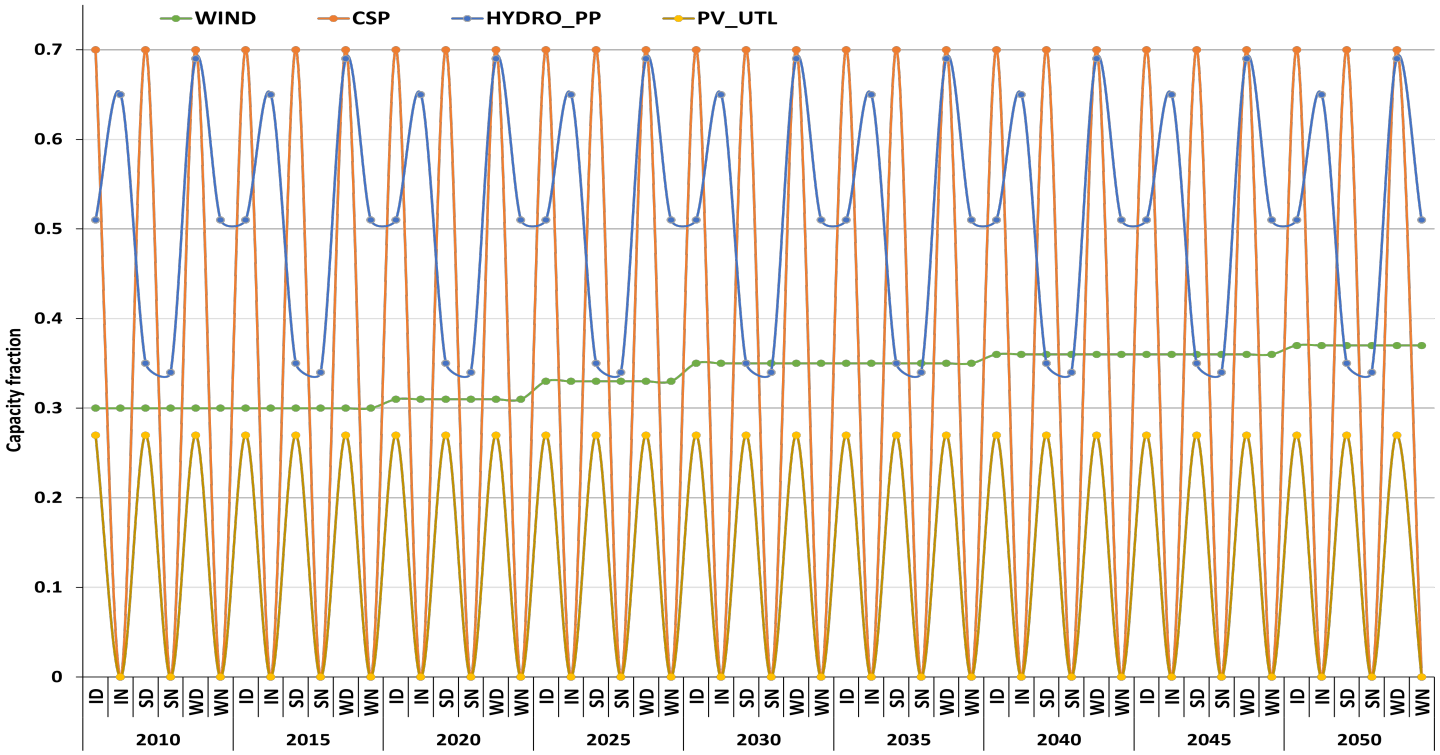


Figure 7: Capacity Factor Renewable Energy

Furthermore, we note that only two of the three possible renewable technologies were chosen, Wind energy and solar photovoltaic energy. However, none of the possible results involve using concentrated solar electricity. A finding that sheds light on Morocco's decisions in 2010 regarding the technologies it plans to deploy.

Figure 4 demonstrates that the majority of the electricity mix in the BAU scenario is composed of conventional technologies. Renewable energy sources account for just 6% of the total capacity in the year 2030, and this percentage will not reach 42% until the year 2050. In 2035, pumped storage technology will account for 48% percent of total installed capacity, making it one of the technologies with the most significant proportion of installed capacity. This focus on PHS can be explained by the fact that the capacity and availability factors were estimated at 100%. In addition, the model did not stipulate an upper capacity limit for the system. In light of these assumptions, one can question the veracity of the results. These assumptions, however, may be appropriate even in the lack of reliable facts on the issue. For starters, pumped storage power stations can run as long as the upstream reservoir is adequately replenished. This can be achieved by either flowing water from a river or pumping water from the downstream tank. Then, for the availability factor and unexpected shutdowns and maintenance, two pumping units and two turbine units are sufficient to ensure the station's operation. Furthermore, regarding available capacity, Morocco has a relatively vast coastline that may be utilized as marine PHS, considerably improving Morocco's potential in this technology.

In contrast to scenario 01, solar PV is the first renewable technology to be integrated into the electricity mix in the 2e scenario (Figure 9), with wind beginning to gain a significant role from 2035 to 2050. We also find that the share of gas power plants is directly proportional to the integration rate of renewable energy sources. Indeed, in 2050, the share of NG-PP approaches 50% in scenario 02, whereas it does not exceed 40% in scenario 01. This result demonstrates the significance of this technology in the context of the vast development of renewable energies.

This observation is corroborated in scenario 03 (Figure 10), when the rates of renewable energy integration rise substantially. This is owing to the model's need to meet significant minimum capacity demand. This is also something we observe at this time. The built wind capacity had expanded dramatically at the start of the modeling period, in contrast to what we stated in Scenario 2. It should be highlighted that, based on current price trends, solar PV will be pretty competitive for average renewable energy integration between 2020 and 2040. However, technology to offer backup is required. Wind power, however, is critical for rapid integration with a high percentage of renewable energy. Indeed, with rates reaching 60% from the start of the projected period, significant investments in natural gas power plants have been made, highlighting the critical potential of this technology to offer the necessary reserve to deal with renewable energy's intermittent nature. Once the integration rate reaches 80%, establishing PHS facilities becomes imperative. This is what we notice in the year 2050.

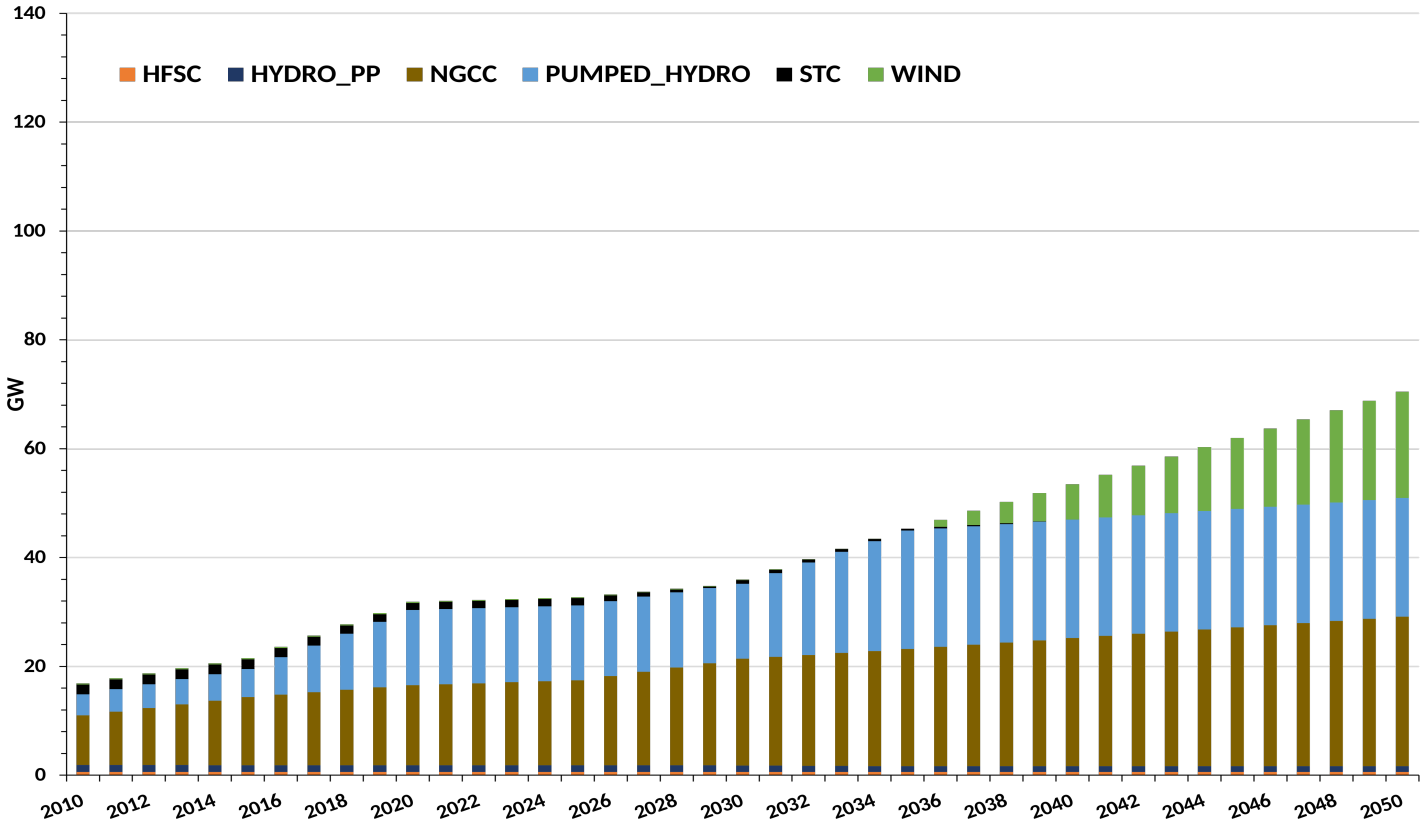


Figure 8: Total Annual Capacity (Scenario 1)

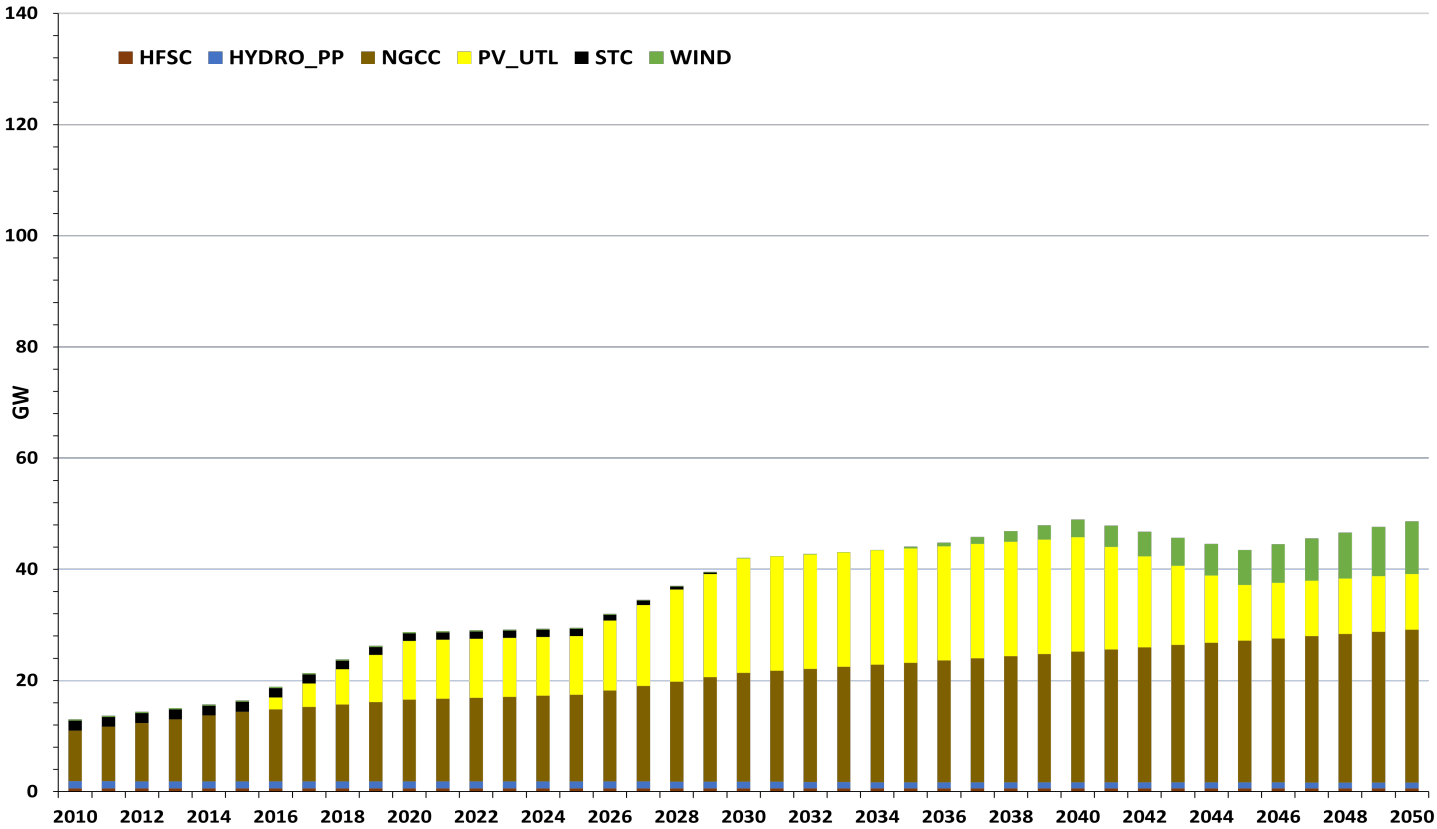


Figure 9: Total Annual Capacity (Scenario 2)

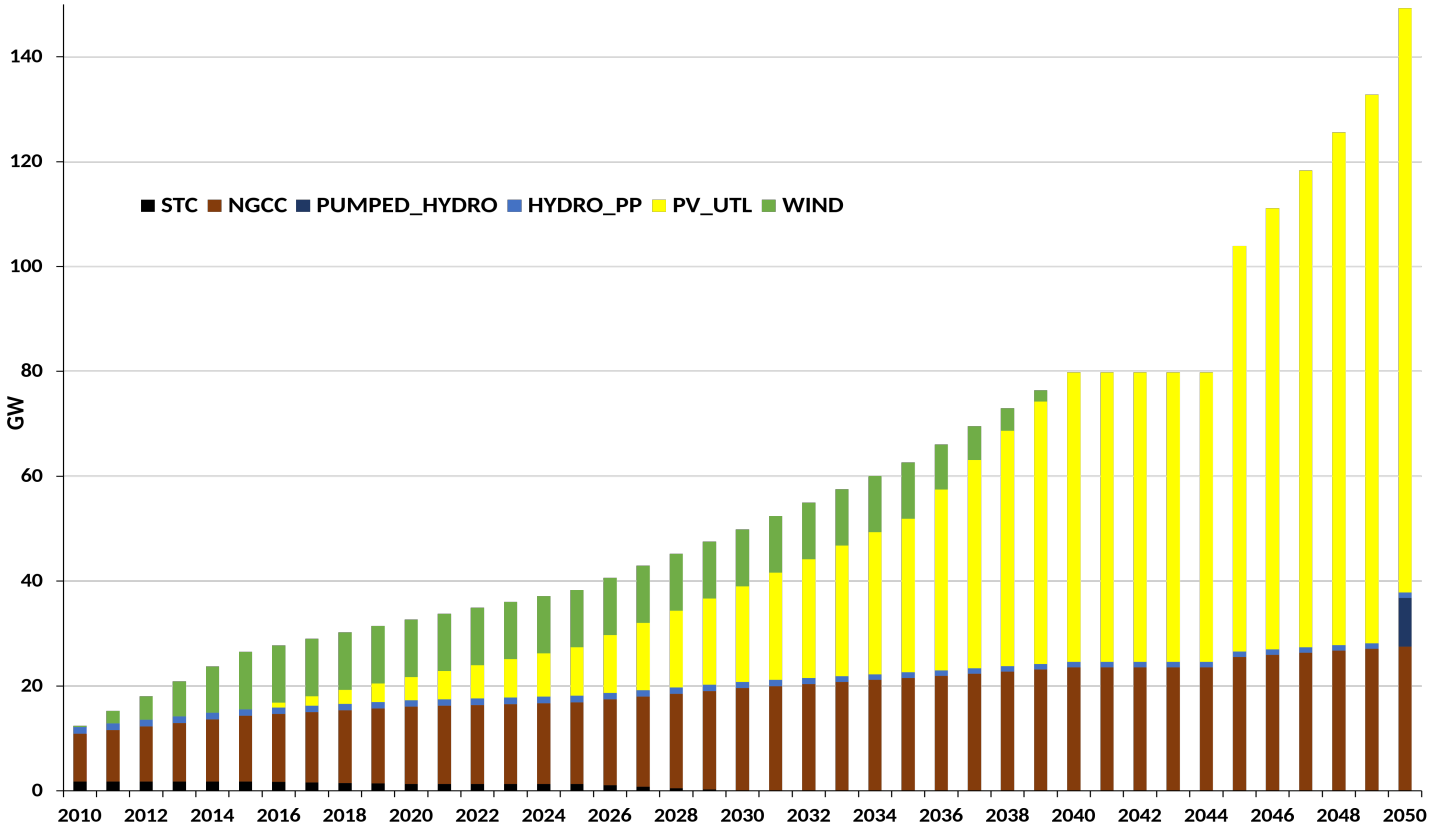


Figure 10: Total Annual Capacity (Scenario 3)

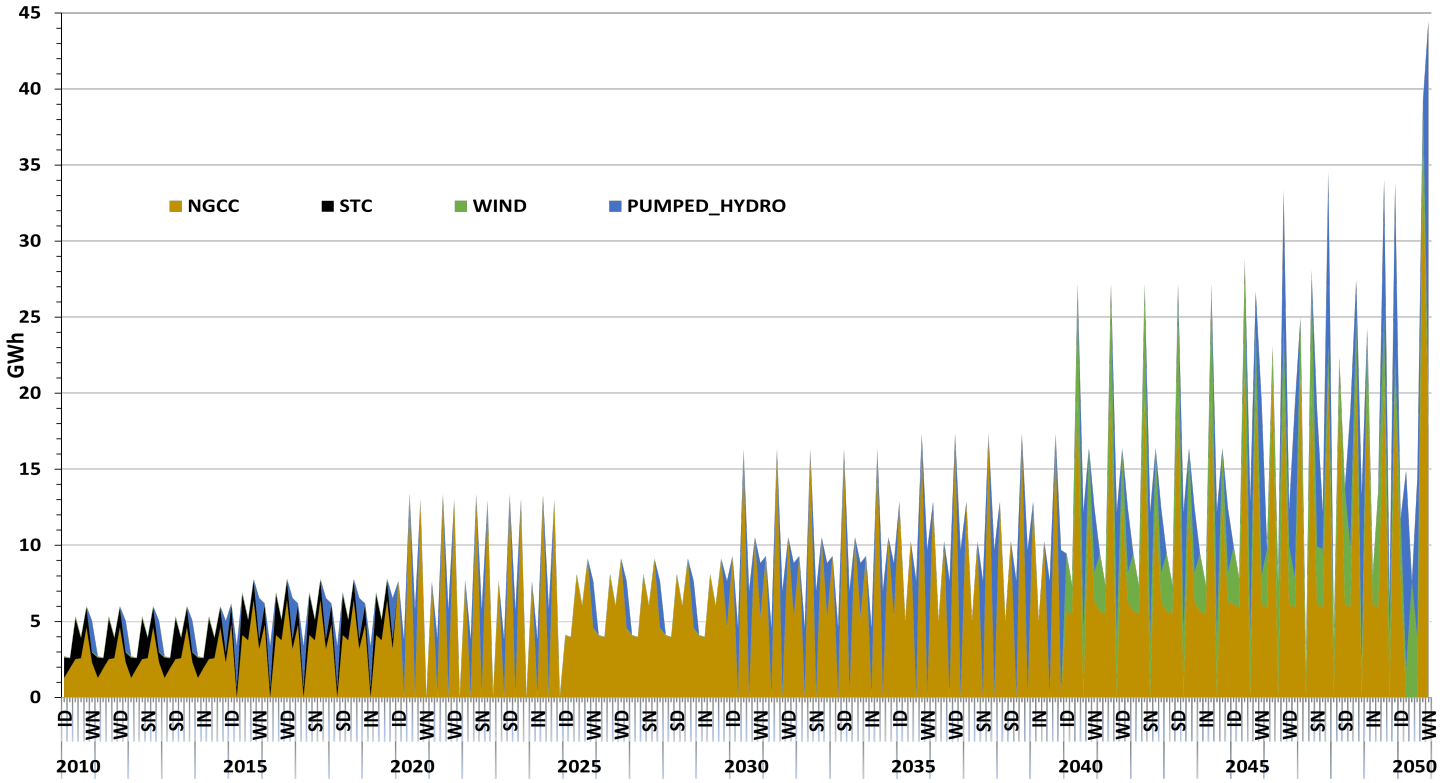


Figure 11: Technology Production by Time Slice (Scenario 1)

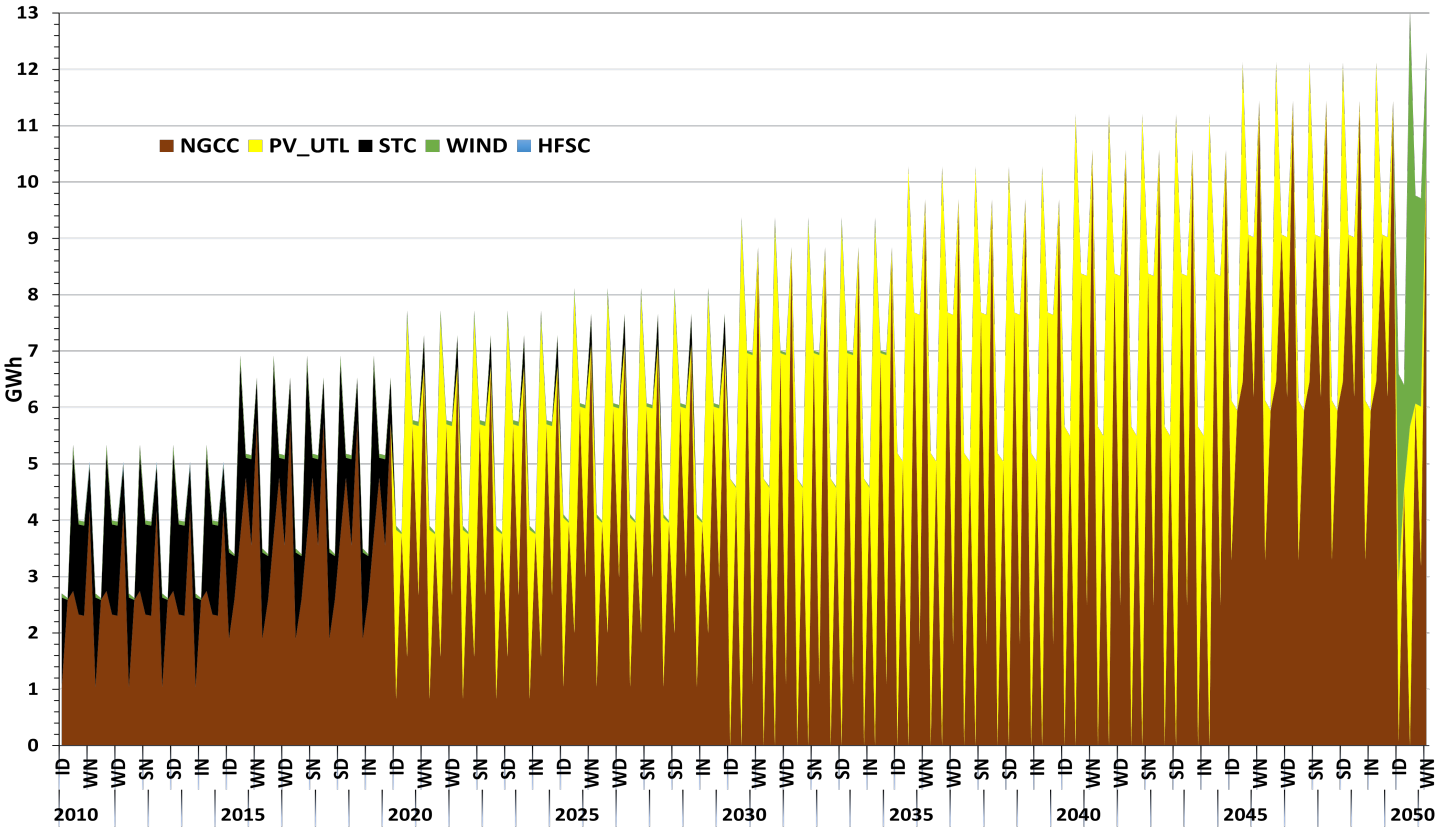


Figure 12: Technology Production by Time Slice (Scenario 2)

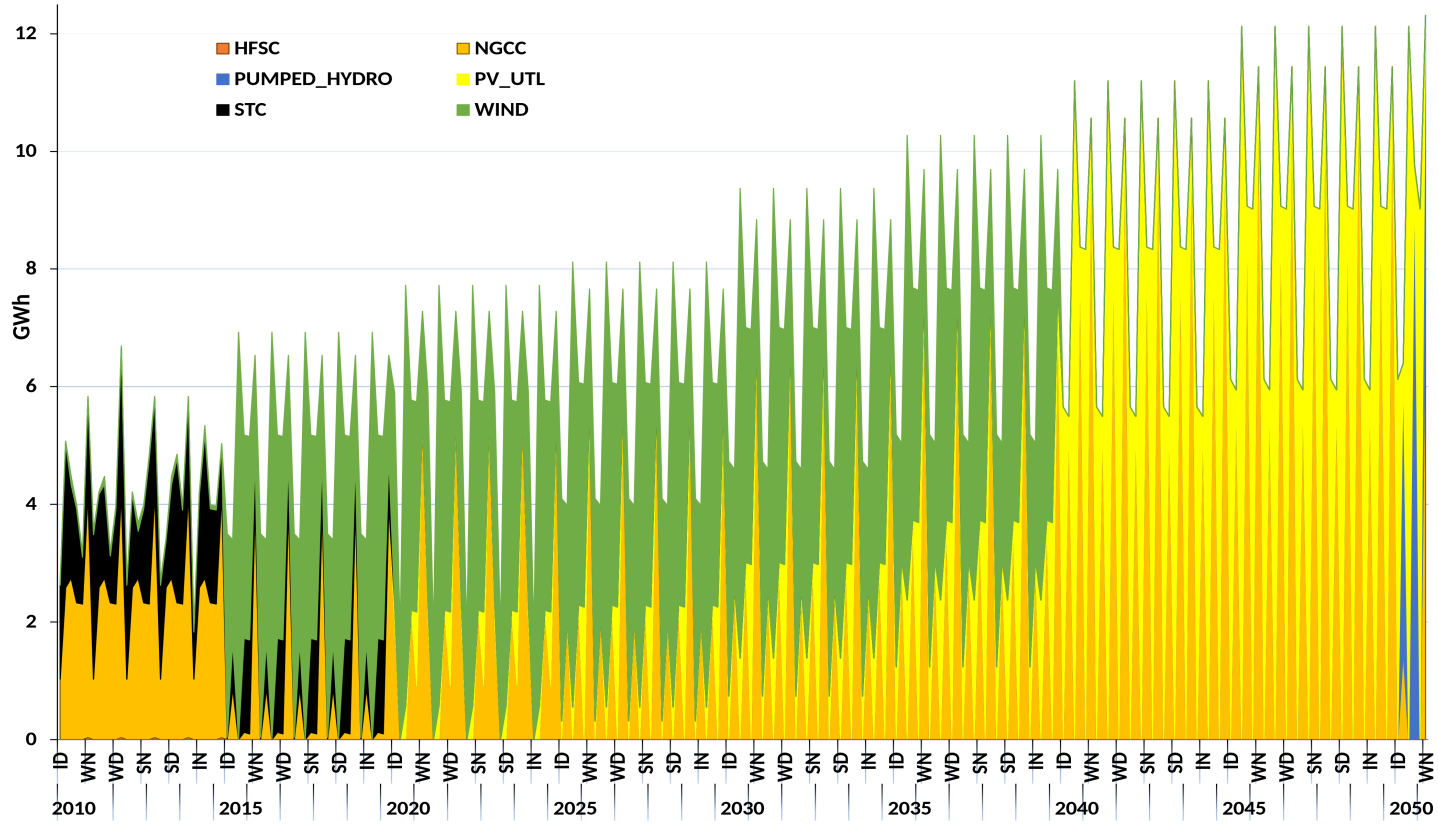


Figure 13: Technology Production by Time Slice (Scenario 3)

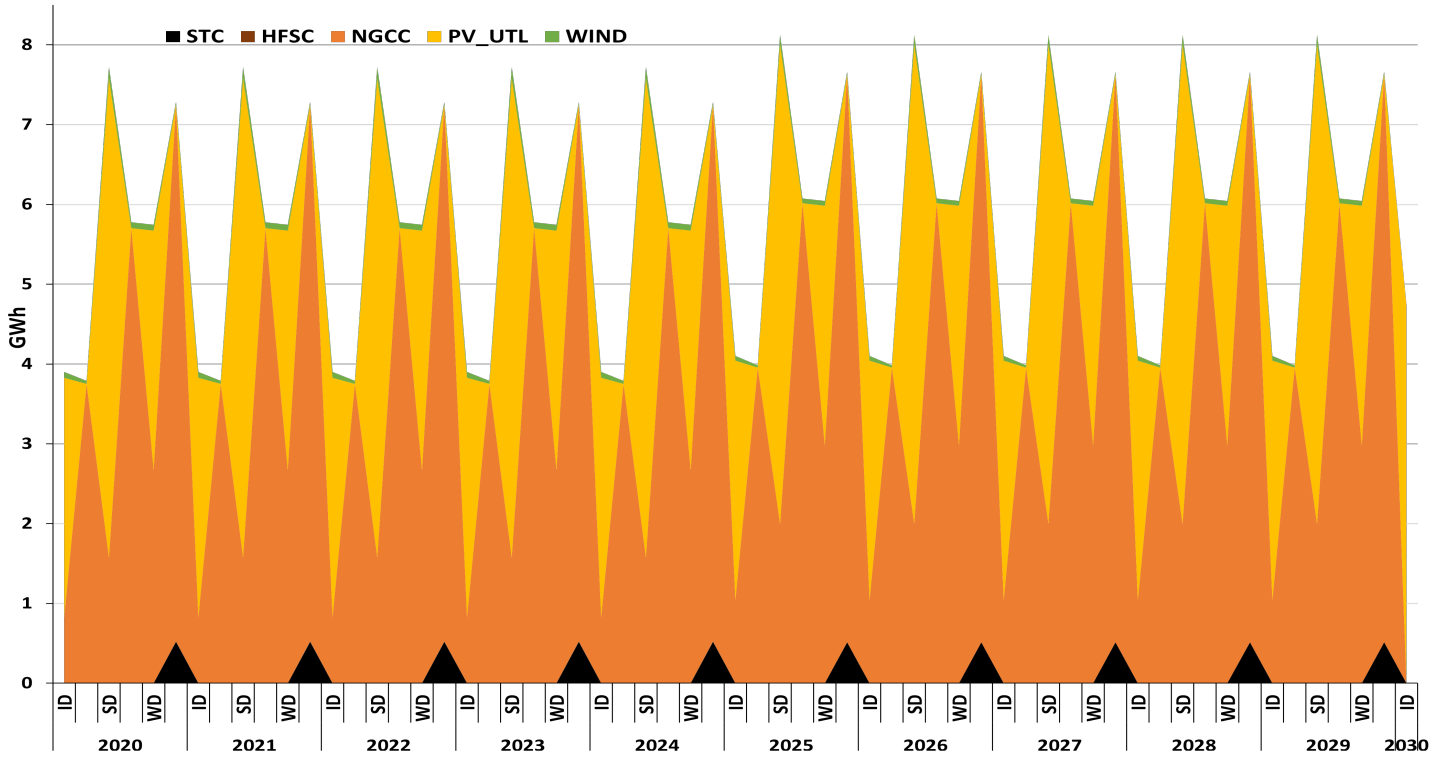


Figure 14: Technology Production by Time Slice 2020-2030 (Scenario 02)

4.2 *Electricity Production Aspects*

Figures 11, 12, and 13 shows how energy output is allocated in three scenarios; We can instantly see that scenario 1 produces substantially more energy than the other two situations. Indeed, scenario 01's total output is 17% higher than scenario 02's and 22% higher than scenario 03's. Although all three scenarios have the same overall demand that must be satisfied, Scenario 01's high use of PHS causes additional demand to be met through the pumping required to fill the quantity tanks. When we look at the breakdown of power generated to fulfill demand between 2010 and 2015, coal and gas come out on top. Second, in Scenarios 02 and 03, conventional generation facilities are predominantly centered on natural gas-fired plants. The second point to note is the variation in generation by time slice in the 02 and 03 scenarios. Because renewable energy sources are intermittent, this was to be expected.

We already mentioned in Scenario 01 that natural gas and coal will be the dominant energy generation sources, as seen in Figure 5. Indeed, the competitiveness of renewable energy in electricity production remains questionable because no minimum production constraint or installed capacity limit is imposed. Wind energy, on the other hand, will become more competitive by 2040 and will naturally be included in the optimal electricity mix. This is attributed to lower wind energy costs and higher generating capacity.

On the modeled days, photovoltaic (PV) generation is the primary contributor to satisfying power demand first thing in the morning in the second scenario. Gas-fired power plants increase output to meet the increased demand for electricity in the evening. Between 2010 and 2020, there will be virtually no visible difference in power generation between conventional and renewable energy sources. However, between 2020 and 2030, the difference between the two will expand slightly. The fundamental cause of this increase is the amplitude of photovoltaic generation, which decreases dramatically when the sun sets. Figure 14, which focuses solely on this period, clearly illustrates this argument. Wind farms will be used to compensate for the decreasing amount of solar power available in the evenings beginning in 2040. As a result, it is less likely that gas-fired power plants will be used to the same extent until the end of the period predicted.

When compared to the conclusion of the period evaluated in Scenario 2, the amount of power generated by NG-PP plants falls at a much steeper and faster pace in Scenario 3. Furthermore, as illustrated in Figure 7, which depicts the distribution of electricity generation by time slice for Scenario 03, the utilization of NG-PP is primarily related to peak load. The latter is mainly used during peak demand or at night to compensate for the fact that renewable energy sources often do not produce at night.

4.3 *Carbon emissions*

Figure 15 exemplifies this notion. The simulation findings reveal that the initial scenario causes the most severe pollution. Compared to the first scenario, CO2 emissions reported for scenarios 2 and 3 were reduced by 25% and 72%, respectively. It is possible to avoid between 62 and 135 Ktones of CO2 emissions. The operation of thermal power plants is a significant source of carbon dioxide emissions. Given that the 1st scenario is dominated by power plants of this sort and that pulverized coal is a significant component of the production mix, it is not unexpected that these emissions are at their highest level. Consider the objectives that the Moroccan government has set for itself. We can see that Scenarios 2nd and 3rd will considerably contribute to the desired carbon intensity reduc-

tion. It would also be interesting to see how the model reacts when environmental limitations are introduced to the basic scenario.

4.4 Total Costs

As shown in Figure 15, the BAU scenario will result in the lowest total expenditure. In this scenario, the cost of fossil fuels represents more than 80% of the total expenditure, taking into account the absence of any intervention on the installed capacity. The total discounted costs of scenario 3 are 2% higher than the total discounted costs of scenario 1. The total discounted cost of Scenario 2 is only 0.5% higher than the total cost of Scenario 1.

The lower fuel prices almost fully compensate for the investments in the second scenario. This also results in a positive energy independence benefit, at a cost almost comparable to that of Scenario 1. However, this result would not have been achieved without the lower fossil fuel prices. In the third scenario, the discounted costs are the highest of the other scenarios. This is because the model has been forced to make large investments in renewable technologies, the relatively high costs.

Scenarios 02 and 03 are distinguished by considerable growth in the use of renewable energy sources. This results in less dependence on fossil fuels and greater self-sufficiency in imports. On the other hand, scenario 01 is characterized by an increased sensitivity to fluctuations in fossil fuel prices, which have been unstable in recent years and which may create a precarious situation in the long term. Therefore, Morocco relies on policymakers to consider that, even if the scenarios with efficient RE integration are slightly more expensive than those without renewable limits, they will reduce the country's dependence on fossil fuels.

Currently, more studies are being conducted to determine the best integration rates for each renewable energy option. This includes photovoltaics and wind power and selecting the best combination of reserve systems. The aim is to manage the effects of these renewables' intermittent nature while maintaining the national grid's stability [35].

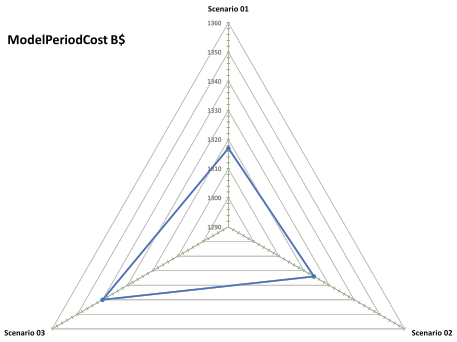


Figure 15: Total discounted cost of the scenarios

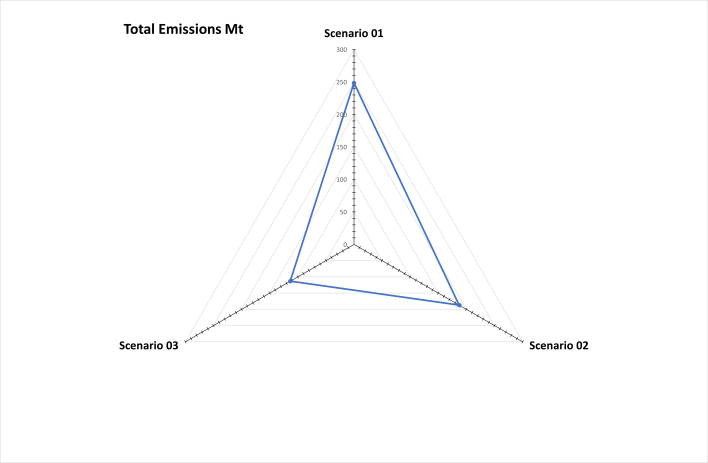


Figure 16: CO2 Emissions by scenario

In view of the results of this study, the following question arises: why has Morocco not created more natural gas power plants and made do with the two power plants of Ain Bnimathar and Tahaddart, which have played an important role in meeting demand and ensuring the stability of the network?

The answer to this question is based on two axes. The first is the absence of this resource on the national territory, while coal has long been extracted from the mines of Jerrada, in the east of the kingdom [36]. The second is the low level of gas infrastructure. Indeed, in our analysis, the infrastructure necessary for the use of natural gas for electricity generation has not been considered, notably in terms of storage capacity and delivery network from the port. This is justified by the fact that the gas-fired power plants in operation so far have benefited from the proximity of the GME (Gazoduc Maghreb Europe) and from the fee in kind paid by Algeria for allowing the use of Moroccan territory for the transport of its natural gas to Europe. All these parameters, as well as others such as the required electrical storage capacity and the optimal storage technology, as well as electricity exchanges with Europe, power to X, and demand management, will be the subject of future analysis.

5 Conclusion

According to the results of this study, the different technological options for electricity generation chosen at the beginning of Morocco's energy transition are not always ideal. Indeed, the decision to develop CSP as the leading renewable source and coal as a backup load source appear controversial. This result leads us to question the primary basis for this choice. This result leads us to question the primary basis for this choice. Actually, at the time, cost reductions of solar PV, derived from its three-decade learning curves, were already far more promising than the modest cost reductions of solar thermal power. These results are from several feedback reports from the Kramer Junction solar thermal power plants in the United States [37].

In addition, while CSP offers enormous storage capacity at acceptable costs, its large-scale growth is limited by the demand for land, making it a centralised power generation technology that requires the creation of a large grid to transmit the output [38].

Furthermore, while coal-fired power plants are an attractive technology for meeting demand and maintaining grid stability, they cannot be used as a back-up technology with a large share of renewables in the electricity mix [39]. This will limit the use of renewable capacity because coal is not very flexible. The third scenario also showed that RE has the potential to become the main source of energy in the Moroccan electricity system. However, for this to happen, they will need to be integrated with flexible generation systems such as gas-fired power plants and energy transfer stations [40, 41].

Besides, as the third scenario shows, increases in renewable energy capacity targets have not always been reflected in total costs. In fact, the total costs remained relatively similar to those in the second scenario. Our results are consistent with the findings of experts in the field. For example, [42] estimated that if Morocco had made the necessary efforts to accelerate the development of power transfer stations, notably the Abdelmoumen project, which was to be commissioned in 2008, it would not have needed as much storage as the CSP plants offer. Moreover, PHS could have been used to compensate for the intermittency of wind power, which is not possible with CSP storage. On the other hand, [42] states that if Morocco had advanced the wind projects, it could have had about 3,000 GWh/year of wind energy, acquired at a price (0.03 USD/kWh). This amount corresponds to the 1050 GWh/year of solar thermodynamic energy currently purchased (at about 0.14 USD/kWh on average).

Further Research Other concerns regarding the energy sector in Morocco will be studied in the near future as part of our ongoing research. A new configuration of the model will be carried out to achieve this objective. This will involve taking into account new parameters, such as the possibility of exporting electricity and the impact that this option will have on the Kingdom's trade balance. In addition, other new parameters will be analyzed, such as the exchange of electricity with neighboring countries. In addition, we will study the storage capacities needed to support the development of renewable energy sources, among others.

Various improvements will also be made to the OSeMOSYS model to analyse concepts such as the stochastic aspect of raw material costs, the fluctuations of RE, as well as the impact of the energy strategy on other sectors such as air quality and environmental preservation, water resource management, or industry and its competitiveness in the context of new carbon taxes introduced and planned by the European market.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This project would not have been possible without the assistance of many people. I'd like to thank my three advisors, Abdeslam KADRANI, Imad EL HARRAKI, and El Hadj EZZAHID, for their invaluable support.I would like to take this opportunity to acknowledge the time and effort devoted by reviewers to improving the quality of this work.

References

[1] J. Slimani, A. Kadrani, I. E. Harraki, E. H. Ezzahid, "Renewable Energy Development in Morocco: Reflections on Optimal Choices through Long-term Bottom-up Modeling," International Conference on Electrical, Computer, and Energy Technologies, ICECET 2021, 2021, doi:10.1109/ICECET52533.2021.9698630.

[2] M. Chentouf, M. Allouch, "Renewable and alternative energy deployment in Morocco and recent developments in the national electricity sector," Open Access Journal Photoenic, **2**(1), 00017, 2018.

[3] R. Madomercandy, "Investigation of The Feasibility of Renewables in Meeting The Moroccan Primary Energy Demand," Napier University, 2007.

[4] B. J. Sellers, J. A. Voltaire, S. M. Gustafson, W. A. Hartman, "Energy Sustainability in Morocco," Energy, 2015.

[5] T. Kousksou, A. Allouhi, M. Belattar, A. Jamil, T. El Rhafiki, A. Arid, Y. Zeraouli, "Renewable energy potential and national policy directions for sustainable development in Morocco," Renewable and Sustainable Energy Reviews, **47**, 46–57, 2015.

[6] A. Haddouche, M. Taoumi, "Turning energy constraints into opportunities: the case of Morocco; Transformer la contrainte energetique en opportunites: le cas du Maroc," Liaison Energie-Francophonie, 2008.

[7] H. Ben Sassi, C. Alaoui, F. Errahimi, N. Es-Sbai, H. B. Sassi, C. Alaoui, F. Errahimi, N. Es-Sbai, "Vehicle-to-grid technology and its suitability for the Moroccan national grid," Journal of Energy Storage, **33**(October), 102023, 2021, doi:10.1016/j.est.2020.102023.

[8] M. Chentouf, M. Allouch, "Assessment of renewable energy transition in Moroccan electricity sector using a system dynamics approach," Environmental Progress and Sustainable Energy, **40**(4), 0–2, 2021, doi:10.1002/ep.13571.

[9] "Stratégie énergétique: Trois ans de retard — L'Economiste," .

[10] LesEco.ma, "Énergies renouvelables : objectif (presque) atteint ! - LesEco.ma," .

[11] "Point de vue : Le Maroc va-t-il réaliser ses objectifs en production d'énergies renouvelables d'ici 2020 ?" .

[12] T. Amegroud, "Morocco's power sector transition: Achievements and potential," 2018.

[13] "Au Maroc, les ratés de la stratégie solaire," .

[14] C. A. Floudas, P. M. Pardalos, Encyclopedia of optimization, Springer Science & Business Media, 2008.

[15] T. Niet, A. Shivakumar, F. Gardumi, W. Usher, E. Williams, M. Howells, "Developing a community of practice around an open source energy modelling tool," Energy Strategy Reviews, **35**, 100650, 2021, doi:10.1016/J.ESR.2021.100650.

[16] M. Howells, H. Rogner, N. Strachan, C. Heaps, H. Huntington, S. Kypreos, A. Hughes, S. Silveira, J. DeCarolis, M. Bazillian, A. Roehrl, "OSeMOSYS: The Open Source Energy Modeling System: An introduction to its ethos, structure and development," Energy Policy, **39**(10), 5850–5870, 2011, doi:10.1016/J.ENPOL.2011.06.033.

[17] M. Welsch, M. Howells, M. Bazilian, J. F. DeCarolis, S. Hermann, H. H. Rogner, "Modelling elements of Smart Grids – Enhancing the OSeMOSYS (Open Source Energy Modelling System) code," Energy, **46**(1), 337–350, 2012, doi:10.1016/J.ENERGY.2012.08.017.

[18] A. Dhakouani, E. Znouda, C. Bouden, "Impacts of energy efficiency policies on the integration of renewable energy," Energy Policy, **133**, 110922, 2019.

[19] "Introduction to OSeMOSYS — OSeMOSYS 0.0.1 documentation," .

[20] ONEE, "Chiffres Clés," 2019.

[21] A. Hermelee, "Regional Reference Energy Systems: electric utility applications," 1979, doi:10.2172/6301105.

[22] ONEE, "Depliant Statistiques FR 2019," 2019.

[23] "Analyse de la demande énergétique à l'horizon 2050," .

[24] "Annual Energy Outlook 2020," .

[25] "Welcome to ENERGYDATA.INFO - ENERGYDATA.INFO," .

[26] "Homepage - U.S. Energy Information Administration (EIA)," .

[27] "Data — Electricity — 2021 — ATB — NREL," .

[28] "Welcome to TAQAMOROCCO — TAQAMOROCCO," .

[29] P. A. Bennouna, "19 Avril 2021," 2021.

[30] I. International Energy Agency, I. E. Agency, "Energy Policies Beyond IEA Countries: Morocco 2019 Review," 2019.

[31] C. Cannone, L. Allington, "Selected 'Starter Kit' energy system modelling data for Morocco (CCG)," 2021, doi:10.21203/RS.3.RS-480023/V2.

[32] B. Amin, "Nul besoin de nouvelles centrales thermiques électriques - ni charbon ni fuel ni gaz naturel," Technical report, 2021, doi:10.13140/RG.2.2.29241.90720.

[33] P. Sullivan, W. Cole, N. Blair, E. Lantz, V. Krishnan, T. Mai, D. Mulcahy, G. Porro, "2015 Standard Scenarios Annual Report: U.S. Electric Sector Scenario Exploration," 2015.

[34] A. Derbali, "Determinants of the performance of Moroccan banks," Journal of Business and Socio-economic Development, **1**(1), 102–117, 2021, doi:10.1108/JBSED-01-2021-0003.

[35] J. Slimani, A. Kadrani, I. El Harraki, E. H. Ezzahid, "Long-Term Wind Power Development in Morocco: Optimality Assessment using Bottom-up Modeling," Proceedings of the 2021 Ural-Siberian Smart Energy Conference, USSEC 2021, 215–220, 2021, doi:10.1109/USSEC53120.2021.9655736.

[36] H. Ez-zaki, A. Diouri, A. Bouregba, F. Amor, S. Chhaiba, O. Sassi, Y. E. Rhaffari, "Ecofriendly bricks elaborated from coal waste of Moroccan Jerrada Mining," MATEC Web of Conferences, **149**, 01043, 2018, doi:10.1051/MATECCONF/201814901043.

[37] C. Philibert, "The present and future use of solar thermal energy as a primary source of energy," International Energy Agency, 1–16, 2005.

[38] M. T. Islam, N. Huda, A. B. Abdullah, R. Saidur, "A comprehensive review of state-of-the-art concentrating solar power (CSP) technologies: Current status and research trends," Renewable and Sustainable Energy Reviews, **91**, 987–1018, 2018, doi:10.1016/J.RSER.2018.04.097.

[39] Y. Zhao, M. Liu, C. Wang, X. Li, D. Chong, J. Yan, "Increasing operational flexibility of supercritical coal-fired power plants by regulating thermal system configuration during transient processes," Applied energy, **228**, 2375–2386, 2018.

[40] M. Simão, H. M. Ramos, "Hybrid Pumped Hydro Storage Energy Solutions towards Wind and PV Integration: Improvement on Flexibility, Reliability and Energy Costs," Water 2020, Vol. 12, Page 2457, **12**(9), 2457, 2020, doi:10.3390/W12092457.

[41] M. A. Gonzalez-Salazar, T. Kirsten, L. Prchlik, "Review of the operational flexibility and emissions of gas- and coal-fired power plants in a future with growing renewables," Renewable and Sustainable Energy Reviews, **82**, 1497–1513, 2018, doi:10.1016/J.RSER.2017.05.278.

[42] "Stratégie énergétique nationale : Interview avec M. Amin BENNOUNA," .

Appendix

```
# Parameters
model.RETagTechnology = Param(model.REGION, model.TECHNOLOGY, model.YEAR, default=0)
model.RETagFuel = Param(model.REGION, model.FUEL, model.YEAR, default=0)
model.REMinProductionTarget = Param(model.REGION, model.YEAR, default=0)
#   Model Variables
model.TotalREProductionAnnual = Var(model.REGION, model.YEAR, initialize=0.0)
model.RETotalProductionOfTargetFuelAnnual = Var(model.REGION, model.YEAR,initialize=0.0)

def FuelProductionByTechnologyAnnual_rule(model, r, t, f, y):
  return (
    sum(
      model.ProductionByTechnology[r, l, t, f, y]
      for l in model.TIMESLICE
    )
    == model.ProductionByTechnologyAnnual[r, t, f, y]
  )

model.FuelProductionByTechnologyAnnual = Constraint(
  model.REGION, model.TECHNOLOGY, model.FUEL, model.YEAR, rule=FuelProductionByTechnologyAnnual_rule
)

def TechIncluded_rule(model, r, y):
  return (
    sum(
      model.ProductionByTechnologyAnnual[r, t, f, y]
      * model.RETagTechnology[r, t, y]
      for t in model.TECHNOLOGY
      for f in model.FUEL
    )
    == model.TotalREProductionAnnual[r, y]
  )


model.TechIncluded = Constraint(
  model.REGION, model.YEAR, rule=TechIncluded_rule
)


def FuelIncluded_rule(model, r, y):
  return (
    sum(
      model.RateOfProduction[r, l, f, y]
      * model.RETagFuel[r, f, y]
      * model.YearSplit[l, y]
      for f in model.FUEL
      for l in model.TIMESLICE
    )
    == model.RETotalProductionOfTargetFuelAnnual[r, y]
  )


model.FuelIncluded = Constraint(
  model.REGION, model.YEAR, rule=FuelIncluded_rule
)


def EnergyConstraint_rule(model, r, y):
  return (
    model.REMinProductionTarget[r, y]
    * model.RETotalProductionOfTargetFuelAnnual[r, y]
    <= model.TotalREProductionAnnual[r, y]
  )


model.EnergyConstraint = Constraint(
  model.REGION, model.YEAR, rule=EnergyConstraint_rule
)
```

Listing 1: Original Renewable energy constraint coding on Pyomo

```
# Parameters
model.PowerTagTechnology = Param(model.REGION, model.TECHNOLOGY, model.YEAR, default=0)
model.REMinCapacityTarget = Param(model.REGION, model.YEAR, default=0)

#   Model Variables
model.TotalRECapacityAnnual = Var(model.REGION, model.YEAR, initialize=0.0)

##################  RE Capacity Target equations  ###########
def PWTechIncluded_rule(model, r, y):
  return(
    sum (
        model.TotalCapacityAnnual[r, t, y]
        * model.PowerTagTechnology[r, t, y]
        for t in model.TECHNOLOGY
       )
      == model.TotalPowerCapacityAnnual[r,  y]
    )
model.PWTechIncluded = Constraint(
  model.REGION, model.YEAR, rule=PWTechIncluded_rule
)

def RETechIncluded_rule(model, r, y):
    return (
      sum(
        model.TotalCapacityAnnual[r, t, y]
        *model.RETagTechnology[r, t, y]
        for t in model.TECHNOLOGY
        )
      == model.TotalRECapacityAnnual[r, y]
    )
model.RETechIncluded = Constraint(
  model.REGION, model.YEAR, rule=RETechIncluded_rule
)

def RECapacityConstraint_rule(model, r, y):
  return (
    model.REMinCapacityTarget[r, y]*
        model.TotalPowerCapacityAnnual[r, y]
    <= model.TotalRECapacityAnnual[r, y]
  )
model.RECapacityConstraint = Constraint(
  model.REGION, model.YEAR, rule=RECapacityConstraint_rule
)
```

Listing 2: New Renewable energy constraint coding on Pyomo

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 3, 139-151 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science and Technology

ASTESJ

ISSN: 2415-6698

Automated Robotic System for Sample Preparation and Measurement of Heavy Metals in Indoor Dust Using Inductively Coupled Plasma Mass Spectrometry (ICP-MS)

Heidi Fleischer[1,*], Sascha Statkevych[2], Janne Widmer[2], Regina Stoll[3], Thomas Roddelkopf[1], Kerstin Thurow[4]

[1]*University of Rostock, Faculty of Computer Science and Electrical Engineering, Institute of Automation, Rostock, 18119, Germany*

[2]*Innerstädtisches Gymnasium (Senior High School), Rostock, 18055, Germany*

[3]*University Medical Center, Institute for Preventive Medicine, Rostock, 18055, Germany*

[4]*Center for Life Science Automation (celisca), Rostock, 18119, Germany*

ARTICLE INFO

Article history:
Received: 05 April, 2022
Accepted: 25 May, 2022
Online: 24 June, 2022

Keywords:
Automated measurement
Compound-oriented measurement
Environmental samples
Heavy metals
ICP-MS
Indoor dust
Robotic system

ABSTRACT

Dust is ubiquitous in our daily environment—outdoor and indoor. In modern times, people often spend the majority of their time at home, in offices, at work or in schools. Suspended particles such as tiny crumbs up to long fibers generate indoor dust deposits. Inhouse sources are the interior releasing abraded fibers from carpets, bedding and clothing as well as the human itself distributing skin cells, lost hairs and food residues. External sources are finest sand, pollen, exhaust particles and microorganisms (e.g., dust mites). An exposure to heavy metals in certain concentrations may affect the human health and may lead to intoxication, allergies or carcinogenic effects. The heavy metals amount in indoor dust depends on the environmental conditions, requiring a sampling with adequate sampling points and numbers. High sampling numbers ensure good coverage of the area to be examined. Therefore, fast and reliable measurement methods for identification and quantification of the elemental composition are needed. To meet these requirements, a robotic system for automated sample preparation and determination of heavy metals in indoor dust using ICP-MS was developed. The values for repeatability, recovery rate, within-laboratory precision, measurement precision and the limits of detection and quantification were determined for both, the manual and the automated process. Furthermore, the individual process steps and times were compared. Besides the processing of certified reference material, inhouse dust samples from different origin were prepared and measured to give a first overview of the inhouse dust composition. The results show, that the robot-assisted system is well-suited for the heavy metal screening in indoor dust.

1. Introduction

This paper is an extension of "Robot-assisted Measurement of Heavy Metals in Indoor Dust Using ICP-MS" presented at the IEEE International Instrumentation and Measurement Technology Conference (I2MTC) 2021 in Glasgow (virtual) [1]. The focus of this study is related to the determination of the elemental composition of dust in our close environment. In the daily life, dust is omnipresent—outdoors and indoors. Especially, the indoor air quality should be focused since their contamination can be higher than outdoors [2, 3]. Today, urban people spend more than 90% of their time indoors—at home, in schools, offices etc. [4]. Furthermore, indoor dust contains suspended particles from internal origins such as smoking, cooking, fuel combustion and decorative material (e.g., paints, carpets, furniture etc.) and external origins as soil, mining, smelting, industrial activities and vehicular emissions [4-6]. And the human itself and pets can be a dust source by distributing food residues, loose hair and skin particles as well as fibrous material from clothes. Both, inorganic and organic contaminants (e.g., biological material, pesticides, PCBs, PAHs, and heavy metals) can be found in indoor dust and may be absorbed and accumulated in the human organism. The heavy metal amount in household dust can be elevated in comparison to outdoor dust such as pavement dust, road dust or garden soil [4, 7]. Heavy metals cannot be degraded. An exposure to individual high concentrations can impact the human health significantly. Allergies, signs of poisoning and cancer diseases may result [2, 4, 8-10]. Several carcinogenic effects to the

*Corresponding Author: Heidi Fleischer, University of Rostock, Institute of Automation, Friedrich-Barnewitz-Strasse 8, 18119 Rostock, Germany, email: Heidi.Fleischer@uni-rostock.de

respiratory organs, the cardiovascular and nervous system as well as to the growth development were reported [2]. Heavy metals can be uptaken into the body in different ways—oral, over the respiratory organs or the skin. Particularly, kids are endangered since they often play on floors [2, 11].

Due to the wide variety of origins, heavy metals cannot completely be avoided in household dust. Therefore, a monitoring of heavy metal concentrations to determine influencing factors and potential health risks is required. A review from the year 2016 gives a comprehensive insight into multiple studies to determine heavy metals in indoor dust [2]. A widespread study from the year 2021 is based on more than 127 articles published between 1985 and 2019 and shows the spatial distribution, sources, and consequences of heavy metal concentrations in indoor dust [4].

The high number of studies in the field of heavy metal determination in indoor dust show an increasing demand for fast and reliable sample preparation and measurement methods able to process multiple samples in a short time frame. Such measurement methods are "compound-oriented measurements" following the "concept of the pre-, intra- and post-sensory selectivity" [12]. In the next section of this publication, this concept is explained regarding the heavy metal determination in household dust samples. Furthermore, an overview of conventional elemental measurement techniques for heavy metal determination as well as of current automation approaches is given. The existing gap and the aim of this study is explained. The system concept and the system design is printed in the third section. The materials and methods are described in the fourth section. The fifth section shows the validation results, a comparison of the automated analytical measurement and the manual process, including the results for the first exemplarily dust samples collected at various inhouse places. A summary and a forecast to future investigations are provided in the last section.

2. Heavy Metal Measurement

2.1. Compound-oriented Measurements

From a metrological perspective, the qualitative and quantitative analysis of heavy metals in indoor dust is assigned to "compound-oriented measurements" [12]. A wide variety of elemental (e.g., heavy metals) and compound species is existing, and no specific single sensors are available for the individual species. Therefore, a generally applicable concept was developed for determination of single elements and chemical compounds in complex mixtures to reach the required selectivity [12]. The "concept of the pre-, intra-, and post-sensory selectivity" is mainly focused to species which are included in a complex matrix causing interfering signals [12]. In our case, the heavy metals to be measured are embedded in dust—particles, crystals, fibers, biologic material, cells etc.

The "pre-sensory selectivity" is reached in the first step of the concept. The heavy metals to be determined must be separated from the matrix dust. This includes sampling, homogenizing by milling and mixing, weighing, a microwave-assisted acid digestion and a final dilution. After these sample preparation steps, the heavy metals were transferred from the solid dust matrix into a liquid measurement solution. This separation step takes place before the sensor system (pre-sensory selectivity). Inductively coupled plasma mass spectrometry (ICP-MS) was used to achieve a high degree of "intra-sensory selectivity". The ICP-MS data must be evaluated, interpreted and visualized in the last step to achieve qualitative (identification of metal species) and quantitative information—the "post-sensory selectivity". Fig. 1 visualizes the individual process steps for heavy metal measurements in dust samples and shows the relationship to the "concept of the pre-, intra- and post-sensory selectivity".

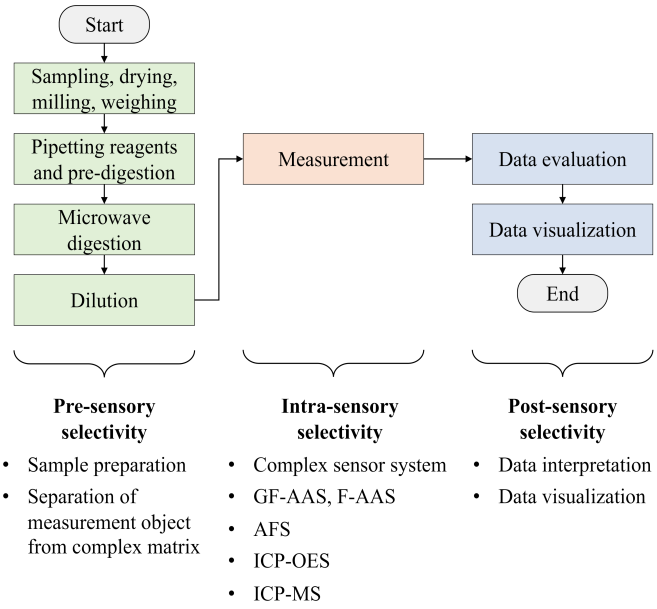


Figure 1: Process steps for heavy metal measurements in dust samples and their relationship to the "concept of the pre-, intra-, and post-sensory selectivity" (abbreviations: GF-AAS—graphite furnace atomic absorption spectroscopy, F-AAS—flame atomic absorption spectroscopy, AFS—atomic fluorescence spectroscopy, ICP-OES—optical emission spectroscopy with inductively coupled plasma, ICP-MS—mass spectrometry with inductively coupled plasma)

2.2. Measurement Methods

The decision for the selection of an analytical measurement technique always depends on the elements (metals) and the aim of investigation [13, 14]. Standardized methods are available from the German Institute for Standardization (Deutsches Institut für Normung e.V., DIN) for the determination of heavy metals in particles of airborne dust, in exhaust gases and atmospheric air [15-17]. The Association of German Engineers (Verein Deutscher Ingenieure, VDI) published standardized methods in German and English for the determination of multiple single elements in particulate matter and in suspended matter in ambient air [18-26]. The solid particles require a special sample preparation before the measurement. Some elements, such as arsenic, antimony, and selenium were separated from the dust matrix into their volatile hydrides for measurement in gaseous form [19]. Mercury can be sorbed into amalgam and then be converted into mercury cold vapor for measurement [25]. For the most elements, such as heavy metals, a digestion is required to transfer them from the solid matrix into a liquid form. An acid digestion is a common procedure performed in different ways such as open, thermal or microwave-assisted digestions using a wide variety of acids or acid mixtures (e.g., hydrofluoric acid, aqua regia or nitric acid with hydrogen peroxide) [13, 23]. The qualitative and quantitative determination of heavy metals is mostly performed using optical spectroscopy

such as graphite-furnace atomic absorption spectroscopy (GF-AAS), flame atomic absorption spectroscopy (F-AAS), atomic fluorescence spectroscopy (AFS), and optical emission spectroscopy with inductively coupled plasma (ICP-OES). Mass spectrometry (ICP-MS) has gained more and more importance in the last decades and have been established for heavy metal measurements [15-26].

2.3. Automated Sample Preparation and Measurement

Today, screening systems show a high degree of automation in the fields of biotechnology, pharmacy and high-throughput applications. In contrast, the automation degree in the field of compound-oriented measurements is still low. Multiple—and often changing—process steps with complex sub-processes, single vessels varying in size and design, harsh processing environments (e.g., high pressures, high temperatures, toxic reagents, high-concentrated acids, organic solvents etc.) and special laboratory devices such as heating plates, coolers, incubators, shakers, centrifuges etc. must be integrated into the automation system. A further challenge is the flexibility—to run multiple applications at one automation system to make it more economically for analytical laboratories.

Only few automation solutions have been established in elemental measurements. Typically, the existing automation systems are often specialized to an individual application. Mercury analyzers are one example for automated sample preparation, sample introduction and measurement using AAS and AFS—sometimes coupled with the cold vapor technique. The advanced mercury analyzer AMA 254 (LECO company, Geleen, Netherlands) was designed for the determination of mercury trace amounts in liquid and solid samples [6]. It uses a technique of direct combustion to decompose the sample in an oxygen-rich environment and to remove interfering elements. The mercury amount is determined using a gold amalgamator trap and a spectrophotometer [27]. The advantage is a very sensitive mercury determination in different samples (solids and liquids) without additional chemical sample preparation [6]. Another automated mercury analyzer is the DMA-80 L (MLS-MWS Laboratory Solutions, Leutkirch im Allgäu, Germany) which uses the cold vapor technique [28]. Both devices are limited to the element mercury and therefore limited in their flexibility.

ICP-OES and ICP-MS are predestinated for measurement of multiple elements in a short time frame, since they allow simultaneous determination of numerous elements. In general, those instruments process samples in liquid and gaseous state. The measurement of solid samples typically requires a previous sample pretreatment, such as a microwave-assisted acid digestion. One example of a commercial automation system is the MiniLab (Rohasys Robotic and Handling Systems & Rouwette Consultancy, Spaubeek, Netherlands) for the automated sample preparation of solid samples using an aqua regia digestion [29]. Mercury and other heavy metals can then be measured using ICP-OES or ICP-MS. Another example for automation in elemental analysis is the implementation of flow-injection systems for automated sample introduction and adding of the internal standard prior to the measurement using ICP-OES [30]. This automation solution significantly increases the instrument's performance and precision and is today integrated into the majority of commercial ICP-MS and ICP-OES instruments.

The automation solutions mentioned before are limited in their flexibility—the systems are designed for a special functionality or for a special application. The aim of the authors of this study is the development of flexible automation systems for "compound-oriented measurements", which are not limited to a special application and which can be applied to elemental and structural measurement tasks. A first version of the robotic system was introduced for the sample preparation of waste wood samples and the determination of mercury using ICP-OES and ICP-MS [31-34]. The automation system was extended for the preparation of animal and human bone samples for the determination of calcium and phosphor using ICP-MS to support the recent osteoporosis research [35]. For a further enhancement of the methodological scope, the system was modified and extended to perform the sample preparation for both elemental and structural measurements. Incrustations of clogged biliary endoprostheses as well as pig bile were automatically prepared for the determination of trace metals using ICP-MS and for the determination of cholesterol using GC/MS to support the development of more compatible biliary endoprostheses [36-39]. The robotic system was recently extended to perform the sample preparation of human tissue samples for subsequent heavy metal measurements to support arthrosis treatment by joint implants [40, 41].

This study is focused on the investigation of the close indoor environment. The robotic system was expanded to include an automated screening setup for heavy metal determination in indoor dust. New labware was designed and integrated. New sample processing procedures for the individual automation stations and process control methods were developed. The method development was challenging since common sample preparation procedures for dust samples include a digestion using hydrofluoric acid—recommended for certain heavy metals [13, 14, 42, 43]. Special trained staff as well as special safety devices for the laboratory equipment and instruments are therefore required. In this study, an automated screening method for the determination of heavy metals in indoor dust samples was developed as previously presented [1, 44]. This includes the development of the manual sample preparation method and ICP-MS measurement method [1, 44]. The sample pretreatment (microwave-assisted digestion) was done with nitric acid and hydrogen peroxide; the use of hydrofluoric acid was avoided. Samples with conspicuous results acquired in the screening can then be confirmed by the manual standard method with hydrofluoric acid. As a result, overall processing times and financial resources can be reduced. Furthermore, this study presents a very flexible robot-assisted sample preparation system, which is adaptable to multiple application areas. This promising automation concept enables small and middle analytical laboratories to install such a robotic system and to use it for different tasks, increasing the overall laboratories' efficiency, economy, and safety.

3. System Concept and System Design

3.1. Robotic Sample Preparation System

The robotic sample preparation system was designed according to the automation concept of the "central system integrator" with a decentralized, open structure [45, 46]. For connecting the

individual automation stations (e.g., storage system, different liquid handling devices etc.) two ORCA laboratory robots (Beckman Coulter, Krefeld, Germany) were integrated. These two articulated robots move on orthogonal linear rails and transfer labware with samples and reagents between the stations. A labware transfer station is mounted at the crossing area of the linear rails, enabling the take-over of labware from one robot to the other one. A barcode reader is included for identification of the labware. Two further stations are used for liquid handling and dosing. The Biomek 2000 (Beckman Coulter, Krefeld, Germany) is a liquid handling station for dosing liquids in volumes up to 1,000 µL. The single-vial liquid handler is used for higher liquid volumes (max. 10 mL). Especially in dilution tasks, this large-volume dispenser reduces pipetting steps and processing time. The automation system was designed in a decentralized structure distributed over three laboratories. The microwave digestion device Mars 5 (CEM, Kamp-Lintfort, Germany) is located in a separate laboratory close to the robotic system. The measurement instrument ICP-MS 7700x (Agilent Technologies, Waldbronn, Germany) is installed in another laboratory. The automation system and the additional instruments are equipped with a labware transfer station. These stations allow the access by both—by laboratory assistants and alternatively by mobile robots. The storage system consists of 196 positions for labware in the standard microplate footprint to provide chemicals, standards and labware required for the process run. Figure 2 shows the robotic sample preparation system as a 3D CAD drawing.

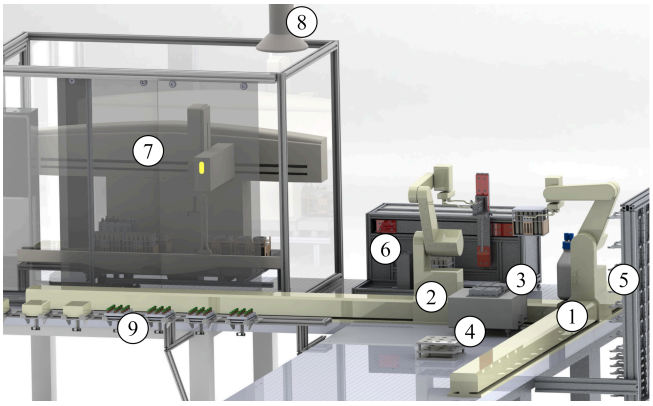
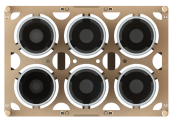
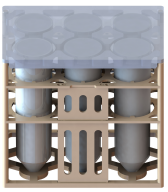
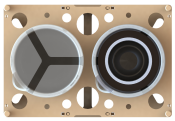
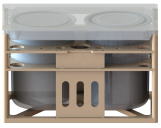
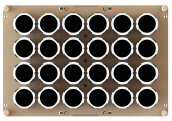
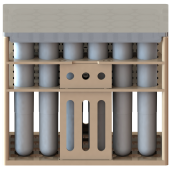


Figure 2: Robotic system for automated sample preparation—laboratory robots moving on linear rail (1) in front of the storage system and (2) in front of liquid handling stations, (3) labware transfer station between two robots, (4) additional rack and thermo shaker on workbench, (5) storage system for chemicals and labware, (6) single-vial liquid handler, (7) liquid handler Biomek 2000 with safety housing and, (8) exhaust system, (9) labware transfer station to laboratory staff or to mobile robots

In general, the labware required in sample preparation tasks for analytical measurements (e.g., microwave vessels, autosampler vials) are not designed in the standard microplate format. Several single vessels, beakers, flasks, and vials with different designs, volumes, lids, and screw caps are required. To ensure a safe labware handling by the ORCA laboratory robots, several racks were designed in the standard microplate format to arrange multiple single vessels. For the sample preparation process of dust samples four racks each with six microwave digestion vessels Xpress (vol. 25 mL, CEM, Kamp-Lintfort, Germany) and one solid lid (cover) for simultaneous covering all six vessels were constructed. These vessels are used to store the powdery samples and for the microwave digestion. Additional covers with six holes were constructed to reduce the evaporation of acid fumes and nitrous gases during the pipetting steps after the microwave digestion [1]. The digestion reagents are provided in another rack with two beakers (PFA, vol. 100 mL) for nitric acid and hydrogen peroxide (the beaker contains a smaller insert with a lower volume of 25 mL). The final—with ultrapure water diluted—measurement solutions are stored in 24 tubes (PP, vol. 14 mL) arranged at a further rack with a solid lid. Table 1 gives an overview of the labware arranged in the special designed racks.

Table 1: Special designed racks in microplate footprint (aluminum body, chemical resistant lid made of PTFE) enabling robotic handling of different numbers of single vessels with various volumes and shapes

Description	CAD drawing	
	unlidded (top view)	closed (front view)
6 microwave vessels Xpress (vol. 25 mL)		
2 beakers for storing nitric acid (left: vol. 100 mL) and hydrogen peroxide (right: insert with 25 mL)		
24 tubes for diluted samples/measurement solutions (vol. 14 mL)		

3.2. Multi-level Software System

A multi-level software system is required to schedule and to control multiple automated workflows supported by robotic systems, on measurement instrumentation as well as in interaction with human operators and mobile transportation robots [45]. The overall process with individual subprocesses—sample preparation of dust specimens including robot-assisted liquid handling and microwave digestion, the ICP-MS measurements, transportation steps as well as data processing and visualization—is managed by the high-level workflow management system [47]. The subprocesses supported by robots were controlled by the process control software "SAMI Workstation Ex 4.0" (Beckman Coulter, Krefeld, Germany) representing the middle-level software. The human operator defines the type of the labware, their source position and their destination. Furthermore, the individual substations (e.g., liquid handling) were selected. Finally, the entire workflow of each subprocess is saved in a SAMI method file, which will be executed in the process run. The lowest level of the software system is formed by the device software of the individual substations. The Biomek 2000 software (Beckman Coulter, Krefeld, Germany) is used to control the liquid handler Biomek 2000. The human operator defines the deck layout as well as the detailed pipetting parameters and saves them in a Biomek method

file. The single vial-liquid handler is controlled by an inhouse-developed software module [45].

3.3. Automation

The entire automated process of dust sample preparation and measurement is divided in multiple subprocesses. In the first subprocess "Pre-digestion", the reagents and samples were transported by the ORCA laboratory robots to the liquid handler Biomek 2000. The reagents are added to the solid samples and the pre-digestion over 20 min is performed using the Biomek 2000. In the second subprocess "Microwave digestion" a human operator closes the microwave vessels, starts the automated digestion run and opens the vessels after cooling down. The third subprocess "Dispensing" is carried out in parallel the robotic system. The ORCA laboratory robots transport empty vials to the single-vial liquid handler and ultrapure water is automatically dispensed. Finally, one ORCA laboratory robot transports the rack with water to the Biomek 2000. In the fourth subprocess "Dilution", the digestion solutions were pipetted into the vials—previously filled with water—and mixed the liquids using the Biomek 2000. After this, the samples are ready for the subprocesses "Measurement" and "Data Evaluation and Visualization". Two top processes were created—for simultaneously processing of 12 and alternatively 24 samples. A more detailed description of the automated subprocesses is given in section 5.2.

4. Materials and Methods

4.1. Standards and Chemicals

The certified reference material (CRM) of indoor dust SRM 2583 (NIST, Gaithersburg (MD), United States) was purchased from Sigma Aldrich/Merck (Darmstadt, Germany). Nitric acid (HNO_3) and 30% stabilized hydrogen peroxide (H_2O_2)—both in suprapure quality (Rotipuran®)—were purchased from Carl-Roth (Karlsruhe, Germany). The ICP multi-element standard IV and ICP single-element standards for As, Hg, Lu and Re were from Merck (Darmstadt, Germany).

4.2. Sample Preparation Using Microwave Digestion

The certified reference material SRM 2583 (fine dust powder) was directly weighted (30 mg) into a microwave digestion vessel Xpress with a volume of 25 mL (CEM, Kamp-Lintfort, Germany). Dust samples—collected by the authors at different inhouse locations—were previously dried and pulverized using an oscillating mill MM2000 (Retsch, Haan, Germany). Three stirring balls (PTFE, diameter 6 mm) from Bola (Grünsfeld, Germany) were added to each sample in a microwave digestion vessel. Nitric acid was used as digestion reagent. To correct evaporation effects during the microwave digestion, Rhenium (Re) was added to the nitric acid as internal standard (ISTD) with a concentration of 1.25 mg/L (after sample processing 80 µg/L). A volume of 2.4 mL of HNO_3 with ISTD and 0.6 mL of H_2O_2 were pipetted to the powdery samples. The vessels were closed after a rest of 20 minutes without lids. The Mars 5 device (CEM, Kamp-Lintfort, Germany) is used for the microwave digestion with the following temperature-time program: room temperature to 180 °C (356 °F) in 20 min, 180 °C to 220 °C (428 °F) in 20 min and a hold time at 220 °C (428 °F) for 20 min. After cooling down and opening the vessels, the samples were diluted with ultrapure water.

4.3. ICP-MS Measurement Parameters

The ICP-MS 7700x was operated with the autosampler ASX-500 (Cetac, Omaha (NE), United States) for automated sample introduction. The data were acquired using the MassHunter Workstation Software for 7700 ICP-MS G7201C Version C.01.04 (Agilent Technologies, Waldbronn, Germany) and the data interpretation and statistical evaluations were done using the in-house developed software module "Analytical Data Evaluation" (ADE)—a web application running at all operating systems [48, 49].

The calibration was generated with five standard solutions (concentrations 1, 5, 10, 50 and 100 µg/L for Li, Be, V, Cr, Co, Ni, As, Sr, Ag, Cd, Tl, and Pb). Mercury was added to the respective standard solution with 0.01, 0.05, 0.10, 0.50 and 1.00 µg/L. The ICP-MS was operated with the following parameters: 1,550 W RF power, 10 mm sample depth, 1.05 L/min nebulizer gas flow, 0.10 rps nebulizer pump speed, 13 °C (55.4 °C) spray chamber temperature, argon as plasma and nebulizer gas, and 4.3 mL/min collision cell helium flow. The data were acquired with three repetitions (0.3 sec integration time). The autosampler probe was flushed after each measurement first with HNO_3/HCl (5%/1%) followed by HNO_3 (8%). After the last measurement, the probe was finally flushed with water.

5. Results and Discussion

5.1. Manual Sample Preparation and Measurement

The manual sample preparation and measurement method was validated before the transfer to the automation system. The validation was focused on several parameters [33, 36, 45]. The repeatability and the recovery rate were determined with 23 standard samples of the certified reference material (CRM) NIST SRM 2583 and one blank—prepared and measured at one day. The within-laboratory precision was determined with ten daily prepared and measured CRM samples—this procedure was performed on five different days. The measurement precision of the ICP-MS was determined by measuring one sample ten times. The limit of detection (LOD) and the limit of quantification (LOQ) were established by measuring ten blanks samples, which previously gone through the whole sample preparation process. The validation results were previously published [44].

5.2. Automated Sample Preparation and Measurement

In the first subprocess "Pre-digestion" the racks providing reagents in beakers and solid samples in microwave digestions vessels were transported from the storage system to the deck of the liquid handling station Biomek 2000 using the two ORCA laboratory robots and the robot transfer station. At the Biomek 2000, the reagents (nitric acid and hydrogen peroxide) were added to the samples. After a rest over 20 min with open vessels, the racks were covered and provided on the sample transfer station. The second subprocess "Microwave digestion" requires the support of a human operator for closing the individual microwave vessels and starting the microwave digestion run. After cooling down, the screw caps of the vessels were manually removed, the vessels arranged in the racks and the racks lidded with a special perforated cover [1].Then the racks were directly provided on the deck of the Biomek 2000. In parallel processing to the microwave digestion, the third subprocess "Dispensing" is carried out using the robotic

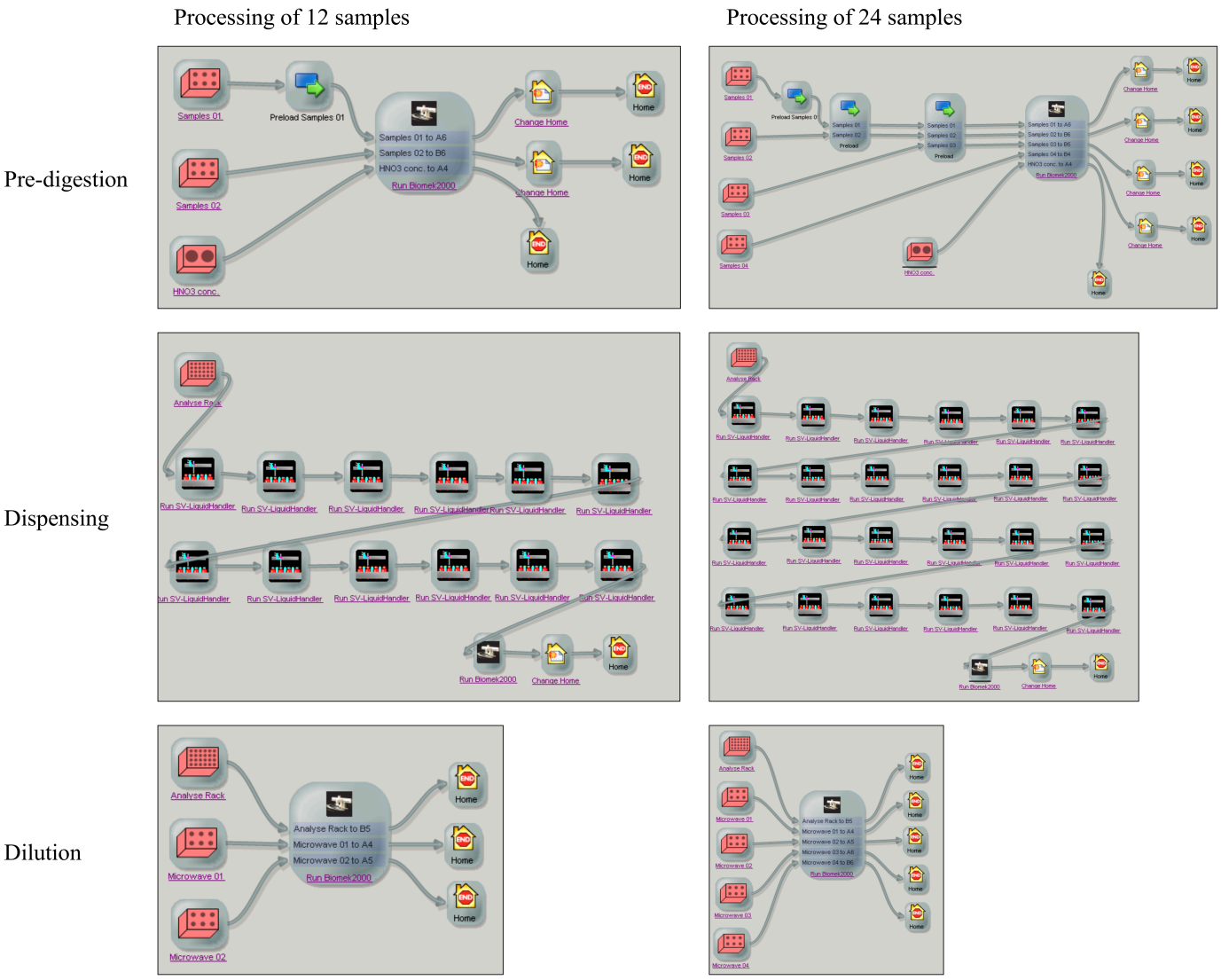


Figure 3: SAMI Ex Editor with subprocesses of the automated preparation of 12 samples arranged on two racks (left) and 24 samples arranged on four racks (right)—Pre-digestion, dispensing and dilution

system. First, the lidded rack with empty vials is transported from the storage system to the deck of the single-vial liquid handler station using the ORCA laboratory robots. The lid is removed by one ORCA laboratory robot, ultrapure water automatically dispensed into the empty vials and the rack is covered again by the robot. Finally, the rack with the water-filled vials is transported by the robot to the deck of the liquid handler Biomek 2000. Figure 3 shows the SAMI methods for the preparation of 12 and 24 samples.

The digestion solutions were diluted with ultrapure water in the fourth subprocess "Dilution" to adjust the acid concentration for the subsequent measurement. The digestion solutions were aspirated through the perforated lids and dispensed into the water-filled vials. The liquids were well mixed by multiple aspiration/dispensing. After this, the rack is lidded and the samples are ready for the subprocesses "Measurement" and "Data Evaluation and Visualization".

The rack with the measurement solutions is transported to the ICP-MS, which is located in a separate laboratory. The transport can be performed either manually or automated using mobile robots [50].

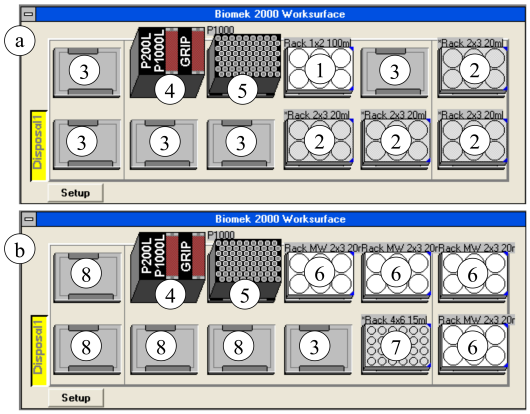


Figure 4: Deck layout of the Biomek 2000 for processing 24 dust samples—(a) pre-digestion, (b) dilution, (1) rack with nitric acid and hydrogen peroxide, (2) lidded racks with samples, (3) free positions for lids, (4) tool rack, (5) tip box, (6) racks with digestion solutions and perforated lids, (7) water filled tubes for measurement solutions, (8) free positions

The sample rack is positioned into the ICP-MS autosampler and the automated sample introduction and measurements are

started. After finishing, the data were automatically evaluated [49]. Figure 4 shows the deck layout of the liquid handler Biomek 2000 for the subprocesses "Pre-digestion" and "Dilution" for the processing of 24 dust samples.

5.3. Validation of the Automated Process

The validation parameters were determined similar to the manual method validation. The repeatability and the recovery rate were determined with 23 CRM samples and one blank arranged in four racks. The repeatability showed coefficients of variation (CV) for the elements chromium, arsenic, cadmium, and mercury which were in good accordance to the certified values. The CV values were slightly higher (+0.26%) for lead. The average recovery rates were determined for arsenic with 93.9%, for cadmium with 82.9% and for lead with 82.04%. The average recovery rates of chromium and mercury were lower, with values of 56.3% (Cr) and 66.4% (Hg). To increase the recovery rates, the use of a stronger acid mixture with hydrofluoric acid in the microwave digestion is recommended [13, 14, 42, 43]. In this automation approach, a hydrofluoric acid digestion was avoided for safety reasons. The automated process is performed on a technical system with mechanical and electronical components sensitive to corrosive acids. The technical equipment and the laboratory staff must be protected, and additional safety equipment would be required.

The within-laboratory precision was determined with 11 CRM samples and one blank prepared at five days. Two racks were used in each method run. The resulting CV values are in good accordance to the certified values for Cr, As, and Cd. The concentration uncertainties of the CRM were given for chromium (±27.5%) arsenic (±22.9%) and cadmium (±50.7%). The CV values for the lead concentration were slightly higher than certified. The method presented is a rapid screening method—therefore, a CV value lower than 15% can be accepted. The measurement precision—determined using ten measurements of one sample—showed results in the expected range (<2%). The limits of detection and quantification were determined with the preparation and measurement of 10 blank samples. The LOD values were determined for the measurement solutions with 180.7 ng/L (Cr), 10.2 ng/L (As), 3.7 ng/L (Cd), 6.5 ng/L (Hg), and 14.5 ng/L (Pb) and for the solid dust material with 225.9 µg/kg (Cr), 12.7 µg/kg (As), 4.6 µg/kg (Cd), 8.1 µg/kg (Hg), and 18.2 µg/kg (Pb). The LOQ values were determined for the measurement solutions with 423.2 ng/L (Cr), 25.8 ng/L (As), 9.2 ng/L (Cd), 12.1 ng/L (Hg), and 27.3 ng/L (Pb) and for the solid dust material with 529.0 µg/kg (Cr), 32.2 µg/kg (As), 11.6 µg/kg (Cd), 15.2 µg/kg (Hg), and 34.2 µg/kg (Pb). The validation results for the certified elements Cr, As, Cd, Hg, and Pb are summarized in Table 2.

In this extended version of the study, the results of additional elements are presented (Li, B, V, Co, Ni, Sr, Ag, and Tl). The average concentrations range from 0.045 mg/kg (Tl) to 65.25 mg/kg (Sr). The repeatability showed coefficients of variation ranging from 3.97% (Li) to 13.28% (Ni). Only Ag has a higher CV value of 36.75%. Certified values for these additional elements are not available for the CRM used in this study. For this reason, the values provided have merely informative character. The precision values of the non-certified elements were similar to the certified elements, and this shows the ability of the automated method to measure more than the certified elements. Additional determination of the recovery rate would be required to confirm the true value.

The measurement precision was determined with 10 repetition measurements of the same sample. The LOD and LOQ values were determined with 10 blank samples for both the measurement solutions and the solid dust material. The entire validation results for the non-certified elements are also included in Table 2.

5.4. Comparison of Manual and Automated Measurements and Literature Values

Comparison of the process steps: A volume of 2.4 mL HNO_3 is given to the powdery samples. Only one step is needed in manual processing. The liquid handler Biomek 2000 has a maximum pipetting volume of 1 mL. For this reason, three steps each with 0.8 mL are required to add the digestion acid. After the microwave digestion, a volume of 2 mL sample solution is manually diluted with 23 mL water. Therefore, water is dispensed in three steps with volumes of 10 mL, 10 mL, and 3 mL. The sample solution is then added in one step and mixed with the water. This dilution step was miniaturized in the automated procedure. The single-vial liquid handler dispenses a volume of 5.75 mL water in one step. Then a volume of 0.5 mL sample solution is added to the water and mixed by the Biomek 2000. This miniaturization has an additional advantage. Due to the smaller total volume of the final measurement solutions—vials with a volume of 30 mL were reduced to vials with 14 mL—more vials can be arranged on one rack.

Only six large volume vials can be placed on one rack, but 24 low-volume vials. This allows a higher sample throughput due to the reduced transportation steps for multiple racks. Table 3 summarizes the numbers of manual and automated process steps.

Comparison of the processing times: The entire processing times of the manual and automated processing were determined with the preparation of 12 samples. The duration of the automated measurement process is 15.5 min longer than the manual procedure executed with trained laboratory staff. Typically, the human operator arranges all required equipment on the workbench. Special transportation steps between a storage system and individual liquid handling stations are not required.

Furthermore, the tool change process of the liquid handler Biomek 2000 (e.g., from gripper tool for lid handling to pipetting tool and return) needs additional time. A human uses it hands for lid handling and operating the pipette. This time delay—between manual and automated operation—can be avoided by a suitable process scheduling with parallel and overlapping runs. Nevertheless, the laboratory assistant is free in the time of the automated run and can perform other tasks in parallel. Furthermore, the automated process can run during the staff's break times. In summary, despite the slightly longer processing time of the automated process, the laboratories efficiency and the throughput can be increased. Table 4 compares the processing times for the individual process steps and for the entire process as well as the operators in the manual and the automated procedure.

Table 2: Validation results of the automated measurement method in comparison to the certified values of the indoor dust CRM NIST SRM 2583

	^{7}Li	^{9}Be	^{51}V	^{52}Cr	^{59}Co	^{60}Ni	^{75}As	^{88}Sr	^{107}Ag	^{111}Cd	^{202}Hg	^{205}Tl	^{208}Pb
Repeatability (n=23)													
Average [mg/kg]	12.48	0.23	13.13	45.02	3.73	44.41	6.57	65.25	1.13	6.05	1.04	0.045	70.48
STDEV [mg/kg]	0.49	0.02	0.54	3.09	0.33	5.90	0.26	5.02	0.41	2.15	0.11	0.004	6.11
CV [%]	3.97	7.24	4.09	6.87	8.76	13.28	3.94	7.70	36.75	35.45	10.47	7.93	8.66
Recovery rate (n=23)													
Average [%]	n.a.	n.a.	n.a.	56.27	n.a.	n.a.	93.90	n.a.	n.a.	82.89	66.44	n.a.	82.04
Min [%]	n.a.	n.a.	n.a.	48.55	n.a.	n.a.	86.97	n.a.	n.a.	56.11	60.58	n.a.	71.73
Max [%]	n.a.	n.a.	n.a.	64.84	n.a.	n.a.	100.33	n.a.	n.a.	171.37	94.01	n.a.	99.73
Within-laboratory precision (n=11, 5 days)													
Max. average [mg/kg]	14.98	0.29	16.15	56.18	4.46	54.10	7.42	81.51	1.29	8.31	1.43	0.061	79.15
Min. average [mg/kg]	13.90	0.25	15.07	50.64	3.87	45.85	6.74	70.69	0.94	6.10	0.92	0.056	68.68
Max. CV [%]	11.29	11.40	11.47	14.18	28.32	30.78	10.28	19.12	40.70	58.92	49.39	10.30	14.69
Min. CV [%]	1.85	5.37	3.16	5.77	4.88	14.65	4.04	1.17	24.50	21.88	5.26	5.20	5.70
Measurement precision (n=1, 10 measurements)													
CV [%]	0.59	5.41	0.94	0.66	0.72	0.77	1.16	0.46	0.74	0.80	1.99	2.58	0.41
Analytical LOD and LOQ (in measurement solution)													
LOD [ng/L]	28.1	3.8	121.5	180.7	6.3	109.4	10.2	178.6	11.7	3.7	6.5	9.8	14.5
LOQ [ng/L]	45.7	9.1	152.5	423.2	13.6	253.9	25.8	418.3	20.8	9.2	12.1	14.2	27.3
Methodical LOD and LOQ (in solid dust samples)													
LOD [µg/kg]	35.1	4.8	151.8	225.9	7.9	136.7	12.7	223.3	14.7	4.6	8.1	12.3	18.2
LOQ [µg/kg]	57.1	11.4	190.6	529.0	16.9	317.4	32.2	522.8	26.0	11.6	15.2	17.8	34.2
Certified values of NIST SRM 2583													
Average [mg/kg]	n.a.	n.a.	n.a.	80	n.a.	n.a.	7	n.a.	n.a.	7.3	1.56	n.a.	85.9
Uncertainty [mg/kg]	n.a.	n.a.	n.a.	±22	n.a.	n.a.	±1.6	n.a.	n.a.	±3.7	±0.19	n.a.	±7.2
Uncertainty [%]	n.a.	n.a.	n.a.	±27.5	n.a.	n.a.	±22.9	n.a.	n.a.	±50.7	±12.80	n.a.	±8.4

Table 3: Manual and automated process steps (1 sample) [1]

Process Step	Manual		Automated	
	Volume [mL]	Steps	Volume [mL]	Steps
Pipetting HNO_3	2.40	1	0.80 0.80 0.80	3
Pipetting H_2O_2	0.60	1	0.60	1
Dispensing H_2O	10.00 10.00 3.00	3	5.75	1
Pipetting digested sample solution	2.00	1	0.50	1

Table 4: Processing times and operators in manual and automated sample preparation and measurement (12 samples) [1]

Process step	Manually		Automated	
	Processing time [min]	Operator	Processing time [min]	Operator
Initial transport (labware, reagents)	2.0	Human laboratory assistant	3.5	2 ORCA laboratory robots
Pipetting HNO_3	1.0	Human laboratory assistant	8.5	Biomek 2000 liquid handler

Process step	Manually		Automated	
	Processing time [min]	Operator	Processing time [min]	Operator
Pipetting H_2O_2	1.0	Human laboratory assistant	2.0	Biomek 2000 liquid handler
Pre-digestion	20.0	Waiting time	20.0	Waiting time
Closing vessels (manually)	2.0	Human laboratory assistant	2.0	Human laboratory assistant
Microwave digestion, cool down	90.0	Microwave device	90.0	Microwave device
Opening vessels (manually)	4.0	Human laboratory assistant	4.0	Human laboratory assistant
Dispensing water (performed in parallel to microwave digestion)	(3.5) not included in calculation	Human laboratory assistant	(6.5) not included in calculation	Single-vial liquid handler
Pipetting of digestion solution, mixing	4.5	Human laboratory assistant	10.0	Biomek 2000 liquid handler
ICP-MS measurement (12 samples)	54.0	ICP-MS equipped with autosampler	54.0	ICP-MS equipped with autosampler
Total processing time	178.5		194.0	

Table 5: Measurement results of dust samples collected in laboratories and offices (samples 1-7, 12), in private households (samples 8-11, 21), in old disused garden houses (samples 13-20), and in high school locations (22-32); (n.a.: not acquired)

No.	^{7}Li [mg/kg]	^{9}Be [mg/kg]	^{51}V [mg/kg]	^{52}Cr [mg/kg]	^{59}Co [mg/kg]	^{60}Ni [mg/kg]	^{75}As [mg/kg]	^{88}Sr [mg/kg]	^{107}Ag [mg/kg]	^{111}Cd [mg/kg]	^{202}Hg [mg/kg]	^{205}Tl [mg/kg]	^{208}Pb [mg/kg]
1	18.29	n.a.	n.a.	34.86	0.99	5.87	n.a.	334.81	0.28	0.21	n.a.	n.a.	9.66
2	26.99	n.a.	n.a.	28.42	0.58	4.43	n.a.	15.73	0.28	0.13	n.a.	n.a.	3.41
3	62.28	n.a.	n.a.	23.79	0.55	14.97	n.a.	14.52	0.35	3.87	n.a.	n.a.	12.39
4	7.33	n.a.	n.a.	97.22	1.11	9.23	n.a.	41.59	0.35	0.19	n.a.	n.a.	13.07
5	12.89	n.a.	n.a.	89.73	1.77	19.60	n.a.	22.86	0.58	0.83	n.a.	n.a.	40.13
6	7.80	n.a.	n.a.	195.61	1.49	6.99	n.a.	17.22	0.39	0.23	n.a.	n.a.	7.07
7	6.48	n.a.	n.a.	63.48	0.81	4.60	n.a.	11.43	0.13	0.17	n.a.	n.a.	3.45
8	7.92	n.a.	n.a.	30.67	1.14	6.87	n.a.	37.33	0.20	0.21	n.a.	n.a.	5.13
9	0.76	n.a.	n.a.	5.47	0.64	5.52	n.a.	13.24	0.18	0.20	n.a.	n.a.	5.79
10	1.07	n.a.	n.a.	5.87	0.47	6.11	n.a.	16.46	0.43	0.15	n.a.	n.a.	6.68
11	2.18	n.a.	n.a.	15.12	1.30	4.03	n.a.	17.73	0.36	0.10	n.a.	n.a.	8.09
12	14.61	n.a.	n.a.	69.44	3.36	45.90	n.a.	91.79	1.59	0.84	n.a.	n.a.	35.28
13	3.20	n.a.	n.a.	39.31	1.86	6.24	n.a.	37.48	0.09	6.11	n.a.	n.a.	2,779.47
14	6.79	n.a.	n.a.	36.77	14.13	14.07	n.a.	180.51	0.23	4.38	n.a.	n.a.	158.43
15	6.21	n.a.	n.a.	100.81	5.33	15.20	n.a.	98.84	0.36	2.47	n.a.	n.a.	114.51
16	3.65	n.a.	n.a.	29.61	5.34	17.89	n.a.	65.91	0.46	5.94	n.a.	n.a.	620.00
17	6.56	n.a.	n.a.	227.95	6.82	129.53	n.a.	127.61	0.08	1.80	n.a.	n.a.	164.50
18	10.90	n.a.	n.a.	29.78	3.83	27.52	n.a.	183.32	0.23	5.86	n.a.	n.a.	68.08
19	4.21	n.a.	n.a.	43.85	2.88	189.28	n.a.	72.15	0.34	15.39	n.a.	n.a.	354.79
20	0.46	n.a.	n.a.	13.93	0.30	1.81	n.a.	55.88	<LOD	0.04	n.a.	n.a.	3.77
21	0.57	n.a.	n.a.	10.59	0.25	2.91	n.a.	11.95	0.55	0.11	n.a.	n.a.	3.20
22	2.88	0.09	5.37	49.84	1.67	48.20	1.15	33.74	2.46	0.75	0.15	0.04	13.93
23	3.12	0.04	2.48	54.06	4.05	13.25	9.89	40.08	2.59	1.12	0.20	0.03	53.30
24	1.96	0.07	2.45	56.36	1.28	14.38	0.97	46.72	5.83	1.79	0.15	0.04	20.70
25	13.68	0.58	10.82	37.93	4.29	10.76	5.18	54.26	1.19	0.70	0.11	0.14	9.92
26	3.38	0.11	3.89	86.86	1.33	8.31	0.94	46.84	0.55	0.26	0.05	0.03	7.74
27	7.37	0.10	4.49	41.27	2.67	78.93	0.97	57.52	4.49	0.54	0.28	0.04	21.93
28	0.60	0.01	0.63	117.17	1.05	8.59	0.20	10.17	2.16	0.75	0.03	0.01	3.43
29	3.18	0.10	3.57	64.98	1.44	16.04	0.84	44.85	0.49	1.09	0.04	0.03	23.94
30	3.56	0.13	6.97	81.12	3.59	16.13	1.92	73.19	0.88	1.17	0.10	0.06	196.30
31	2.62	0.09	2.97	122.83	1.75	8.17	0.68	33.72	1.36	0.62	0.08	0.03	9.87
32	11.97	0.14	7.04	67.86	5.34	263.70	2.30	158.35	34.67	3.83	9.92	0.07	273.90

Comparison of repeatability and recovery: A close correlation between the validation results of the manual and the automated process was reached. The individual validation results were presented and compared in previous publications [1, 44]. These

results show the successful transfer of the manual measurement method to the automation system. In the two measurement processes—manual and automated—relatively high CV values were observed in the determination of the repeatability (35.45%) and the recovery rate (min. 56.11%, 171.37%) for the element cadmium. Reasons can be found in the inhomogeneity of the CRM. The CRM used in this study certifies Cd uncertainties of ±3.7 mg/kg (±50.7%). In the literature, similar results were reported for other dust reference materials such as ERM-CZ120 [14].

5.5. Measurement Results of Collected Dust Samples

The automated measurement system was tested with a wide variety of real samples. Dust samples were collected in rooms of the institute including robotic laboratories, chemical laboratories, analytical laboratories, and offices (samples 1-7, 12), in private households (samples 8-11, 21), in old disused garden houses (samples 13-20), and at several places in a high school including class and storage rooms, cafeteria, auditorium, and the sports hall. The indoor dust was taken with hand gloves and plastic forceps and was stored in plastic containers to avoid contamination during the sampling.

Selected results were previously presented in [1, 44]. Overall, relatively low heavy metal concentrations were acquired, which were within the expected range. Higher concentrations of some heavy metals were found in dust samples collected in old garden houses. Further, elevated heavy metal concentrations were monitored in access-controlled storage cabinets for chemicals. Table 5 summarizes the results—previously published and extended with results of additional heavy metals. In the first 21 samples no recording of the concentrations of Be, V, As, Hg, and Tl is available since these elements were later included into the measurement method to extend the methodological scope. The concentrations of all elements shown in Table 5 were determined for the subsequent sample set (samples 22-32). The measured heavy metal concentrations in the collected indoor dust samples show a wide concentration range. Elevated Pb concentrations were determined in the samples from old garden houses.

Elevated As and Hg concentrations were monitored in dust collected in storage cabinets of chemicals. The concentrations of Fe were higher than the highest calibrations standard in all samples measured. If the aim of investigation is determination of Fe, then an additional dilution step must be included in the sample preparation to adjust the concentration in the measurement solutions to the typical working range of the ICP-MS detector. In this study, the determination of Fe was not the focus and no additional dilution was performed. Figure 5 shows the mass spectra of a calibration standard with an element concentration of 100 µg/L and of a dust sample collected in an old garden house (sample 14). An exemplarily overview of the heavy metal concentrations (Li, Cr, Co, and Sr) in different samples gives Figure 6.

A more detailed exploration of the heavy metal distribution requires the analysis of a higher sample number per location. This will be subject to future investigations. The results of this study show the suitability and performance of the automated sample pretreatment and the analytical measurement setup for its application in the determination of heavy metals in indoor dust—and possibly in outdoor dust.

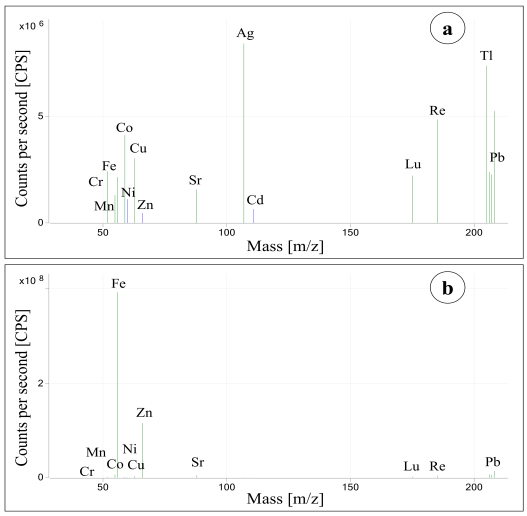


Figure 5: ICP-MS mass spectra—(a) calibration standard with element concentrations of 100 µg/L, (b) collected dust sample 14.

6. Conclusion

In this extended paper, the entire study of the development and application of an automated robot-assisted system for determination of multiple heavy metals in indoor dust was presented. The study started with the development and the validation of the manual sample preparation (microwave-assisted acid digestion) of dust samples and the measurement method (ICP-MS measurements). The manual procedure was validated with the following parameters determined using certified reference material: repeatability, recovery rates, within-laboratory precision, measurement precision as well as the limit of detection (LOD) and the limit of quantification (LOQ). In the next step, the automation system was extended, new racks for the used labware were designed and manufactured and new processes were created in the three-level software system. The manual procedure was not identical transferred to the automation system. The subprocess dilution was miniaturized resulting in a higher sample throughput by using a higher number of smaller vessels. Furthermore, the integration of the single-vial liquid handler allows the dispensing of a large volume ultrapure water in one step, reducing pipetting uncertainties. The automated procedure was identically validated and compared with the manual processing.

The developed measurement method can be understood and used as a screening procedure to prepare and measure a high number of dust samples in a relative short time frame. The processing time of the automated process is approx. 15 min longer than the manual procedure. But while running the automated process, the laboratory staff is free for other tasks operated in parallel. The method validation for both the manual and the automated procedure was performed using certified reference material. The results of the two methods were compared with the certified values for Cr, Cd, As, Hg, and Pb. Close results were reached for As, Cd, and Pb. The recovery rates of Cr and Hg were lower than certified, but they are acceptable for a screening method. Conspicuous samples found in the screening can then be analyzed by the standard method using a stronger acid mixture in the microwave digestion procedure.

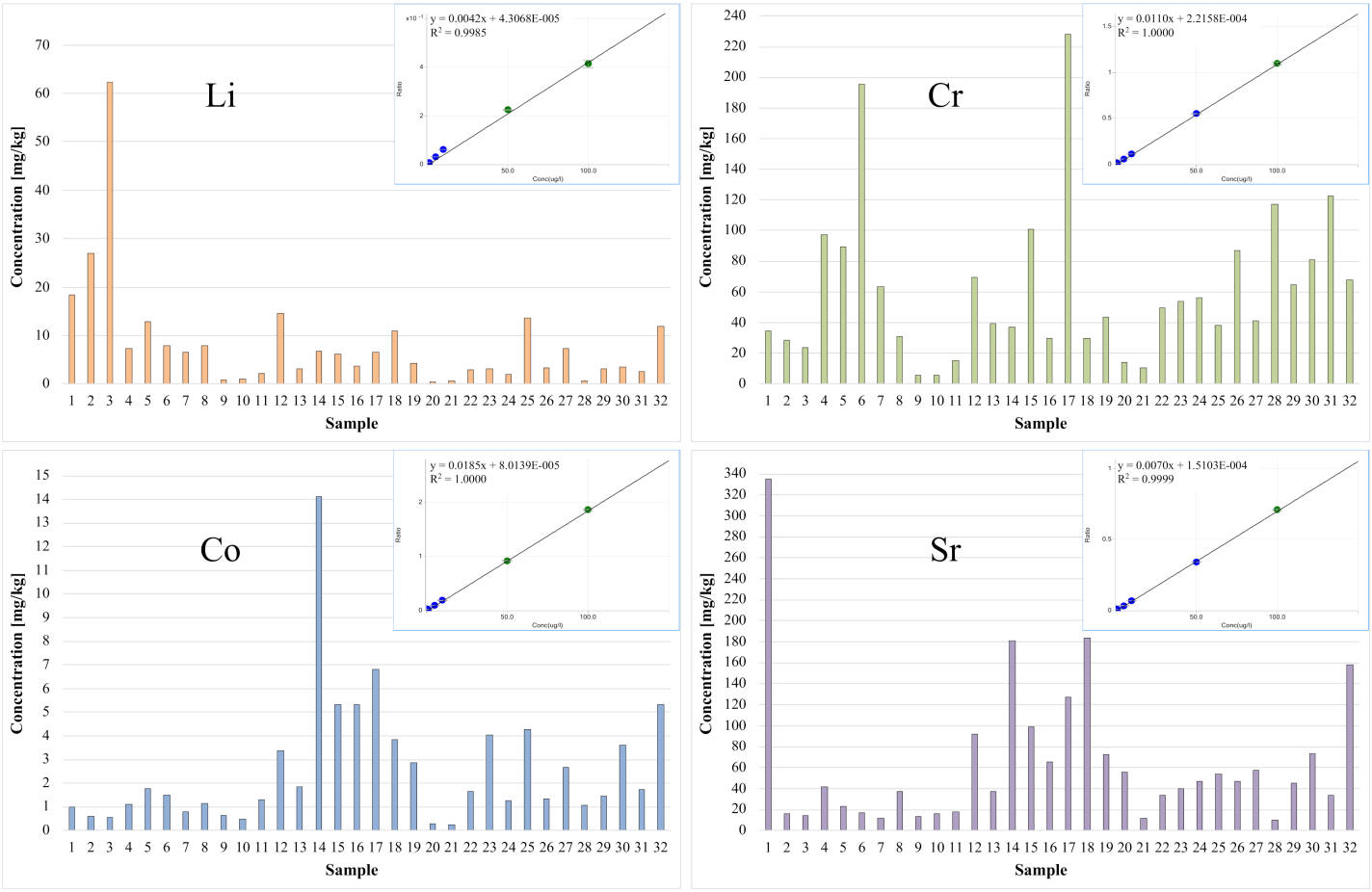


Figure 6: Calibration data and concentrations of the elements Li, Cr, Co, and Sr measured in dust samples which were collected in laboratories and offices (samples 1-7, 12), in private households (samples 8-11, 21), in old disused garden houses (samples 13-20), and in high school locations (samples22-32)

Besides the five elements certified in the CRM, additional elements were measured and validated (except for the determination of the recovery rate). Depending on the element concentration, the recorded CV values show that the automated measurement method is also suitable to screen more than the certified elements.

The automation of the entire process—sample preparation and measurement—has multiple advantages. The human operator is protected from potential hazards caused by harmful reagents such as high concentrated nitric acid and hydrogen peroxide, toxic samples and sample solutions as well as harmful nitrous gases. The laboratory staff is free from highly repetitive tasks like pipetting for multiple times. Other laboratory tasks can be executed in parallel. Overall, this increases the sample throughput as well as the laboratories performance and safety.

The automated measurement system presented is ready for its application in large studies with a high number of samples. In this study, the sample preparation and measurement of collected inhouse dust samples give a first overview of the element concentrations and show the performance of the automation. In further studies, a suitable design of experiment for the desired location is needed together with well-planned sampling methods to acquire large data sets of the area to be investigated. The automation system is here a contribution to the current exploration of the distribution of heavy metals inhouse and for a safer and healthier daily life. Furthermore, the system presented can also be applied in environmental studies screening outdoor dust.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors thank Dr. C. Kupke for organizing the project "Jugend Forscht" (German youth science competition) with the Innerstädtisches Gymnasium (Senior High School) in Rostock, Germany. Special thanks are expressed to H. Engelhardt, Dipl.-Ing. L. Woinar, and Dr. S. Junginger for technical support and suggestions, as well as to S. Horn and B.Sc. A. Reichelt for their excellent laboratory assistance.

References

[1] H. Fleischer, J. Widmer, S. Statkevych, R. Stoll, T. Roddelkopf, K. Thurow, "Robot-assisted Measurement of Heavy Metals in Indoor Dust Using ICP-MS," in 2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), 1-6, 2021, doi: 10.1109/I2MTC50364.2021.9459859.

[2] S. Y. Tan, S. M. Praveena, E. Z. Abidin, M. S. Cheema, "A review of heavy metals in indoor dust and its human health-risk implications," Reviews on environmental health, **31**(4), 447-456, 2016, doi: 10.1515/reveh-2016-0026.

[3] N. Bukowiecki, M. Hill, R. Gehrig, C. N. Zwicky, P. Lienemann, F. Hegedüs, G. Falkenberg, E. Weingartner, U. Baltensperger, "Trace Metals in Ambient Air: Hourly Size-Segregated Mass Concentrations Determined by Synchrotron-XRF," Environmental Science & Technology, **39**(15), 5754-5762, 2005, doi: 10.1021/es048089m.

[4] T. Shi, Y. Wang, "Heavy metals in indoor dust: Spatial distribution, influencing factors, and potential health risks," Science of the total Environment, **755**(2021), 2021, doi: 10.1016/j.scitotenv.2020.142367.

[5] S. T. Y. Tong, "Indoor and outdoor household dust contamination in Cincinnati, Ohio, USA," Environmental Geochemistry and Health, **20**(3), 123-133, 1998, doi: 10.1023/A:1006561832381.

[6] R. Millan, T. Schmid, M. J. Sierra, S. Carrasco-Gil, M. Villadóniga, C. Rico, D. M. S. Ledesma, F. J. D. Puente, "Spatial variation of biological and pedological properties in an area affected by a metallurgical mercury plant: Almadenejos (Spain)," Applied Geochemistry, **26**(2), 174-181, 2011, doi: 10.1016/j.apgeochem.2010.11.016.

[7] D. J. A. Davies, J. M. Watt, I. Thornton, "Lead levels in Birmingham dusts and soils," The Science of the Total Environment, **67**(2-3), 177-185, 1987, doi: 10.1016/0048-9697(87)90210-5.

[8] J. Patocka, K. Cerný, "Inorganic lead toxicology," Acta medica (Hradec Králové) / Universitas Carolina, Facultas Medica Hradec Králové, **46**(2), 65-72, 2003, doi: 10.14712/18059694.2019.8.

[9] C. D. Klaassen, J. Liu, B. A. Diwan, "Metallothionein protection of cadmium toxicity," Toxicology and Applied Pharmacology, **238**(3), 215-220, 2009, doi: 10.1016/j.taap.2009.03.026.

[10] M. P. Waalkes, "Cadmium carcinogenesis in review," Journal of Inorganic Biochemistry, **79**(1-4), 241-244, 2000, doi: 10.1016/S0162-0134(00)00009-X.

[11] D. J. A. Davies, I. Thornton, J. M. Watt, E. B. Culbard, P. G. Harvey, H. T. Delves, J. C. Sherlock, G. A. Smart, J. F. A. Thomas, M. J. Quinn, "Lead intake and blood lead in two-year-old U.K. urban children," The Science of the Total Environment, **90**(C), 13-29, 1990, doi: 10.1016/0048-9697(90)90182-T.

[12] H. Fleischer, K. Thurow, "Compound-oriented Measurement Processes: Elements, Molecules, Structures," in 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), 1-6, 2020, doi: 10.1109/I2MTC43012.2020.9128468.

[13] R. Lumpp, M. Klein, E. Bieber, F. Bunzel, U. Eckermann, C. Frels, W. Günther, C. Hagemann, C. Koch, A. Olschewski, C. Temme, "Comparison of different digestion methods for elemental analysis of airborne dust and dust deposition," Gefahrstoffe - Reinhaltung der Luft, 72(1/2), 64-71, 2012 (article in German with an abstract in English).

[14] R. Lumpp, M. Klein, K. Berger, E. Bieber, F. Bunzel, M. Ernst, C. Frels, W. Günther, C. Hagemann, C. Koch, A. Olschewski, "A study on the certified reference material ERM®-CZ120 – Results of round robin tests," Gefahrstoffe - Reinhaltung der Luft, **75**(1/2), 35-40, 2015 (article in German with an abstract in English).

[15] Deutsches Institut für Normung e.V., "Ambient air quality - Standard method for the measurement of Pb, Cd, As and Ni in the PM10 fraction of suspended particulate matter," (DIN EN 14902), 2005.

[16] Deutsches Institut für Normung e.V., "Stationary Source Emissions - Determination of the total emission of As, Cd, Cr, Co, Cu, Mn, Ni, Pb, Sb, Tl and V," (DIN EN 14385), 2004.

[17] Deutsches Institut für Normung e.V., "Ambient air quality - Standard for determination of arsenic, cadmium, lead and nickel in atmospheric deposition," (DIN EN 15841), 2010.

[18] Verein Deutscher Ingenieure, "Chemical Analyis of Particulate Matter - Determination of Ba, Be, Cd, Co, Cr, Cu, Ni, Pb, Sr, V, Zn in Particulate Emissions by Atomic Spectrometric Methods," (VDI 2268 Part 1), 1987.

[19] Verein Deutscher Ingenieure, "Chemical Analysis of Particulate Matter - Determination of Arsenic, Antimony and Selenium in Dust Emissions by Atomic Absorption Spectrometry after Separation of their Volatile Hydrides," (VDI 2268 Part 2), 1990.

[20] Verein Deutscher Ingenieure, "Chemical Analysis of Particulate Matter - Determination of Arsenic, Antimony and Selenium in Dust Emissions by Graphite-Furnace Atomic Absorption Spectrometry," (VDI 2268 Part 4), 1990.

[21] Verein Deutscher Ingenieure, "Determination of suspended matter in ambient air - Measurement of the element concentration after sampling on filters - Determination of Al, As, Ba, Ca, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Na, Ni, Pb, Sb, Se, Sn, Tl, V, and Zn by GF-AAS, ICP-OES, or ICP-MS," (VDI 2267 Part 1), 2019.

[22] Verein Deutscher Ingenieure, "Determination of suspended matter in ambient air - Measurement of Al, As, Ba, Ca, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Na, Ni, Pb, Sb, Se, Sn, Tl, V, and Zn as part of the atmospheric deposition after sampling with bulk and wet-only collectors using GF-AAS, ICP-OES, and ICP-MS," (VDI 2267 Part 2), 2019.

[23] Verein Deutscher Ingenieure, "Determination of suspended matter in ambient air - Digestion variants for dust samples for subsequent determination of the mass concentration of Al, As, Ba, Ca, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Na, Ni, Pb, Sb, Se, Sn, Tl, V and Zn," (VDI 2267 Part 3), 2015.

[24] Verein Deutscher Ingenieure, "Determination of suspended particles in ambient air - Measurement of the mass concentrationof mercury - Sampling by sorption as amalgam and determination by atomic absorption spectrometry (AAS)," (VDI 2267 Part 8), 2000.

[25] Verein Deutscher Ingenieure, "Determination of suspended particulate matter in ambient air - Measurement of the mass concentration of mercury - Sampling by sorption as amalgam and determination by atomic fluorescence spectrometry (AFS)," (VDI 2267 Part 9), 2002.

[26] Verein Deutscher Ingenieure, "Determination of suspended particulate matter in ambient air - Measurement of the mass concentration of As, Ba, Cd, Co, Cr, Cu, Ni, Pb, Sb, V und Zn as part of dust precipitation by atomic absorption spectrometry (AAS)," (VDI 2267 Part 16), 2007.

[27] W. A. Willford, R. J. Hesselberg, H. L. Bergman, "Versatile combustion amalgamation technique for the photometric determination of mercury in fish and environmental samples," Journal of the Association of Official Analytical Chemists, **56**(4), 1008-1014, 1973, doi: 10.1093/jaoac/56.4.1008.

[28] G. Kopp, "Quecksilberanalytik - Für jede Applikation die optimale Methode," LABO - Instrumentelle Analytik, **2012**(6), 20–23, 2012.

[29] Rohasys Robotic and Handling Systems, Rouwette Consultancy. "Probenvorbereitung für Schwermetallanalysen," 08.03.2022; http://issuu.com/rouwetteconsultancy/docs/probenvorbereitung_f__r_schwermetal.

[30] J. W. Milburn, "Automated addition of internal standards for axial-view plasma ICP spectrometry using the Optima 3000 XL," Atomic Spectroscopy, **17**(1), 9–14, 1996.

[31] H. Fleischer, K. Thurow, "Determination of Total Mercury Content and Trace Metal Analysis in Wood Materials Part 1: ICP-OES Using Mercury Cold Vapor," American Laboratory, **45**(8), 6–9, 2013.

[32] H. Fleischer, K. Thurow, "Determination of Total Mercury Content in Wood Materials - Part 2: ICP-MS - A Multi-element Method," American Laboratory, **45**(9), 6–11, 2013.

[33] H. Fleischer, E. Vorberg, K. Thurow, "Determination of Total Mercury Content in Wood Materials - Part 3: Miniaturization Using ICP-MS," American Laboratory, **46**(6), 16–20, 2014.

[34] E. Vorberg, H. Fleischer, S. Junginger, N. Stoll, K. Thurow, "Automated Sample Preparation for Mercury Analysis in Wood Materials," IET Science, Measurement and Technology, **10**(5), 398-404, 2016, doi: 10.1049/iet-smt.2015.0036.

[35] H. Fleischer, E. Vorberg, M. Warkentin, D. Behrend, K. Thurow, "Determination of Calcium and Phosphor in Bones Using Microwave Digestion and ICP-MS: Comparison of Manual and Automated Methods using ICP-MS," 5th IMEKO TC19 Symposium on Environmental Instrumentation and Measurement), 94–99, 2014.

[36] H. Fleischer, K. Ramani, S. Bauer, M. Warkentin, D. Behrend, K. Thurow, "Determination of Calcium and Trace Metals in Clogged Biliary Endoprostheses Using ICP/MS," American Laboratory, **49**(6), 6-9, 2017.

[37] H. Fleischer, K. Ramani, K. Blitti, M. Warkentin, D. Behrend, K. Thurow, "Flexible Automation System for Determination of Elemental Composition of Incrustations in Clogged Biliary Endoprostheses Using ICP-MS," SLAS Technology, **23**(1), 83-96, 2018, doi: 10.1177/2472630317727451.

[38] H. Fleischer, K. Ramani, K. Thurow, "Sample Preparation and Measurement of Cholesterol in Pig Bile Using SPE and GC/MS," American Laboratory, **52**(2), 10-13, 2019.

[39] H. Fleischer, T. Roddelkopf, R. Stoll, K. Thurow, "Automated Analytical Measurement System for Determination of Cholesterol in Pig Bile," in 2019 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), 1510-1515, 2019, doi: 10.1109/I2MTC.2019.8826917.

[40] H. Fleischer, C. Lutter, A. Buttner, W. Mittelmeier, K. Thurow, "Semi-Automated Determination of Heavy Metals in Autopsy Tissue Using Robot-Assisted Sample Preparation and ICP-MS," Molecules, **26**(13, 2021, doi: 10.3390/molecules26133820.

[41] H. Fleischer, T. Roddelkopf, C. Lutter, A. Büttner, W. Mittelmeier, K. Thurow, "Automation in arthrosis research," SLAS Technology, **27**(1), 97-99, 2022, doi: 10.1016/j.slast.2021.10.009.

[42] T. Schwank, K. Pitzke, K. Gusbeth, K. Ashley, D. Breuer, "Comparison of Microwave-Assisted Digestion and Consensus Open-Vessel Digestion Procedures for Evaluation of Metalliferous Airborne Particulate Matter," Annals of Work Exposures and Health, **63**(8), 950-964, 2019, doi: 10.1093/annweh/wxz068.

[43] P. E. Rasmussen, K. S. Subramanian, B. J. Jessiman, "A multi-element profile of house dust in relation to exterior dust and soils in the city of Ottawa, Canada," Science of the total Environment, **267**(1-3), 125-140, 2001, doi:

10.1016/S0048-9697(00)00775-0.
[44] H. Fleischer, S. Statkevych, J. Widmer, K. Thurow, "Allergene und Schwermetalle in Hausstaub - Bestimmung der elementaren Zusammensetzung mittels ICP-MS," GIT Labor-Fachzeitschrift, **65**(5-6), 36-38, 2021.
[45] H. Fleischer, K. Thurow, Automation Solutions for Analytical Measurements - Concepts and Applications, Wiley-VCH, 2017.
[46] H. Fleischer, K. Thurow, "On the Way to Efficient Analytical Measurements: The Future of Robot-Based Measurements," SLAS Technology, **25**(2), 208-211, 2020, doi: 10.1177/2472630319886270.
[47] S. Neubert, X. Gu, B. Göde, T. Roddelkopf, H. Fleischer, N. Stoll, K. Thurow, "Workflow Management System for the Integration of Mobile Robots in Future Labs of Life Sciences," Chemie-Ingenieur-Technik, **91**(3), 294-304, 2019, doi: 10.1002/cite.201800007.
[48] H. Fleischer, M. Adam, K. Thurow, "A Cross-platform Modular Software Solution for Automated Data Evaluation Applied in elemental and Structural Mass Spectrometry," in 2015 IEEE International Conference on Automation Science and Engineering (CASE), 758-763, 2015, doi: 10.1109/CoASE.2015.7294172.
[49] M. Adam, H. Fleischer, K. Thurow, "Generic and Automated Data Evaluation in Analytical Measurement," SLAS Technology, **22**(2), 186-194, 2017, doi: 10.1177/2211068216672613.
[50] H. Liu, N. Stoll, S. Junginger, K. Thurow, "Mobile robot for life science automation," International Journal of Advanced Robotic Systems, **10**, 1-14, 2013, doi: 10.5772/56670.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 4, 140-146 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Scalability of Multi-Stage Nested Mach-Zehnder Interferometer Optical Switch with Phase Generating Couplers

Masayuki Kawasako, Toshio Watanabe*, Tsutomu Nagayama, Seiji Fukushima

Kagoshima University, 1-21-40 Korimoto, Kagoshima, 890-0065, Japan

ARTICLE INFO

Article history:
Received: 31 May, 2022
Accepted: 07 August, 2022
Online: 29 August, 2022

Keywords:
Optical Switch
Optical Waveguide
Mach-Zehnder Interferometer

ABSTRACT

A nested Mach-Zehnder interferometer (MZI) configuration whose phase shifters are placed in parallel is suitable for silicon-silica hybrid structure to realize a high-speed optical switch. Even when the signal wavelength deviates from an optimal wavelength, the crosstalk of the nested MZI optical switch can be suppressed by employing phase generating couplers (PGCs) in place of directional couplers. We calculate the characteristics of a 4-stage nested MZI switch with PGCs, and show that crosstalk is lower than −40 dB over a wavelength range of as wide as 200 nm from 1450 to 1650 nm in six output ports. We also examine the scalability of the multi-stage nested MZI switch, and deduce the required number of switch stages for given output port counts with low crosstalk.

1. Introduction

In optical fiber communication systems, optical switches are employed to route the optical signals passing through a node without any conversion into electrical signals. Among various types of optical switches, those based on silica waveguides have many advantages [1-5]. Wafer-level fabrication process enables large-scale integration and mass productivity. Polarization and temperature insensitive operation eliminates the need of polarization diversity optics and thermo-electric cooler. Physical and chemical stability of silica as well as the absence of moving mechanical parts offer high reliability. The silica-based optical switches with thermo-optic phase shifters have been investigated since as early as 1980s [1], and their excellent performance as well as high feasibility has been well qualified [2]. They have been deployed in reconfigurable optical add/drop multiplexing (ROADM) systems in 2000s [3], where their millisecond response time is sufficient for this application. Their versatility has been successfully demonstrated thanks to the fact that they can integrate with various kinds of optical components [4]. They have been also deployed in multi-degree ROADM systems as a multicast switch in 2010s [5].

Recently, an application of optical switches to data center network is intensively considered [6-8], where switching speed is a critical issue and microsecond response time is required. A new optical switch architecture based on wavelength routing was proposed, which is suitable for data center network [6]. A fast-optical switch employing digital micromirror device (DMD)-based wavelength selective switch was demonstrated [7], and a review on optical switches for data centers was reported [8]. To realize a waveguide-based optical switch with high speed, silicon-silica hybrid structure [9] is promising since silicon provides excellent thermal properties, although it was reported the investigation to enhance the switching speed of the silica-based optical switch by optimizing the waveguide structure and using an over-driving technique [10].

In the waveguide-based optical switches, a Mach-Zehnder interferometer (MZI) is commonly employed as a basic element [1]. An MZI consists of two directional couplers (DCs) connected by arm waveguides with phase shifters. It acts as 1 × 2 or 2 × 2 optical switches, and we can construct 1 × N or M × N optical switch by cascading them [11-13]. However, conventional cascaded MZI optical switches are not suitable for the hybrid structure because their phase shifters are placed in series. In contrast, a nested MZI configuration [14, 15], whose phase shifters are placed in parallel, is suitable for the hybrid structure.

Crosstalk of the nested MZI optical switch as well as the conventional MZI optical switch becomes larger when the input signal wavelength deviates from an optimal wavelength because both the coupling ratio of the DC and the phase given by the arm waveguides vary as wavelength. To mitigate this wavelength dependence, a previous study [16] proposed a nested MZI optical switch employing phase generating couplers (PGCs) [17] in place of DCs. PGCs create the phase to cancel the wavelength dependence of the arm waveguides. Calculation results showed

*Corresponding Author: Toshio Watanabe, wata104@eee.kagoshima-u.ac.jp

that a single-stage MZI switch with PGCs has low crosstalk less than −30 dB over a wavelength band of 200 nm ranging from 1450 to 1650 nm, whereas a conventional one with DCs has crosstalk lower than −30 dB within 60 nm wavelength range from 1520 to 1580 nm [16]. It was also found that crosstalks in a 2-stage nested MZI switch are further suppressed in two output ports chosen among four output ports. A 3-stage nested MZI optical switch has three output ports with low crosstalk among eight output ports [16]. In addition, we have reported that a 4-stage nested MZI switch has low crosstalk less than −40 dB over a wavelength range of as wide as 200 nm from 1450 to 1650 nm by choosing six ports among 16 output ports [18].

In this paper, we further examine the crosstalk characteristics of the multi-stage nested MZI optical switch and reveal how crosstalk is suppressed in selected output ports. Section 2 describes the difference between a conventional cascaded MZI switch and a nested MZI switch. Section 3 shows the simulation method for calculating crosstalk characteristics of this optical switch, including the PGC circuit configuration and their parameters used in this study. We discuss the calculation results and exhibit the scalability of the multi-stage nested MZI switch in Section 4. Finally, we conclude this paper in Section 5.

2. Conventional MZI Switch and Nested MZI Switch

Figure 1 shows a conventional cascaded MZI optical switch [11-13]. It has two output ports (E3, E5) with low crosstalk among five output ports, because crosstalks to these two ports are blocked by the MZI in through state twice, respectively. This is a common way to suppress crosstalks to unrouted output ports in the conventional cascaded MZI switch. The through and cross ports in the MZI are never symmetric in terms of their optical transmittance when the coupling ratio of DC deviates from 50 % and the phase

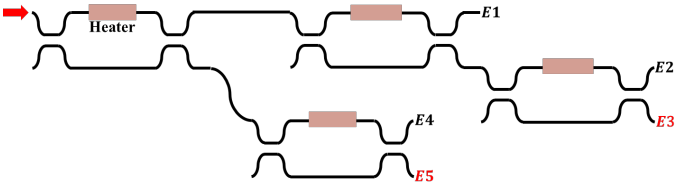


Figure 1: Conventional cascaded MZI optical switch

given by the arm waveguide does from π due to their wavelength dependence. The transfer matrix of DC is expressed by

$$T_{\mathrm{DC}}=\begin{bmatrix} \kappa' & -j\kappa \\ -j\kappa & \kappa' \end{bmatrix}, \tag{1}$$

where κ' and κ are the amplitudes of optical field coupled to the through and cross ports, respectively, and $\kappa'^2+\kappa^2=1$ for a lossless DC. Then the optical output powers in the through and cross ports are given as

$$|E_1|^2 = 1-4K(1-K)\cos^2\left(\frac{\psi}{2}\right) \tag{2}$$

and

$$|E_2|^2 = 4K(1-K)\cos^2\left(\frac{\psi}{2}\right), \tag{3}$$

respectively, where $K=\kappa^2=1-\kappa'^2$ is the power coupling ratio of DC and ψ is the phase given by the arm waveguide [1]. Equations (2) and (3) indicate that crosstalk to through port becomes zero only if $K=0.5$ in cross state (ψ=0), while crosstalk to cross port vanishes regardless of K in through state (ψ=π). In general, the wavelength dependence of K is greater than that of ψ, cross port (in through state) offers lower crosstalk than through port (in cross state). In a 1 × N switch composed of cascaded MZIs shown in Figure 1, the number of output ports with low crosstalk is less than half of the total optical output counts.

The conventional cascaded MZI optical switch, however, is not suitable for the hybrid structure because its phase shifters are placed in series. This makes it difficult to fabricate the phase shifter and other waveguide regions with different material. As the output port counts and the number of MZIs with the phase shifter increase, the transmission loss becomes larger since the coupling loss between the waveguides with different materials is accumulated every time passing through the MZI. In contrast, a nested MZI switch [14-16, 18] is suitable for the hybrid structure because its phase shifters are placed in parallel. It is easy to fabricate the phase shifter and the other waveguide regions separately with different materials, and transmission loss does not increase with the number of output ports.

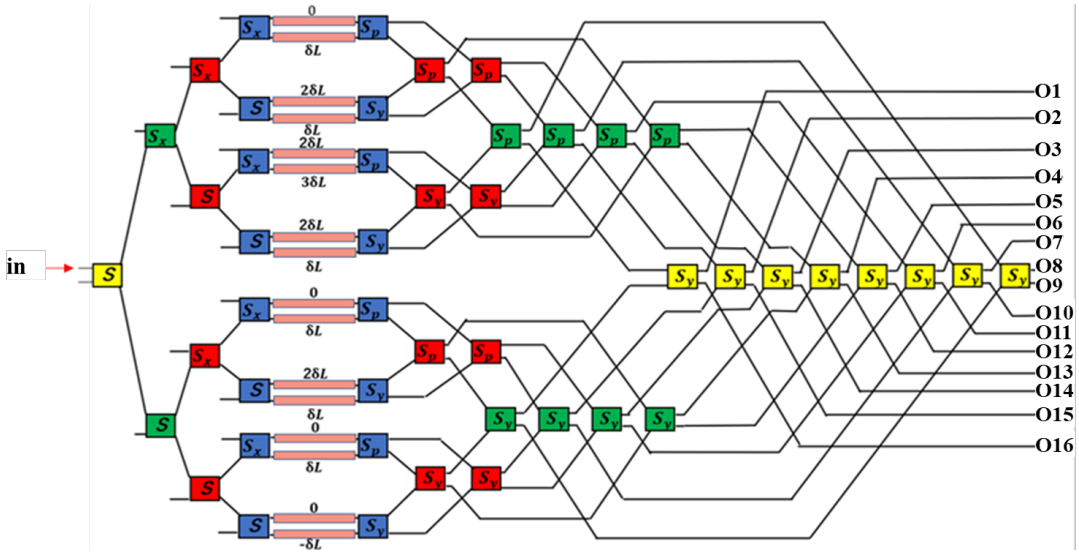


Figure 2: Circuit configuration of 4-stage nested MZI switch with PGCs

Figure 2 shows a circuit configuration of the optical switch we examined in this paper. This circuit corresponds to an MZI switch arranged in a nested architecture with four stages [18]. The PGCs indicated in different colors (blue, red, green and yellow) construct the first, second, third and fourth stages, respectively. An inherent symmetry of this circuit configuration makes it easy to increase the number of stages and output port counts [16]. However, the crosstalk characteristics of the nested MZI optical switch with a large number of stages has not been clarified yet.

3. Simulation Method

We employ PGCs instead of the DCs in all nested MZIs. Figure 3 shows the circuit configurations of the PGCs: a basic element as well as those having the line symmetries with respect to X- or Y-axes and point symmetry. The PGCs in each MZI stages are arranged in mirror symmetry as shown in Figure 2. In order to give an optical phase shift of π, appropriate optical path length differences are set between the arm waveguides connecting the PGCs, as also shown in Figure 2. Table 1 shows the MZI state of each stage for given optical output ports. In Table 1, C denotes that the corresponding MZI stage is in cross state while T does that it is in through state. Table 2 shows the phase settings to route each optical output port, where ϕ_i $(i = 1, 2, \dots, 16)$ is that of i-th phase shifter, and $\phi_i = 0$ and 1 indicate the phase shift of 0 and π, respectively.

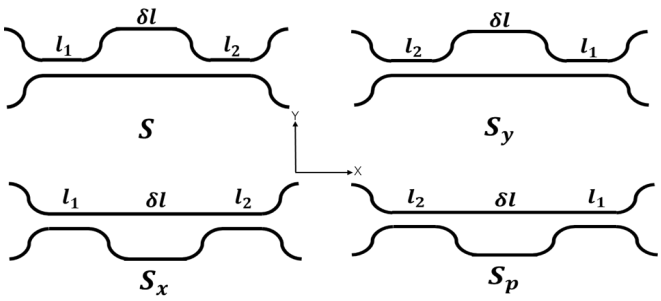


Figure 3: Basic PGC element, line-symmetric (with X- or Y-axes) and point-symmetric PGCs

We calculated the optical transmittance of the nested MZI to each output port by the same method as the previous study [16]. The circuit parameters used in this study was also the same as those given previously. The power coupling ratio of the PGCs was designed to be 50 % (3 dB) at 1550 nm, the center wavelength of

Table 1: MZI states

Stage / Output	1st stage	2nd stage	3rd stage	4th stage
O1	T	T	T	T
O2	C	T	T	T
O3	C	C	T	T
O4	T	C	T	T
O5	T	C	C	T
O6	C	C	C	T
O7	C	T	C	T
O8	T	T	C	T
O9	T	T	C	C
O10	C	T	C	C
O11	C	C	C	C
O12	T	C	C	C
O13	T	C	T	C
O14	C	C	T	C
O15	C	T	T	C
O16	T	T	T	C

MZI state: C = cross, T = through

Table 2: Phase settings

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	ϕ_8	ϕ_9	ϕ_{10}	ϕ_{11}	ϕ_{12}	ϕ_{13}	ϕ_{14}	ϕ_{15}	ϕ_{16}
O1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
O2	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0
O3	0	1	0	1	1	0	1	0	0	1	0	1	1	0	1	0
O4	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0
O5	0	0	1	1	0	0	1	1	1	1	0	0	1	1	0	0
O6	0	1	0	1	0	1	0	1	1	0	1	0	1	0	1	0
O7	0	1	1	0	1	0	0	1	1	0	0	1	0	1	1	0
O8	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0
O9	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
O10	0	1	1	0	1	0	0	1	0	1	1	0	1	0	0	1
O11	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
O12	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
O13	0	0	1	1	1	1	0	0	1	1	0	0	0	0	1	1
O14	0	1	0	1	1	0	1	0	1	0	1	0	0	1	0	1
O15	0	1	1	0	0	1	1	0	1	0	0	1	1	0	0	1
O16	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1

operating range. The parameters of the PGC are as follows: coupling lengths l_1 = 20 μm and l_2 = 10 μm, delay line length δl = 0.086 μm, and optical path length difference δL = 0.4 μm.

The optical output field of a single-stage MZI switch with PGCs (configuration S and S_y) is given as

$$\begin{pmatrix} E_1 \\ E_2 \end{pmatrix} = T_{\text{latter}} T_{\text{phase}} T_{\text{former}} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad (4)$$

where E_1 and E_2 indicate output fields in through and cross ports, respectively. In equation (4),

$$T_{\text{latter}} = \begin{bmatrix} H & F \\ -F^* & H^* \end{bmatrix}, \quad (5)$$

$$T_{\text{phase}} = \begin{bmatrix} \exp(-j\psi(\lambda)/2) & 0 \\ 0 & \exp(j\psi(\lambda)/2) \end{bmatrix}, \quad (6)$$

$$T_{\text{former}} = \begin{bmatrix} H & -F^* \\ F & H^* \end{bmatrix} \quad (7)$$

indicate transfer matrix of S_y, the arm waveguides, and S, respectively [16]. Here, H and F indicate transfer functions of through and cross ports in the PGC, respectively. H^* is a complex conjugate of the transfer function H, and $\psi(\lambda)$ is a phase difference between the arm waveguides connecting the PGCs. We simulated the optical output power of a multi-stage MZI switch with PGCs by multiplying these transfer matrices in appropriate order.

4. Results and Discussion

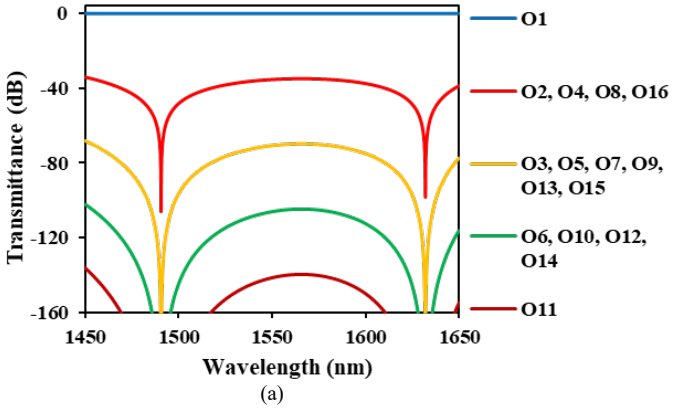
4.1. Results

Figures 4(a)−4(p) show the calculated optical transmittance of the switch we propose when the optical signal is routed to output ports O1−O16, respectively. Here, the transmittance is defined as the ratio of the optical output power to the optical input power. Wavelength dependence was computed in increments of 0.1 nm.

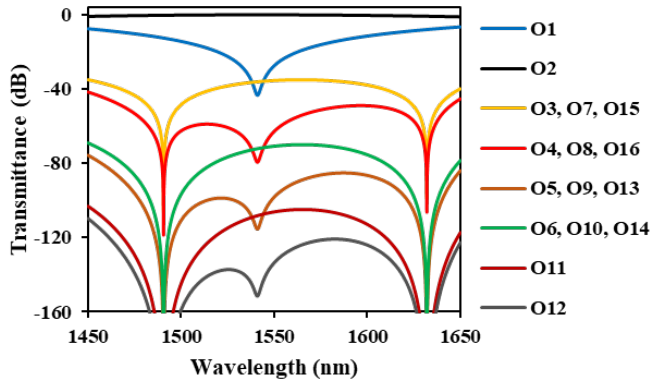
In Figure 4(a) where the optical input signal is routed to O1, crosstalks to unrouted ports of O2, O4, O8 and O16 are the same and less than −30 dB in 200 nm wavelength range from 1450 to 1650 nm. Crosstalks to O3, O5, O7, O9, O13 and O15 are lower than −60 dB, while those to O6, O10, O12 and O14 are below −100 dB. In particular, O11 has quite low crosstalk below −130 dB. In all O2−O16, crosstalks are dropped at two wavelengths of 1490 and 1630 nm, where the PGCs create the phase that exactly cancels the wavelength dependence of the arm waveguides.

However, in Figure 4(b) where the optical signal is directed to O2, crosstalk to O1 increases and becomes higher than −20 dB as the signal wavelength is away from a center wavelength. This is because the power coupling ratio of the PGC deviates from 50 % (3 dB) due to its wavelength dependence. Crosstalks to O3, O7 and O15 are identical and less than −30 dB within the 200 nm wavelength range around 1550 nm, while those to O4, O8 and O16 are lower than −40 dB. Other ports have much lower crosstalks below −60 dB.

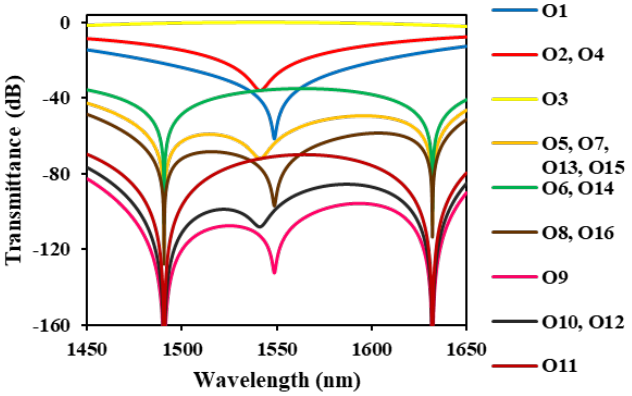
When the optical path is set toward O4, O8 and O16, respectively, O1 has high crosstalk as same as the output is set at O2. Crosstalk to O1 is a little bit low but higher than −20 dB when the optical input signal is directed to O3, O5, O7, O9, O13 and O15, respectively. When the optical signal is routed to O6, O10, O12 and O14, respectively, crosstalk to O1 is identical and also exceeds −20 dB. In the case that the routed port is O11, it remains higher than −20 dB within the wavelength range of 1550 ± 100 nm. These results indicate that we should avoid to use O1 as the output port of the nested MZI optical switch.



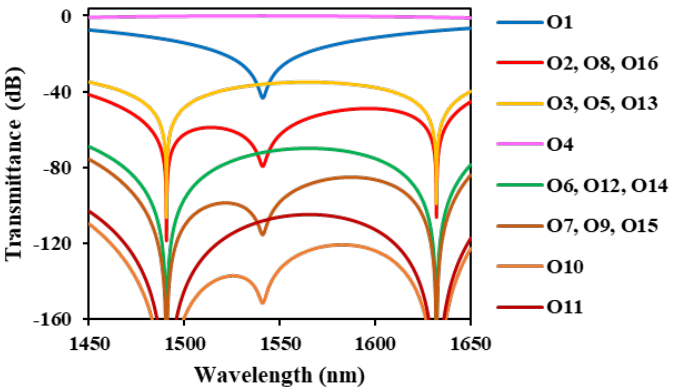
(a)



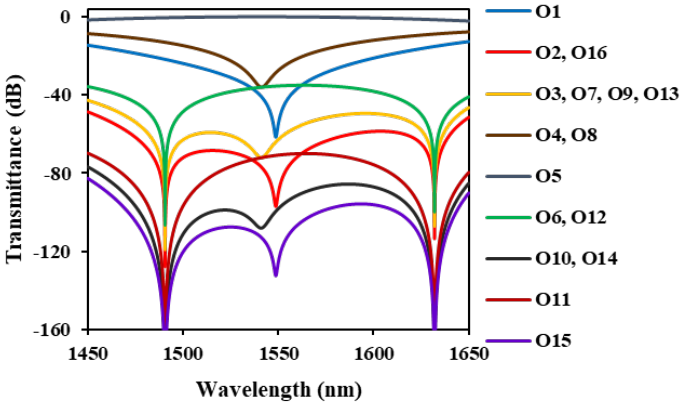
(b)



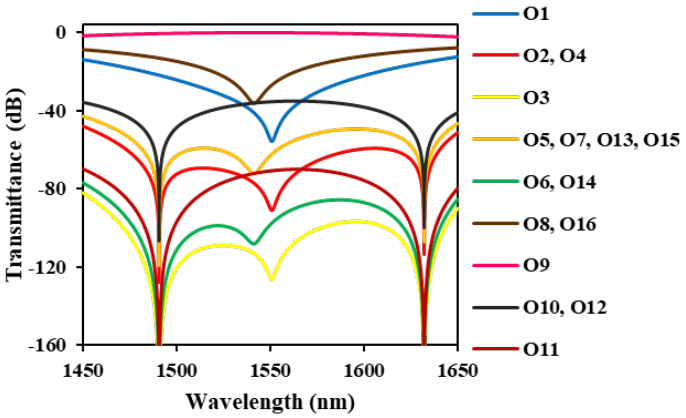
(c)



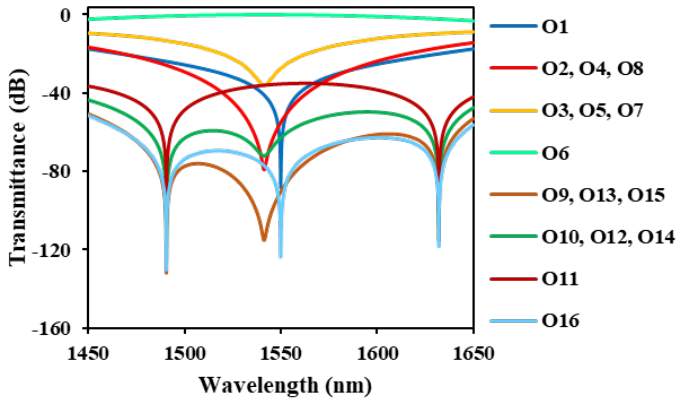
(d)


Transmittance (dB)
0
-40
-80
-120
-160
1450
1500
1550
1600
1650
Wavelength (nm)
O1
O2, O16
O3, O7, O9, O13
O4, O8
O5
O6, O12
O10, O14
O11
O15

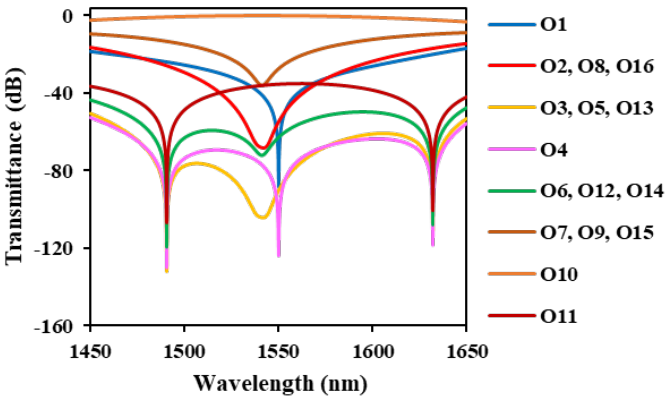
(e)


Transmittance (dB)
0
-40
-80
-120
-160
1450
1500
1550
1600
1650
Wavelength (nm)
O1
O2, O4
O3
O5, O7, O13, O15
O6, O14
O8, O16
O9
O10, O12
O11

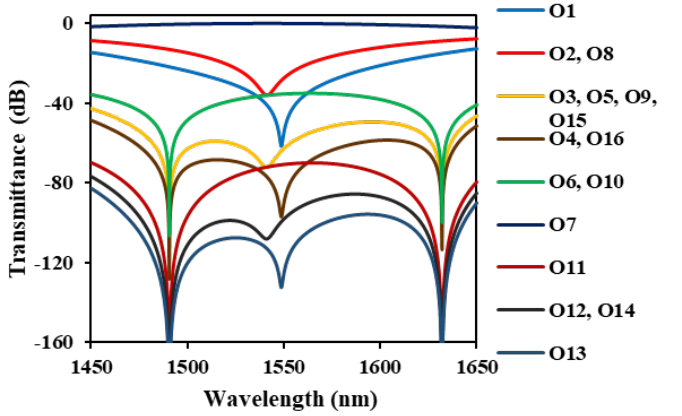
(i)


Transmittance (dB)
0
-40
-80
-120
-160
1450
1500
1550
1600
1650
Wavelength (nm)
O1
O2, O4, O8
O3, O5, O7
O6
O9, O13, O15
O10, O12, O14
O11
O16

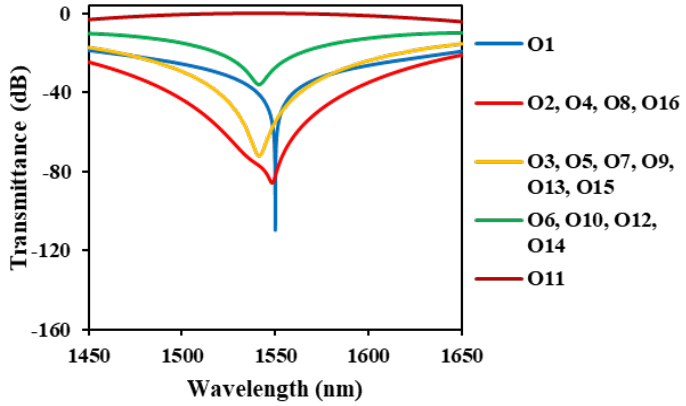
(f)


Transmittance (dB)
0
-40
-80
-120
-160
1450
1500
1550
1600
1650
Wavelength (nm)
O1
O2, O8, O16
O3, O5, O13
O4
O6, O12, O14
O7, O9, O15
O10
O11

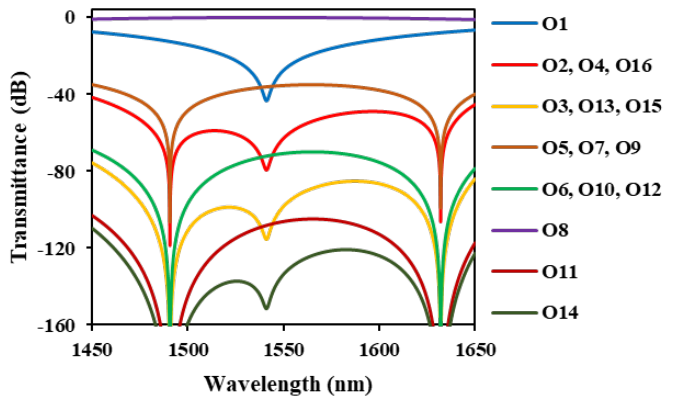
(j)


Transmittance (dB)
0
-40
-80
-120
-160
1450
1500
1550
1600
1650
Wavelength (nm)
O1
O2, O8
O3, O5, O9, O15
O4, O16
O6, O10
O7
O11
O12, O14
O13

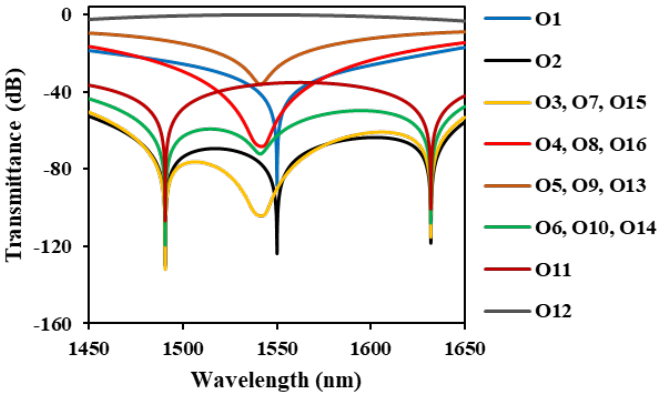
(g)


Transmittance (dB)
0
-40
-80
-120
-160
1450
1500
1550
1600
1650
Wavelength (nm)
O1
O2, O4, O8, O16
O3, O5, O7, O9, O13, O15
O6, O10, O12, O14
O11

(k)


Transmittance (dB)
0
-40
-80
-120
-160
1450
1500
1550
1600
1650
Wavelength (nm)
O1
O2, O4, O16
O3, O13, O15
O5, O7, O9
O6, O10, O12
O8
O11
O14

(h)


Transmittance (dB)
0
-40
-80
-120
-160
1450
1500
1550
1600
1650
Wavelength (nm)
O1
O2
O3, O7, O15
O4, O8, O16
O5, O9, O13
O6, O10, O14
O11
O12

(l)

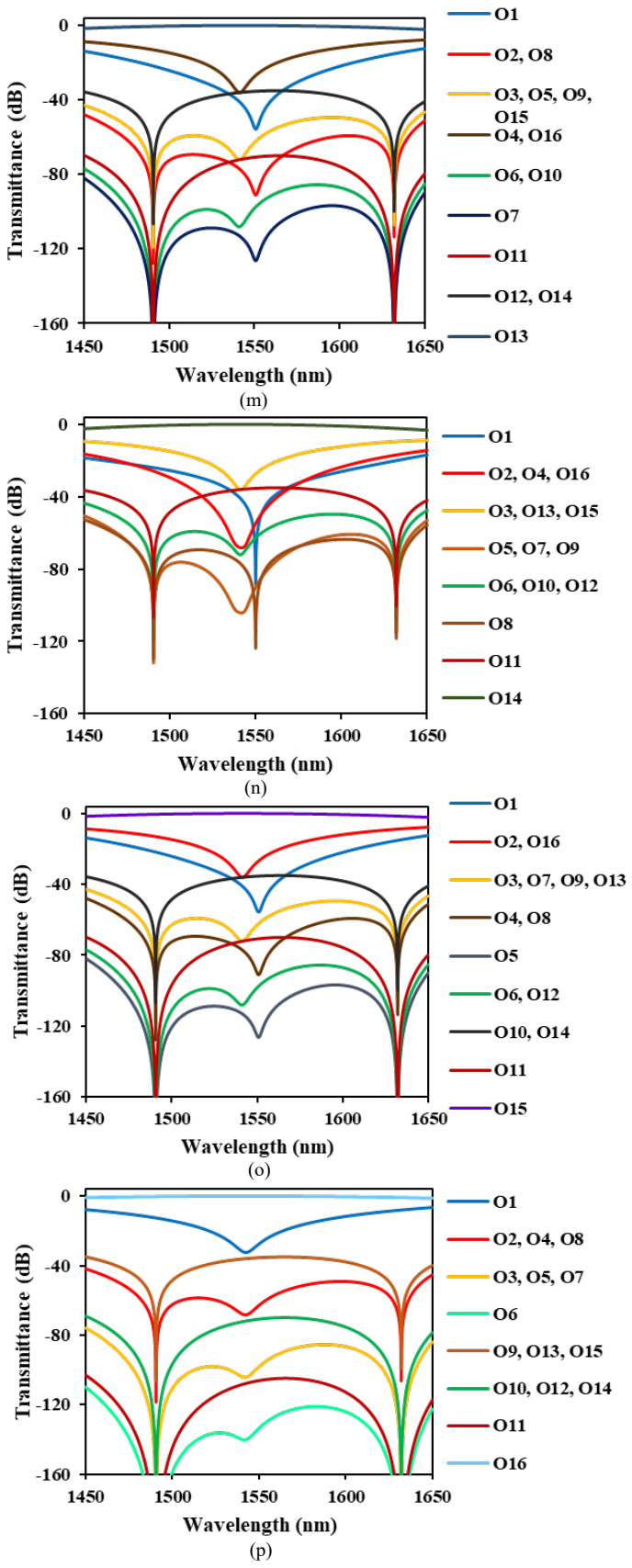


Figure 4: Transmittance of 4-stage nested MZI switch with PGCs when the optical path is set toward O1−O16 (a) O1 (b) O2 (c) O3 (d) O4 (e) O5 (f) O6 (g) O7 (h) O8 (i) O9 (j) O10 (k) O11 (l) O12 (m) O13 (n) O14 (o) O15 (p) O16

We examined the crosstalk characteristics shown in Figures 4(a)−4(p), and found that crosstalks to unrouted ports are suppressed to −40 dB or less over the wavelength band of 200 nm ranging from 1450 to 1650 nm when we choose only O3, O5, O7, O9, O13 and O15 as the output port of the nested MZI optical switch. In Figure 4(c) where the optical input signal is routed to O3, crosstalks to O5, O7, O13 and O15 are identical and less than −40 dB. O9 has much lower crosstalk below −80 dB. When we avoid to use other ports as output, crosstalks to unrouted ports (O5, O7, O9, O13 and O15) never exceed −40 dB. In Figures 4(e), 4(g), 4(i), 4(m) and 4(o), crosstalks remain lower than −40 dB when the optical signal is routed to O5, O7, O9, O13 and O15, respectively.

4.2. Discussion

The crosstalk characteristics of 16 ports shown in Figures 4(a)−4(p) is classified into 5 groups. One is O1, second is O2, O4, O8 and O16, third is O3, O5, O7, O9, O13 and O15, forth is O6, O10, O12 and O14, and the last is O11. Each group has the same number of the MZI states (cross or through) as shown in Table 1. This means that the port settings in the same group have two or more different MZI states among four stages each other. For example, the port settings in the third group (O3, O5, O7, O9, O13 and O15) have two cross states and two through states among four MZI stages. O3 has the MZI states of CCTT, where C or T denotes each MZI state of four stages as shown in Table 1. This state differs in two stages to those of O5 (TCCT), O7 (CTCT), O13 (TCTC) and O15 (CTTC), while in all four stages to O9 (TTCC). Thus, when the optical path is set toward O3, crosstalks to O5, O7, O13 and O15 are blocked in two MZI stages where one is in cross state and the other is in through state, respectively, and that to O9 in four MZI stages where two are in cross state and the other two are in through state. Crosstalks to unrouted ports are blocked similarly when the optical signal is routed to O5, O7, O9, O13 and O15, respectively. This is why the lower crosstalk is given when we choose only O3, O5, O7, O9, O13 and O15 as the output ports of the nested MZI optical switch.

Following the analysis above, we can deduce how many output ports in multi-stage nested MZI switch has low crosstalk. We should choose the port settings which have the same number of the MZI states (cross or through) and equalize the cross and through states. Then the number of output ports N with low crosstalk in an n-stage switch ($n \geq 2$) is given as

$$N = {}_nC_{[n/2]} = \binom{n}{[n/2]}, \tag{8}$$

where $[x]$ is a floor function. In the previous study [16], it has been shown that a 2-stage nested MZI switch has two ports with low crosstalk among four output ports. A 3-stage nested MZI optical switch gives three ports with low crosstalk in eight output ports. This study shows that 4-stage nested MZI switch has six ports with low crosstalk in 16 output ports. Equation (8) agrees with these results, and indicates that 5-stage nested MZI switch must have ten ports with low crosstalk in 32 output ports. Figure 5 shows the required number of stages n for given output port counts N with low crosstalk. The ratio of N to 2^n is less than 50 % and decreases as n increases. However, the conventional MZI optical switch commonly employ two cascaded MZIs per output port to suppress crosstalk sufficiently [16-18]. In that case the ratio $N/2^n$ is around 50 % for the conventional cascaded MZI

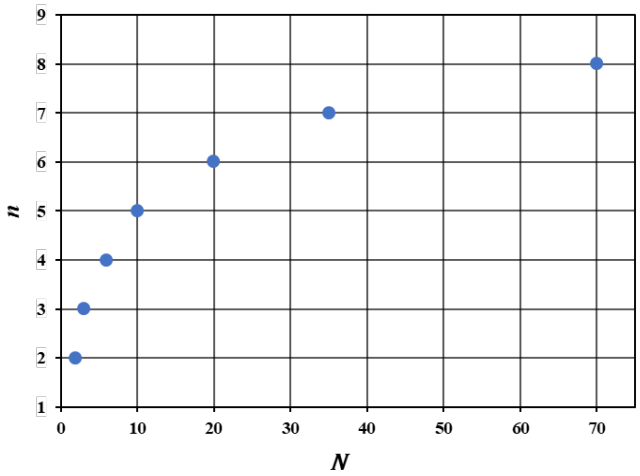


Figure 5: Required number of stages n for given output port counts N with low crosstalk

switch. Although the nested MZI switch has somewhat smaller ratio of the output ports with low crosstalk, it has advantages that its phase shifters are placed in parallel and gives a shorter circuit length than the cascaded MZI optical switch.

Recently, a new concept was demonstrated that effectively reduces the wavelength dependence of an MZI switch [19]. This technique employs no PGCs but arm waveguides with different waveguide widths in each cascaded MZI. It is a future work whether this technique is possible to apply to the nested MZI switch with arrayed arm waveguides.

5. Conclusion

We examined the crosstalk characteristics of a 4-stage nested MZI optical switch with PGCs. It was confirmed that crosstalk is lower than −40 dB over a wavelength band of as wide as 200 nm ranging from 1450 to 1650 nm in six output ports of the 4-stage switch. We showed that crosstalks to these six ports are blocked in two MZI stages where one is in cross state and the other is in through state, respectively. Based on the results, we deduced the required number of stages for given output port counts with low crosstalk in multi-stage nested MZI switch. It is possible to construct 1 × 10, 1 × 20, 1 × 35 optical switches with 5, 6, 7-stages, respectively, and output port count reaches 70 in 8-stage nested MZI switch.

Conflict of Interest

We declare no conflict of interest.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Number JP18K04286.

References

[1] N. Takato, K. Jinguji, M. Yasu, H. Toba, and M. Kawachi, “Silica-based single-mode waveguides on silicon and their application to guided-wave optical interferometers,” Journal of Lightwave Technology, **6**(6), 1003-1010, 1988, doi: 10.1109/50.4091.

[2] S. Sohma, S. Mino, T. Watanabe, M. Ishii, T. Shibata, and H. Takahashi, “Solid-state optical switches using planar lightwave circuit and IC-on-PLC technologies,” Proceedings of SPIE, **5625**, 767-775, 2004, doi: 10.1117/12.579684.

[3] H. Takahashi, T. Watanabe, T. Goh, S. Sohma, and T. Takahashi, “PLC optical switch that enhances the optical communication network,” NTT Technical Review, **3**(7), 17-21, 2005.

[4] C. R. Doerr and K. Okamoto, “Advances in silica planar lightwave circuits,” Journal of Lightwave Technology, **24**(12), 4763-4789, 2006, doi: 10.1109/JLT.2006.885255.

[5] T. Watanabe, K. Suzuki, T. Goh, K. Hattori, A. Mori, T. Takahashi, T. Sakamoto, K. Morita, S. Sohma, and S. Kamei, “Compact PLC-based transponder aggregator for colorless and directionless ROADM,” Technical Digest of 2011 Optical Fiber Communication Conference/National Fiber Optic Engineers Conference (OFC/NFOEC 2011), paper OTuD3, 2011, doi: 10.1364/OFC.2011.OTuD3.

[6] K. Sato, H. Hasegawa, T. Niwa, and T. Watanabe, “A large-scale wavelength routing optical switch for data center networks,” IEEE Communications Magazine, **51**(9), 46-52, 2013, doi: 10.1109/MCOM.2013.6588649.

[7] N. Farrington, A. Forencich, G. Porter, P.-C. Sun, J. E. Ford, Y. Fainman, G. C. Papen, and A. Vahdat, “A multiport microsecond optical circuit switch for data center networking,” IEEE Photonics Technology Letters, **25**(16), 1589-1592, 2013, doi: 10.1109/LPT.2013.2270462.

[8] Q. Cheng, M. Bahadori, M. Glick, S. Rumley, and K. Bergman, “Recent advances in optical technologies for data centers: a review,” Optica, **5**(11), 1354-1370, 2018, doi: 10.1364/OPTICA.5.001354.

[9] S. Katayose, Y. Hashizume, and M. Itoh, “Fabrication and demonstration of 1 × 8 silicon-silica multi-chip switch based on optical phased array,” Japanese Journal of Applied Physics, **55**, 08RB01, 2016, doi: 10.7567/JJAP.55.08RB01.

[10] O. Moriwaki and K. Suzuki, “Fast switching of 84 μs for silica-based PLC switch,” Technical Digest of 2020 Optical Fiber Communication Conference (OFC 2020), paper Th3B.5, 2020, doi: 10.1364/OFC.2020.Th3B.5.

[11] K. Hattori, M. Fukui, M. Jinno, M. Oguma, and K. Oguchi, “PLC-based optical add/drop switch with automatic level control,” Journal of Lightwave Technology, **17**(12), 2562-2571, 1999, doi: 10.1109/50.809678.

[12] H. Takahashi, T. Goh, T. Shibata, M. Okuno, Y. Hibino, and T. Watanabe, “High performance 8-arrayed 1 × 8 optical switch based on planar lightwave circuit for photonic networks,” Proceedings of 28th European Conference on Optical Communication (ECOC 2002), paper 4.2.6, 2002.

[13] T. Watanabe, Y. Hashizume, and H. Takahashi, “Double-branched 1×29 silica-based PLC switch with low loss and low power consumption,” Technical Digest of 17th Microoptics Conference (MOC 2011), paper J-2, 2011.

[14] K. Suzuki, T. Mizuno, M. Oguma, T. Shibata, H. Takahashi, Y. Hibino, and A. Himeno, “Low loss fully reconfigurable wavelength-selective optical 1 × N switch based on transversal filter configuration using silica-based planar lightwave circuit,” IEEE Photonics Technology Letters, **16**(6), 1480-1482, 2004, doi: 10.1109/LPT.2004.827419.

[15] T. Watanabe, K. Tasaki, T. Nagayama, and S. Fukushima, “Nested Mach-Zehnder interferometer optical switch with low crosstalk,” Technical Digest of 24th Microoptics Conference (MOC 2019), paper P-69, 2019, doi: 10.23919/MOC46630.2019.

[16] K. Tasaki, M. Tokumaru, T. Watanabe, T. Nagayama, and S. Fukushima, “Nested Mach-Zehnder interferometer optical switch with phase generating couplers,” Japanese Journal of Applied Physics, **59**, SOOB04, 2020, doi: 10.35848/1347-4065/ab8f09.

[17] T. Mizuno, H. Takahashi, T. Kitoh, M. Oguma, T. Kominato, and T. Shibata, “Mach–Zehnder interferometer switch with a high extinction ratio over a wide wavelength range,” Optics Letters, **30**(3), 251-253, 2005, doi: 10.1364/OL.30.000251.

[18] M. Kawasako, T. Watanabe, T. Nagayama, and S. Fukushima, “4-stage Mach-Zehnder interferometer optical switch with phase generating couplers,” Technical Digest of 26th Microoptics Conference (MOC 2021), paper P-40, 2021, doi: 10.23919/MOC52031.2021.

[19] T. Goh, K. Yamaguchi, and A. Yanagihara, “Multiband optical switch technology,” Technical Digest of 2022 Optical Fiber Communication Conference (OFC 2022), paper W4B.1, 2022, doi: 10.1364/OFC.2022.W4B.1.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 5, 146-156 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Model Order Reduction and Distribution for Efficient State Estimation in Sensor and Actuator Networks

Ferdinand Friedrich*

University of Augsburg, Chair of Control Theory, Augsburg, 86159, Germany

ARTICLE INFO

Article history:
Received: 31 August, 2022
Accepted: 04 October, 2022
Online: 25 October, 2022

Keywords:
Model Order Reduction
Sensor and Actuator Network
Distributed State Estimation

ABSTRACT

We present in this contribution the distribution of a global multi-input-multi-output system in a sensor and actuator network. Based on controllability and observability, the global system is decentralized and the system properties are preserved as a result. This results in multiple decentralized local single-input-single-output systems with the same system order as the global system. As these local systems are implemented on decentralized CPUs in the network, the computational effort of the nodes has to be minimized. This is achieved by approximating the input and output behavior and reducing the system order of the decentralized local systems. For this purpose, the two most common techniques, Balanced Truncation and Krylov subspace methods, are presented. Kalman filters are used for state reconstruction. To approximate the input/output behavior of the global system, information from all decentralized reduced local systems is necessary, thus a fully interconnected network is used for communication. By decentralized fusion algorithms in the network nodes, the Kalman filter algorithm is separated and distributed in the network.

1 Introduction

This paper is an extension of "Reduced and Distributed Estimation in Sensor and Actuator Networks - Automated Design Based on Controllability and Observability" presented at the IEEE Conference on Control Technology and Applications (CCTA) 2021 in San Diego (virtual) and "Model-Order Reduction and System Distribution Using Krylov Subspaces - An Approach for Efficient State Estimation in Sensor and Actuator Networks" presented at the IEEE Conference on Control Technology and Applications (CCTA) 2022 in Trieste [1], [2].

The requirements for the accuracy of mathematical models are constantly increasing. Also large-scale MIMO[1] systems are not protected from the increasing requirements. These and the increasing complexity of the systems enhance the system order of the dynamic large-scale MIMO systems [3], [4]. The use of such complex and large-scale models in simulation is very cumbersome and sometimes impossible. Moreover, system analysis and controller design using known methods is also impossible. For this reason, methods for linear and nonlinear systems are developed to reduce large models to a manageable size and to approximate the characteristic properties related to the input/output behavior [5]. Memory and computational requirements are reduced by distributing the system so that only parts of the model are available for local systems [6]. However, this requires decentralized estimation procedures.

If the internal model relationships (e.g. mechanical and electrical) are known, the distribution into local systems is performed manually [6]. Furthermore, system digraphs and cut-point sets are used for model distribution [7]. In [8], a manual method for distributing large systems is also presented. In this method, reduced local systems with overlapping state vectors are created. All described manual methods share the property that the actual system order of the global system is preserved in the network despite the distribution into local systems.

In contrast, we use decentralization for distribution based on the controllability and observability of the global MIMO system. By decoupling the inputs and outputs, several local SISO[2] systems are formed from the global MIMO system. These decentralized local systems are implemented in a fully interconnected network. Moreover, we use model order reduction techniques, resulting in the decentralized local systems with much lower order and preserving the system properties [9]. Further, the computational effort in each network node is reduced.

Several techniques for reducing the model order are known [10]. The most common techniques for approximating the input and output behavior are Balanced Truncation and *Krylov* subspace methods.

*Corresponding Author: Ferdinand Friedrich, University of Augsburg, Chair of Control Theory, Universitaetsstrasse 2, 86159 Augsburg, ferdinand.friedrich@uni-a.de
[1]Multi-Input-Multi-Output
[2]Single-Input-Single-Output

In Balanced Truncation, the Hankel singular values are analyzed and the states with a small impact on the energy transport are neglected and truncated [11]. In *Krylov* subspace methods, to approximate the input and output behavior, the transfer functions of the global and reduced systems are expanded at an arbitrary development point. In this process, the first coefficients, also called moments, of the Taylor series of the transfer function are matched [12], [13].

Therefore, the methods are also called moment matching with the *Padé*-type approximation method based on rational interpolation [14], [15]. If only a single development point (several are also possible) $s = 0$ is chosen, the reduction is ascribed to an *padé*-approximation [16]. A more extensive background on *Krylov* subspace methods is given in [10], [17], and [18]. Model order reduction with *Krylov* subspaces requires less memory and computational effort (compared to Balanced Truncation) [19]. Meanwhile, several algorithms for the Krylov subspaces are known. In this paper, we only use the *Arnoldi* algorithm with its extension to MIMO systems [20], [21].

This paper is organized as follows. In Section 2, the systems under consideration are defined. The methods used to reduce the model order are presented in Section 3. The distribution and deployment in a network are demonstrated in Section 4. A demonstrator is used for detailed real-time experimental evaluation of the presented methods and the results are presented in Section 5.

2 Preliminaries

In this contribution, we refer to global multi-input-multi-output linear time-invariant systems of the form

$$\mathbf{\Sigma}_G = \begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) \end{cases} \tag{1}$$

Here, $\mathbf{A} \in \mathbb{R}^{n\times n}$ is the system matrix and $\mathbf{B} \in \mathbb{R}^{n\times p}$ is the input matrix. The matrix $\mathbf{C} \in \mathbb{R}^{q\times n}$ ist the output matrix and $\mathbf{x}$, $\mathbf{u}$ as well as $\mathbf{y}$ are the state vector, the inputs and the outputs of the global continuous-time system $\mathbf{\Sigma}_G$.

3 Model Order Reduction

Precise modeling of technical systems often leads to large systems of differential equations, requiring a high computational effort. For this reason, methods are used to reduce the order of the systems. Thereby the characteristic properties are supposed to be preserved and the input/output behavior has to be maintained as exactly as possible. The linear reduction methods can be categorized into three classes. One class is the modal reduction methods, in which less dominant eigenvalues and eigenmodes are removed from the system representation. The second class of reduction methods, which is also used in this contribution, are the balanced methods. Here, energy-oriented analyses are used to remove an unimportant subsystem. The third class describes the Krylov subspace methods. In this method, certain parameters of the transfer function, the so-called moments, are preserved in the reduced model. This class is also used for decentralization and distribution in this contribution.

3.1 Balanced Truncation

The Balanced Truncation (BT) method is based on the system properties of controllability and observability. Different methods for determining controllability and observability are known. For balanced truncation, the *Gramian* controllability matrix and the *Gramian* observability matrix shall be used.
A system (1) is controllable if a control vector $\mathbf{u}(t)$ exists that transforms the system from an initial state $\mathbf{x}(t = 0) = \mathbf{0}$ to a final state in finite time. The controllability *Gramian* is defined as

$$\mathbf{W}_c = \int_0^\infty e^{\mathbf{A}t}\,\mathbf{B}\mathbf{B}^T\,e^{\mathbf{A}^T t}\,dt. \tag{2}$$

Furthermore, a system (1) is fully observable if the system state $\mathbf{x}(t = 0)$ can be determined by observing the output $\mathbf{y}(t)$ over a finite time interval. The *Gramian* observability matrix is derived as follows

$$\mathbf{W}_o = \int_0^\infty e^{\mathbf{A}^T t}\,\mathbf{C}^T\mathbf{C}\,e^{\mathbf{A}t}\,dt. \tag{3}$$

For the transfer of the system (1) from the initial position $\mathbf{x}_0$ to the final position $\mathbf{x}_e$ in infinite time, the energy $\mathbf{x}_e^T\mathbf{W}_c^{-1}\mathbf{x}_e$ must be applied. In contrast, the energy $\mathbf{x}_0^T\mathbf{W}_o\mathbf{x}_0$ is generated when the system (1) in the initial position $\mathbf{x}_0$ can oscillate uninfluenced ($\mathbf{u}(t) = 0$) to its rest position. This energy-based approach reveals that a difficultly controllable state requires a high level of energy to achieve.This can also be observed in the eigenvalues of the matrix $\mathbf{W}_c^{-1}$. The more energy is needed to reach a specific state, the larger is its eigenvalue. This analysis can also be used for observability. However, here a small eigenvalue of the state determines that it is badly observable, because only a small amount of energy is visible.
By a state transformation the system $\mathbf{\Sigma}_G$ with order n is transformed into a balanced state-space representation, whereby the *Gramian* matrices

$$\mathbf{W}_c = \mathbf{W}_o = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{bmatrix} \tag{4}$$

are diagonal and identical [22]. In [23] it was demonstrated that any complete controllable and observable system can be transformed into a balanced state-space representation. The diagonal elements σ_i of (4) are the *Hankel singular values* (HSV). The HSV can be determined by

$$\sigma_i = \sqrt{\lambda_i \mathbf{W}_c\mathbf{W}_o}. \tag{5}$$

Here, λ_i represents the eigenvalues of the system $\mathbf{\Sigma}_G$. The HSVs are invariant to transformations, so small HSVs correspond to badly controllable and observable states [22], [24]. The transformation matrix to the balanced state-space representation of (1) are determined in four steps:

1. In [25] it is shown that the controllability (2) and the observability matrix (3) are also solutions of the *Lyapunov* equation

$$\mathbf{A}\mathbf{W}_c + \mathbf{W}_c\mathbf{A}^T + \mathbf{B}\mathbf{B}^T = \mathbf{0} \tag{6}$$

$$\mathbf{A}^T\mathbf{W}_o + \mathbf{W}_o\mathbf{A} + \mathbf{C}^T\mathbf{C} = \mathbf{0} \tag{7}$$

2. After solving the *Lyapunov* equation, *Gramian* matrices are positive definite and can be decomposed into *Cholesky* factors: For the solution of the *Lyapunov* equation, one positive

definite solution for $\mathbf{W}_c$ and $\mathbf{W}_o$ exists for every fully controllable and observable system. The *Gramian* matrices can be decomposed into *Cholesky* factors

$$\mathbf{W}_c = \mathbf{S}^T\mathbf{S} \tag{8}$$

$$\mathbf{W}_o = \mathbf{R}^T\mathbf{R} \tag{9}$$

3. The product $\mathbf{SR}^T$ is subjected to a singular value decomposition (SVD),
$$\mathbf{SR}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{10}$$
which yields the orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$, and the diagonal matrix $\mathbf{\Sigma}$ of HSVs.

4. The matrix $\mathbf{T}$ for transformation to balanced state-space representation is
$$\mathbf{T} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{V}^T\mathbf{R} \tag{11}$$
and its inverse is
$$\mathbf{T}^{-1} = \mathbf{S}^T\mathbf{U}\mathbf{\Sigma}^{-\frac{1}{2}} \tag{12}$$

In [26] it is presented that these matrices (11) and (12) transform a fully controllable and observable system in balanced state-space representation.

$$\mathbf{\Sigma}_{Bal} = \begin{cases} \dot{\mathbf{x}}_{Bal}(t) & = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}\mathbf{x}_{Bal}(t) + \mathbf{T}^{-1}\mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) & = \mathbf{C}\mathbf{T}\mathbf{x}_{Bal}(t) \end{cases} \tag{13}$$

The HSVs reveals that not all states are equally well controllable and observable. If only the states which contributes much to the energy transfer are used, then only the first r HSVs of $\mathbf{\Sigma}$ and the first r columns of $\mathbf{U}$ as well as $\mathbf{V}$ need to be considered. BT can also be interpreted as a *Petrov-Galerkin* projection into a reduced subspace [22]. For this projection, the transformation matrices (11) and (12) are modified.

$$\mathbf{W}_{BT}^T = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{U}^T\mathbf{R}^T \tag{14}$$

$$\mathbf{V}_{BT} = \mathbf{S}\mathbf{V}\mathbf{\Sigma}^{-\frac{1}{2}} \tag{15}$$

Where $\mathbf{W}_{BT} \in \mathbb{R}^{r\times n}$ and $\mathbf{V}_{BT} \in \mathbb{R}^{r\times n}$ are the biorthogonal projection matrices. By the projection matrices only the first r columns are considered, thus the system order $r < n$ is reduced. This truncates only the HSVs with a small participation in the energy transfer, thus the input/output behavior is only marginally affected.

3.2 Krylov Subspaces

If numerically robust and efficient computations are required for model order reduction, the reduction must be performed by *Krylov* subspaces. Additionally, this method is used for very large systems. In [21] and [27] the *Krylov* subspace is generally defined. Let $\mathbf{M}$ be an arbitrary constant matrix and $\mathbf{v}$ a constant vector, called starting vector, then the *Krylov* subspace is defined as follows

$$\mathcal{K}_r(\mathbf{M}, \mathbf{v}) = \text{span}\left\{\mathbf{v}, \mathbf{M}\mathbf{v}, \ldots, \mathbf{M}^{r-1}\mathbf{v}\right\}. \tag{16}$$

Here, $\mathbf{M} \in \mathbb{R}^{n\times n}$, $\mathbf{v} \in \mathbb{R}^n$ are the arbitrary matrix and vector as well as $\mathcal{K}_r \subseteq \mathbb{R}^n$ is a r-dimensional subspace. For model order reduction with *Krylov* subspaces also transformation matrices $\mathbf{W}$ and $\mathbf{V}$ are used. The matrices are composed of column vectors, which are also the basis of the *Krylov* subspaces.
Referring to the system (1), the transformation matrix $\mathbf{V}$ as an arbitrary basis of the input subspace is determined by

$$\mathcal{K}_{r1}\left(\mathbf{A}^{-1}, \mathbf{A}^{-1}\mathbf{B}\right). \tag{17}$$

In addition, the transformation matrix $\mathbf{W}$ is derived as the basis of the output subspace

$$\mathcal{K}_{r2}\left(\mathbf{A}^{-T}, \mathbf{A}^{-T}\mathbf{C}^T\right). \tag{18}$$

For the reduction, $r_1 = r_2 = r$ must be required, whereby it must be ensured that both transformation matrices $\mathbf{V}$ and $\mathbf{W}$ have full rank r. The *Arnoldi* algorithm is used to determine the transformation matrices $\mathbf{V}$ and $\mathbf{W}$. Algorithm 1 shows the calculation of the transformation matrix $\mathbf{V}$ based on (17).

Algorithm 1: MIMO Arnoldi Algorithm [21]

Result: Transformation matrix $\mathbf{V}$
Set $\mathbf{v}_1 = \frac{\mathbf{b}_1}{\sqrt{\mathbf{b}_1^T\mathbf{b}_1}}$ and $\mathbf{b}_1$is the first starting vector;
for *i=2,3,...* **do**
  if $i \leq m_i$ **then**
    Next vector is the i-th starting vector
  else
    $\mathbf{r}_i = \mathbf{A}_1\mathbf{v}_{i-m_1}$
  end
  Set $\hat{\mathbf{v}}_i = \mathbf{r}_i$ ▷ Orthogonalization
  for $j = 1, \ldots, i-1$ **do**
    h = $\hat{\mathbf{v}}_i^T\mathbf{v}_j$
    $\hat{\mathbf{v}}_i = \hat{\mathbf{v}}_i - h\mathbf{v}_j$
  end
  if $\hat{\mathbf{v}}_i = 0$ **then**
    Reduce m_1 to $m_1 - 1$
    m_1 is nonzero calculate new $\mathbf{r}_i$
  else
    $\mathbf{v}_i = \frac{\hat{\mathbf{v}}_i}{\sqrt{\hat{\mathbf{v}}_i^T\hat{\mathbf{v}}_i}}$ ▷ i-th column of matrix $\mathbf{V}$
  end
end

The same algorithm and (18) is used to determine the matrix $\mathbf{W}$. The columns of the transformation matrices are spanning the corresponding *Krylov* input and output subspace (Fig. 1). If the model order reduction is performed based on (17) and (18) with matrices $\mathbf{V}$ and $\mathbf{W}$, the reduction procedure is called two-sided. Obviously, only one of the two matrices can be used. The remaining transformation matrix is chosen arbitrarily, but it must have full rank. The reduction procedure is then called one-sided.

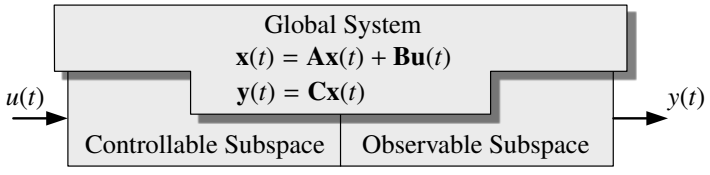


Figure 1: Controllability and observability are major properties and subspaces of dynamical control systems.

In *Krylov* subspace methods, the transfer functions of the original system and the reduced system are expanded at an arbitrary development point s_0, thus the first moments (coefficients of the *Taylor*

series) match. In this contribution, the development point $s_0 = 0$ is chosen, whereby the transfer function of (1) is described by a *Taylor* series around the developing point. The resulting i-th moment of the series around $s = 0$ is

$$\mathbf{m}_i = \mathbf{C}\mathbf{A}^{-(i+1)}\mathbf{B} \quad \text{with } i = 0, 1, \ldots \tag{19}$$

One-sided *Krylov* reduction methods reduce the original system to order r. In this case, the first r moments of the original system and the reduced system are matched. In the case of a two-sided method, the first $2r$ moments are matched.
The moment $\mathbf{m}_0$ corresponds to the stationary gain and is preserved in the *Krylov*-based reduction, so the reduced system is equally stationary accurate. Two-sided *Krylov* methods adjust two times as many moments for the same reduction size compared to one-sided methods. For this reason, two-sided methods are usually preferable. In addition, they offer the advantage that the reduction is independent of the representation of the original model. In contrast, regular state transformations and other equivalent transformations of the original model in one-sided reduction affect the transfer behavior in the reduced model [28]. Using a projection, the state vector $\mathbf{x}(t)$ is approximated by the reduced state vector $\mathbf{x}_r(t)$ as follows

$$\mathbf{x}(t) = \mathbf{V}\mathbf{x}_r(t). \tag{20}$$

Here, $\mathbf{V} \in \mathbb{R}^{n \times r}$ is the basis of the *Krylov* input subspace and is determined using algorithm 1. $\mathbf{x}_r \in \mathbb{R}^r$ is the state vector of the reduced system.
By applying the *Arnoldi* algorithm twice with different subspaces (17) or (18), the reduced system with the transformation matrices can be determined as follows.

$$\boldsymbol{\Sigma}_r = \begin{cases} \dot{\mathbf{x}}_r(t) &= \mathbf{W}^T\mathbf{A}\mathbf{V}\mathbf{x}_r(t) + \mathbf{W}^T\mathbf{B}\mathbf{u}(t) \\ \mathbf{y}_r(t) &= \mathbf{C}\mathbf{V}\mathbf{x}_r(t) \end{cases} \tag{21}$$

4 Model Distribution

Decentralization and distribution of global large-scale MIMO systems in a sensor and actuator network achieve order reduction, as individual nodes never need to estimate the entire global state vector. Often, this distribution occurs due to internal system relationships or due to physical system boundaries. In this section, we present two approaches for decentralization and distribution, both based on controllability and observability of the global system. One approach demonstrates decentralization into local SISO systems based on the inputs and outputs of the global MIMO system. The order of the local systems is reduced to further minimize the computational effort using model order reduction techniques. In a further approach, the local SISO systems will be generated based on the controllable and observable subspace. The order reduction resulting from decentralization and distribution is reflected in a reduced computational effort for each node. As a consequence, the required computational effort in a local node is lower than in a centralized system.

4.1 Input-/Output-based

By decentralizing the inputs and outputs of (1), m local systems with an equal number of sensors and actuators are derived. Thereby, the outputs of the global system (1) are decentralized without further considerations. To decentralize the inputs, the physical couplings must be considered. For this purpose, the decoupling method from [29] can be used. It is assumed that the inputs have a localized impact in large-scale systems. The inputs of the global system are decentralized into local systems as in [1]. Decentralization and distribution provides the ability to distribute the system in a network. In a fully connected network, minimal non-modeled couplings between the inputs are considered by communication.
Thus, m local SISO systems are derived from the global system $\boldsymbol{\Sigma}_G$. During decentralization, local systems with one input and one output each are expanded, but the system order n is not reduced (Fig. 2).

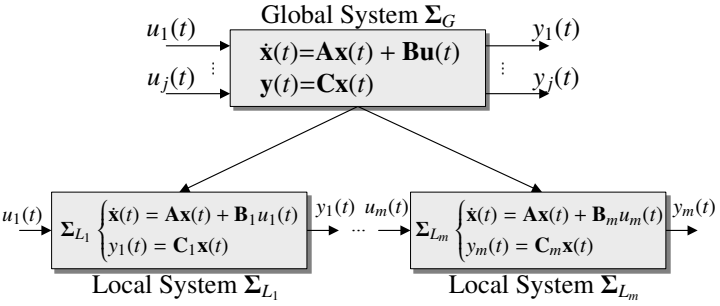


Figure 2: Decentralization of the global MIMO system with j inputs and outputs into m local systems.

The controllability and observability of the global system were not modified by decentralization. It can be demonstrated that by superposing the local *Gramian* controllability matrix $\mathbf{W}_{c_j}$ and the local observability matrix $\mathbf{W}_{o_j}$, the global controllability $\mathbf{W}_\mathbf{c}$ and observability $\mathbf{W}_o$ are recovered. The local systems are described as follows

$$\boldsymbol{\Sigma}_{L_j} = \begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}_j u_j(t) \\ y_j(t) = \mathbf{C}_j\mathbf{x}(t) \end{cases} \quad \text{with } j = 1, \ldots, m. \tag{22}$$

Where m describes the number of sensors and actuators. The index j at the input matrix $\mathbf{B}$ and at the output matrix $\mathbf{C}$ represents the corresponding column and row of the matrices. This defines local SISO systems $\boldsymbol{\Sigma}_{L_j}$ with one sensor and actuator each.
The local systems (22) are implemented on different decentralized CPUs, which communicate through a fieldbus system, and are used in a network.

4.1.1 Application in a Sensor and Actuator Network

Each local system $\boldsymbol{\Sigma}_{L_j}$ describes a network node where a Kalman Filter (KF) is used for state reconstruction [30]. The computational effort in the network nodes is further reduced if the KF algorithm is separated [1]. In the prediction step the system inputs are used and in the correction step, the observations are processed. Both filter steps communicate through a priori and a posteriori estimation $\hat{\mathbf{x}}(k)$ and covariance $\mathbf{P}(k)$. As the inputs are only applied in the prediction step, this part of the algorithm is implemented in the actuator nodes. In the sensor nodes, the observations are applied and only the correction step of the KF algorithm is executed. *Splitting the Kalman Filter* (SKF) provides the distribution of the local system (22) in a

network of sensor and actuator nodes. Each network node only has to calculate a part of the KF algorithm. By exchanging the a priori and a posteriori information, both filtering steps are connected in the network.

Therefore, the continuous-time dynamics of (22) must be transformed into discrete-time dynamics. To further minimize the computational effort, it is necessary to reduce the order of the discrete decentralized local systems. For this purpose, the model order reduction techniques presented in Section 3 are used.

$$\boldsymbol{\Sigma}_{L_{r,j}} = \begin{cases} \dot{\mathbf{x}}_r(t) = \mathbf{W}^T\mathbf{A}\mathbf{V}\mathbf{x}_r(t) + \mathbf{W}^T\mathbf{B}_j u_j(t) \\ y_{r,j}(t) = \mathbf{C}_j\mathbf{V}\mathbf{x}_r(t) \end{cases} \quad (23)$$

with $j = 1, \ldots, m$.

The decentralized reduced local systems (23) can also be used in the SKF. Depending on the reduction method, the projection matrices $\mathbf{V}$ and $\mathbf{W}$ are determined by Section 3.1 or Section 3.2. However, this turns the SKF into the *Splitted Reduced Kalman Filter* (SRKF)

Prediction

$$\hat{\mathbf{x}}_r^-(k) = \mathbf{A}_r(k-1)\hat{\mathbf{x}}_r^+(k-1) + \mathbf{B}_r(k-1)u_j(k-1) \quad (24)$$

$$\mathbf{P}_r^-(k) = \mathbf{A}_r(k-1)\mathbf{P}_r^+(k-1)\mathbf{A}_r^T(k-1) + \mathbf{Q}(k-1) \quad (25)$$

Correction

$$\mathbf{K}_r(k) = \mathbf{P}_r^-(k)\mathbf{C}_r^T(k)\left[\mathbf{C}_r(k)\mathbf{P}_r^-(k)\mathbf{C}_r^T(k) + \mathbf{R}(k)\right]^{-1} \quad (26)$$

$$\hat{\mathbf{x}}_r^+(k) = \hat{\mathbf{x}}_r^-(k) + \mathbf{K}_r(k)\left[y_j(k) - \mathbf{C}_r(k)\hat{\mathbf{x}}_r^-(k)\right] \quad (27)$$

$$\mathbf{P}_r^+(k) = \left[\mathbf{I} - \mathbf{K}_r(k)\mathbf{C}_r(k)\right]\mathbf{P}_r^-(k)\left[\mathbf{I} - \mathbf{K}_r(k)\mathbf{C}_r(k)\right]^T + \mathbf{K}_r(k)\mathbf{R}_r(k)\mathbf{K}_r^T(k). \quad (28)$$

The index r expresses that a reduced model $\boldsymbol{\Sigma}_{L_{r,j}}$ is used in the decentralized local nodes. The index j in (24) and (27) refers to the corresponding local system.

The distribution of prediction (24), (25) and correction (26),(27) as well as (28) of the local reduced systems (23) in sensor and actuator network nodes, is connected by the communication of the reduced a priori and a posteriori information. A fully connected network is formed if all m local reduced systems $\boldsymbol{\Sigma}_{L_{r,j}}$ communicate. This requires each sensor node to communicate its reduced a posteriori information to all local actuator nodes. This information is processed in the actuator nodes in a fusion step of the SRKF and used in determining the reduced a priori information. The SRKF becomes the *Decentralized Splitted Reduced Kalman Filter* (DSRKF) (Fig. 3). In [31], [32] and [33], a different notation is proposed for the a posteriori values, as the correction is performed using only the local observation.

$$\tilde{\mathbf{P}}_{r,j}^+(k) = \left[\left(\mathbf{P}_{r,j}^-(k)\right)^{-1}\mathbf{C}_{r,j}^T(k)\mathbf{R}_{r,j}^{-1}(k)\mathbf{C}_{r,j}(k)\right]^{-1} \quad (29)$$

$$\tilde{\mathbf{x}}_{r,j}^+ = \hat{\mathbf{x}}_{r,j}^-(k) + \mathbf{K}_{r,j}(k)\left[y_i(k) - \mathbf{C}_{r,j}(k)\hat{\mathbf{x}}_{r,j}^-\right] \quad (30)$$

For the interaction in the network between each local system, the information of the reduced covariance error and the reduced state error are used. These errors are determined in the corresponding local sensor nodes and transmitted to the actuator nodes. Thus, all actuator nodes receive the reduced error information of the sensor nodes. The reduced covariance error and reduced state error for the decentral reduced local system $\boldsymbol{\Sigma}_{L_{r,j}}$ is determined by

$$\mathbf{E}_{r,j}(k) = \left(\tilde{\mathbf{P}}_{r,j}^+(k)\right)^{-1} - \left(\mathbf{P}_{r,j}^-(k)\right)^{-1} \quad (31)$$

$$\mathbf{e}_{r,j}(k) = \left(\tilde{\mathbf{P}}_{r,j}^+(k)\right)^{-1}\tilde{\mathbf{x}}_{r,j}^+(k) - \left(\mathbf{P}_{r,j}^-(k)\right)^{-1}\hat{\mathbf{x}}_{r,j}^-(k) \quad (32)$$

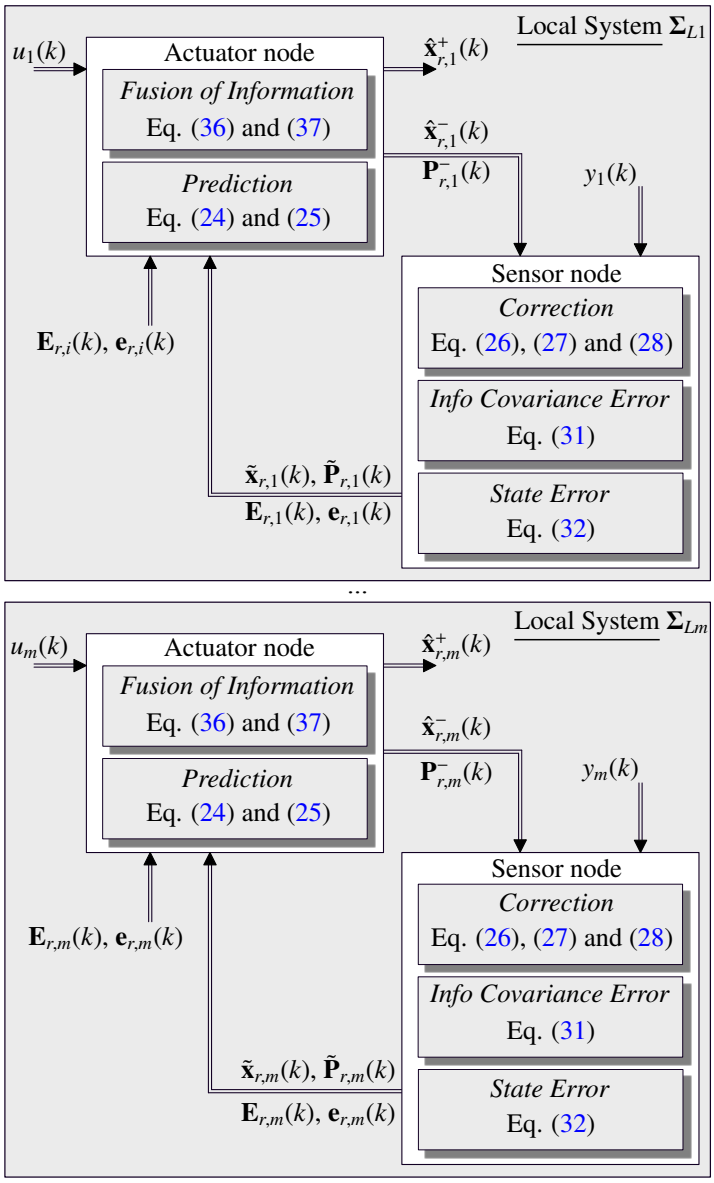


Figure 3: Splitting the KF algorithm on the actuator and sensor nodes. In all nodes reduced models are implemented (SRKF).

Each local system is projected by the matrices $\mathbf{V}$ and $\mathbf{W}$ into its own reduced subspace. In order to exchange information between the subspaces, the information, for instance the reduced error covariances $\mathbf{E}_{r,j}$ and the reduced state error vectors $\mathbf{e}_{r,j}$, must be transformed. The transformation from the transmitter subspace to the receiver subspace is performed in the receiving local actuator node. For the transformation matrices proposed in [31], we use the matrices already determined by the model order reduction in Section 3. These matrices are not symmetric, thus the *Moore-Penrose* inverse is used and specified by the superscript †. If the projection matrices

are determined by the method of Section 3.1, the transformation between the local subspaces are performed as follows

$$\mathbf{V}_{r,ji} = \mathbf{W}_{BT,i}\mathbf{W}^{\dagger}_{BT,j}. \tag{33}$$

For the state of clarity, the transformation using the methods form Section 3.2 is identical. As mentioned above, small couplings between inputs are considered in a fully networked sensor and actuator network. The error information of local system j are transformed by

$$\mathbf{E}_{r,ji}(k) = \left(\mathbf{V}^{T}_{r,ji}\right)^{\dagger}\mathbf{E}_{r,j}(k)\mathbf{V}^{\dagger}_{r,ji} \tag{34}$$

$$\mathbf{e}_{r,ji}(k) = \left(\mathbf{V}^{T}_{r,ji}\right)^{\dagger}\mathbf{e}_{r,j}(k) \tag{35}$$

into the subspace of the system i. By adding fusion equations

$$\left(\mathbf{P}^{+}_{r,j}(k)\right)^{-1} = \left(\mathbf{P}^{-}_{r,i}(k)\right)^{-1} + \sum_{j=1}^{m}\mathbf{E}_{r,ji}(k) \tag{36}$$

$$\hat{\mathbf{x}}^{+}_{r,j}(k) = \mathbf{P}^{+}_{r,j}(k)\left[\left(\mathbf{P}^{-}_{r,j}(k)\right)^{-1}\hat{\mathbf{x}}^{-}_{r,j}(k) + \sum_{j=1}^{m}\mathbf{e}_{r,ji}(k)\right] \tag{37}$$

to each actuator node, a fully interconnected network can be created.

Another approach uses for fusion all state estimates and covariance matrices of the decentralized reduced local systems (23). Here, the estimated a posteriori state vectors $\hat{\mathbf{x}}^{+}$ and a posteriori covariance matrix $\mathbf{P}^{+}$ are transmitted to a central fusion center. There is no direct communication between the locally reduced systems.
The fusion is performed in the center by the *Generalized Millman's Formula* (GMF). The GMF combines $\hat{\mathbf{x}}^{+}$ and $\mathbf{P}^{+}$ for multisensor systems [34], [35]. Assuming that m local systems estimate the vector $\mathbf{x} \in \mathbb{R}^{n}$ with $\hat{\mathbf{x}}$, these estimates are applied to determine the related local error covariance

$$\mathbf{P}_{i,j} = \mathrm{cov}\left(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j\right) \tag{38}$$

$$\tilde{\mathbf{x}}_i = \mathbf{x} - \hat{\mathbf{x}}_i \tag{39}$$

$$i, j = 1, \ldots, m.$$

The objective of GMF is to find the optimal estimate of $\mathbf{x}$. [35] presents that this is achieved if the individual $\boldsymbol{\xi}_i$ are weighted and summed

$$\hat{\mathbf{x}} = \sum_{i=1}^{m}\boldsymbol{\xi}_i\hat{\mathbf{x}}_i. \tag{40}$$

In addition, it is shown that the weighting matrix $\boldsymbol{\xi}_i$ is found by minimizing the mean square error criterion, and that the minimization leads to linear equations

$$\sum_{i=1}^{m-1}\boldsymbol{\xi}_i\left(\mathbf{P}_{i,m-1} - \mathbf{P}_{i,m}\right) + \boldsymbol{\xi}_m\left(\mathbf{P}_{m,m-1} - \mathbf{P}_{m,m}\right) = \mathbf{0} \tag{41}$$

$$\sum_{i=1}^{m}\boldsymbol{\xi}_i = \mathbf{I}. \tag{42}$$

Therefore, the fusion of two local state vectors is proved to be equivalent to the well known *Bar-Shalom-Campo* formula. A more detailed background is given by [34] and [36].

Each decentralized reduced local system uses a KF for state reconstruction and communicates the estimated states and covariances to a fusion center (Fig. 4).

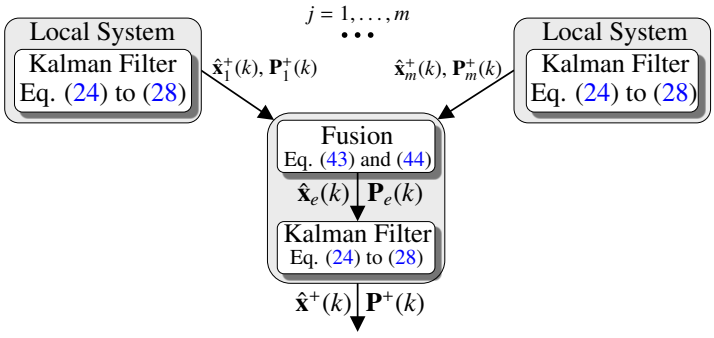


Figure 4: The state reconstruction in the local nodes is performed by a KF with a reduced model.

For independent estimates, GMF has a simple closed form

$$\hat{\mathbf{x}}_e = \mathbf{P}_e\sum_{j=1}^{m}\mathbf{P}^{+^{-1}}_{j}\hat{\mathbf{x}}^{+}_{j} \tag{43}$$

$$\mathbf{P}^{-1}_{e} = \sum_{j=1}^{m}\mathbf{P}^{+^{-1}}_{j}. \tag{44}$$

Due to decentralization and model order reduction, the decentralized reduced systems are transformed into different subspaces. The decentralized reduced estimates of the local systems have to be transformed into a common subspace for fusion. For this purpose, a reference system is derived from the global system (1). The reference system is described as follows

$$\boldsymbol{\Sigma}_{Ref} = \begin{cases}\dot{\mathbf{x}}_{Ref}(t) = \mathbf{W}^{T}\mathbf{A}\mathbf{V}\mathbf{x}_{Ref}(t) + \mathbf{W}^{T}\mathbf{B}\mathbf{u}(t)\\ \mathbf{y}_{Ref}(t) = \mathbf{C}\mathbf{V}\mathbf{x}_{Ref}(t).\end{cases} \tag{45}$$

In the model order reduction, the reference system is adapted to the order of the decentralized reduced local systems. In deriving the reference system, attention was given only to the stability and not to the quality of the approximation. At the fusion center, the transmitted information are considered as uncorrelated as they originate from different subspaces. After the received information is transformed into the reference system subspace, it is fused (43) and (44). Based on the reference model, a KF is implemented in the fusion center. This interprets the fusion $\hat{\mathbf{x}}_e$ as an observation and this improves the fusion result.

4.2 Subspace based

Using the input subspace $\mathcal{K}_{r1}$ and the output subspace $\mathcal{K}_{r2}$ separately as well as algorithm 1, the global system $\boldsymbol{\Sigma}_G$ is reduced and distributed. Here, a one-sided method is chosen, thus the controllable subspace is determined only by using the transformation matrix $\mathbf{V}$ [37]. For reduction, $\mathbf{W}^{*} = \mathbf{V}$ is chosen so that $\mathbf{W}^{*}$ has maximum rank. Thus, a reduced order model is found,

$$\boldsymbol{\Sigma}_{r,C} = \begin{cases}\dot{\mathbf{x}}_{r,C}(t) = \mathbf{W}^{*^{T}}\mathbf{A}\mathbf{V}\mathbf{x}_{r,C}(t) + \mathbf{W}^{*^{T}}\mathbf{B}\mathbf{u}(t)\\ y_{r,C}(t) = \mathbf{C}\mathbf{V}\mathbf{x}_{r,C}(t)\end{cases} \tag{46}$$

which describes the controllable subspace. The application of algorithm 1 is repeated for the *Krylov* output space. Thus the matrix $\mathbf{W}$ is determined. For this reduction $\mathbf{V}^* = \mathbf{W}$ is chosen, thus also here the second matrix has maximal rank. The reduced model

$$\mathbf{\Sigma}_{r,O} = \begin{cases} \dot{\mathbf{x}}_{r,O}(t) = & \mathbf{W}^T\mathbf{A}\mathbf{V}^*\mathbf{x}_{r,O}(t) + \mathbf{W}^T\mathbf{B}\mathbf{u}(t) \\ y_{r,O}(t) = & \mathbf{C}\mathbf{V}^*\mathbf{x}_{r,O}(t) \end{cases} \tag{47}$$

describes the observable subspace. Using (46) and (47), the global system is approximated over the controllable and observable subspace. A reduced system is obtained by the system matrix $\mathbf{A}_{r,C}$ and the input matrix $\mathbf{B}_{r,C}$, of the controllable subspace, and by the output matrix $\mathbf{C}_{r,O}$, of the observable subspace

$$\begin{aligned} \mathbf{A}_{r,C} &= \mathbf{W}^{*^T}\mathbf{A}\mathbf{V} \\ \mathbf{B}_{r,C} &= \mathbf{W}^{*^T}\mathbf{B} \\ \mathbf{C}_{r,O} &= \mathbf{C}\mathbf{V}^* \end{aligned} \tag{48}$$

4.2.1 Application in a Sensor and Actuator Network

A KF is also used for state reconstruction in the subspace-based distribution. After the transformation from continuous to discrete time, the system is used in KF. By distributing into the controllable and observable subspace, this method can also be used in a network. Appling the SKF presented in Section 4.1, the computational effort in the decentralized CPUs is reduced because each network node only needs to determine a part of the KF algorithm. By using the reduced model (48), the SKF becomes the SRKF as described above (Fig. 5). However, information from both subspaces is necessary for a complete approximation. By communicating a priori estimates $\hat{\mathbf{x}}_r^-$ and covariances matrices $\mathbf{P}_r^-$ as well as a posteriori state vectors $\hat{\mathbf{x}}_r^+$ and covariance matrices $\mathbf{P}_r^+$, a interconnected network is created. To achieve this objective, the orthonormal bases from Section 3 are used. For communication between the reduced controllable and reduced observable subspace, the transmitted data must be transformed into the target subspace.

$$\mathbf{T}_c^o = \mathbf{W}\mathbf{V}^\dagger. \tag{49}$$

The index of $\mathbf{T}$ describes the starting point and the superscript indicates the target point of transformation. Due to the lack of symmetry of the orthonormal bases, the *Moore-Penrose* inverse is here also necessary. As in the previous section, we assume that no packets are lost and that both nodes operate with the same sampling time.

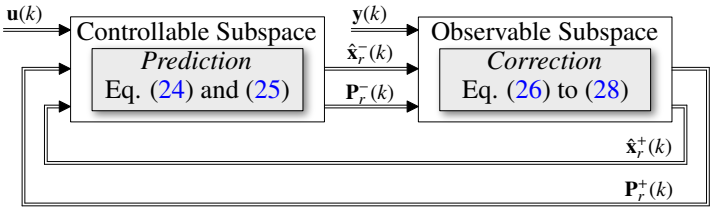


Figure 5: Distribution of the global linear system on the basis of controllability and observability

5 Experimental Evaluation and Results

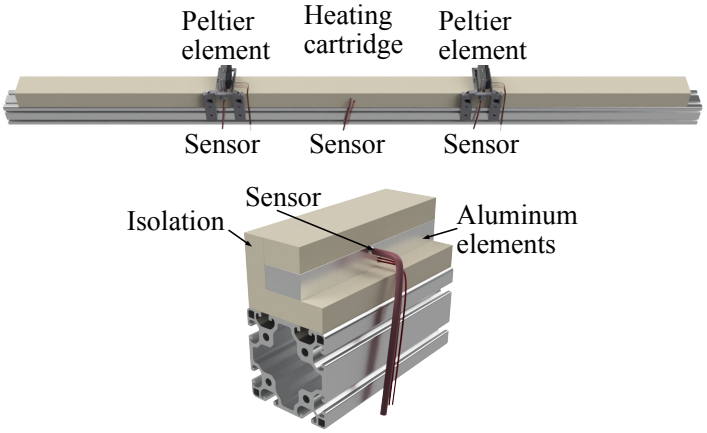


Figure 6: Visualization of a large scaled system. The discrete aluminum rod (l = 2m) is coold on two elements and heated on another.

We have experimentally evaluated the presented methods on an example published in [1] and [2]. Here we analyzed the temperature distribution in a discrete heating rod.
The aluminum rod consists of $n = 100$ elements. The contact surfaces of the aluminum elements have been thermally insulated, minimizing the influence of ambient temperature. The heat conducting rod is cooled by two Peltier elements and heated by a heating cartridge. In addition, the temperature distribution in the aluminum elements is monitored by three PT1000 sensors. The global system of the thermal rod can be described as follows

$$\begin{aligned} &\mathbf{x}_q(k+1) = p_1\mathbf{x}_{q-1}(k) + (2p_2 + p_3)\,\mathbf{x}_q(k) + p_1\mathbf{x}_{q+1}(k) + \mathbf{w}_q \\ &\quad \text{with } q = 1, \ldots, n \end{aligned} \tag{50}$$

Where the index q is the number of the aluminum element. The parameters p_1 and p_2 represents the element geometry, heat capacity as well as the heat transfer coefficient. The elements $q = 1$ and $q = 100$ have only one neighboring element, thus the heat emitting surface is larger than for the other elements. For this purpose, $2p_2 + p_3 \Longrightarrow p_2 + p_4$ is chosen for the elements at the end of the rod. The parameter p_3 and p_4 are the product of the heat transfer coefficient and the heat emitting surface. $\mathbf{w}_q$ is a normally distributed noise, describing uncertainties in the process.
The system order is $n = 100$ and is determined by the elements of the rod. Thus the system matrix is $\mathbf{A} \in \mathbb{R}^{100\times100}$. Here, the system has three actuators and three sensors, so the input matrix is $\mathbf{B} \in \mathbb{R}^{n\times3}$ and the output matrix is $\mathbf{C} \in \mathbb{R}^{3\times n}$. The sensors and actuators are placed in the same aluminum element (25, 50, 75). So the input matrix $\mathbf{B}$ is a zero matrix and is only assigned at the positions $B_{25,1} = -1$, $B_{50,2} = 1$ and $B_{75,3} = -1$ and the output matrix $\mathbf{C}$ is also a zero matrix and is also only assigned at the elements $C_{1,25} = 1$, $C_{2,50} = 1$ and $C_{3,75} = 1$. The global system can be described as follows

$$\mathbf{\Sigma}_{Rod} = \begin{cases} \mathbf{x}(k+1) & = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \\ \mathbf{y}(k) & = \mathbf{C}\mathbf{x}(k) \end{cases} \tag{51}$$

5.1 Parameter Identification

The unknown parameters of the global system are determined by identification in two steps. In the first step, the parameter p_2 is identified by the homogeneous solution of the state space equation. For this purpose, a single aluminum element is used to determine the main diagonal of the system matrix $\mathbf{A}$ (Fig. 7).

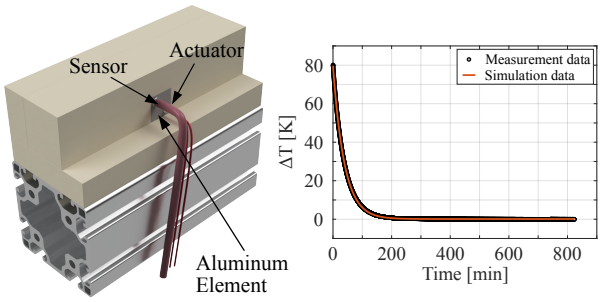


Figure 7: Parameter identification of the main diagonal elements.

This element is equipped with the heating cartridge and a sensor, and the isolation minimizes the influence of the ambient temperature. The heating cartridge will heat the temperature in the element to T=80°C. When the temperature reaches T=80°C, the power of the heating cartridge is switched off and the cooling curve is recorded until the steady state is reached (Fig. 7).
For the identification procedure, a first-order linear model and a *Nelder-Meat* simplex are used in combination with an ode4 solver. The least squares method is used to evaluate the output error and the *Nelder-Meat* simplex minimizes the cost function

$$\min J = \min_{\hat{\mathbf{p}} \in \mathbb{R}^n} \sum_{k=1}^{N} \left[y(k) - \hat{y}(k) \right]^2 . \tag{52}$$

so that the behavior of the model is fitted to the real system.
In the second identification step, the secondary diagonal elements of the system matrix $\mathbf{A}$ are determined. For this purpose, seven aluminum elements are used to reduce the discretization error (Fig. 8). The elements at the edge are equipped with a sensor. The middle element is also equipped together with the heating cartridge. With constant heating power, the elements are heated until the steady state is reached. A seventh order model is used for identification and least squares is used to evaluate the output error. The cost function (52) is also minimized using a *Nelder-Meat* simplex in combination with an ode4 solver.

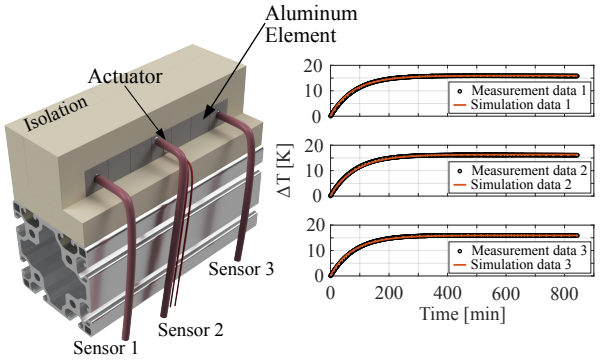


Figure 8: Parameter identification of the secondary diagonal elements.

These identified parameters can be used in the global system

$$\begin{aligned} p_1 &= 4.2812 \cdot 10^{-4} \tfrac{1}{s} \quad p_2 = 4.3048 \cdot 10^{-4} \tfrac{1}{s} \\ p_3 &= 1.8067 \cdot 10^{-4} \tfrac{1}{s} \quad p_4 = 2.2561 \cdot 10^{-4} \tfrac{1}{s} \end{aligned} \tag{53}$$

5.2 Input-/Output-based Distribution

The starting point is a global MIMO system (51) with three actuators and three sensors and a system order of $n = 100$. Considering the controllability and observability of the global system $\mathbf{\Sigma}_{Rod}$, $m = 3$ local systems are decentralized with one actuator and sensor each.

$$\begin{aligned} \mathbf{\Sigma}_{L_1} &\begin{cases} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{B}_j u_1(k) \\ y_1(k) &= \mathbf{C}_j \mathbf{x}(k) \end{cases} \\ \mathbf{\Sigma}_{L_2} &\begin{cases} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{B}_j u_2(k) \\ y_2(k) &= \mathbf{C}_j \mathbf{x}(k) \end{cases} \\ \mathbf{\Sigma}_{L_3} &\begin{cases} \mathbf{x}(k+1) &= \mathbf{A}\mathbf{x}(k) + \mathbf{B}_j u_3(k) \\ y_3(k) &= \mathbf{C}_j \mathbf{x}(k) \end{cases} \end{aligned} \tag{54}$$

The index $j = 1, \ldots, m$ is the corresponding column of the global input matrix $\mathbf{B}$ and the corresponding row of the global output matrix $\mathbf{C}$. The system order is not changed, thus the decentralized local systems are also $n = 100$.
Analyzing the HSVs reveals, that order $r = 6$ is achieved (Fig. 9). The order $r = 6$ allows 93.1217% of the energy to be transported from the inputs to the outputs. Based on the analysis, states with a small impact on the energy transport are cut off and the system order is reduced to $r = 6$. The model order reduction reduces the computational effort in the network node.
The decentralized reduced local systems are extended by the DSRKF presented in Section 4, providing communication between the network nodes and generating a fully interconnected network.

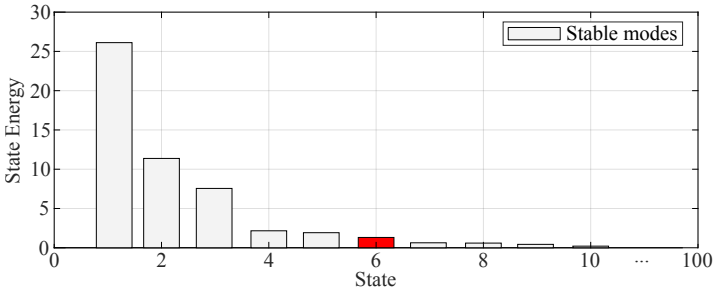


Figure 9: Visualization of the state energy of the global system (50). It is illustrated that the energy decreases strongly with the sixth state. Red bar illustrates the cut-off system order r.

Fig. 3 illustrates that the sensor nodes communicate with the actuator nodes by means of the exchanged a priori and a posteriori information. The full interconnectivity of the network is achieved by exchanging the error information (31) and (32) between the decentralized reduced local systems. This error information is already used in the prediction, so a fusion step has to be added to the algorithm in the actuator nodes (36) and (37).
By using BT, the physical interpretability of the states is lost, thus, for an analysis, the reduced state vector has to be transformed back

to the global state space. For this transformation of the a priori and a posteriori information, the following equations are used

$$\hat{\mathbf{x}}^{-}(k) = \mathbf{V}_{BT}\hat{\mathbf{x}}_r^{-}(k) \quad (55)$$

$$\hat{\mathbf{x}}^{+}(k) = \mathbf{V}_{BT}\hat{\mathbf{x}}_r^{+}(k). \quad (56)$$

The transformation matrix $\mathbf{V}_{BT}$ is determined by (15).

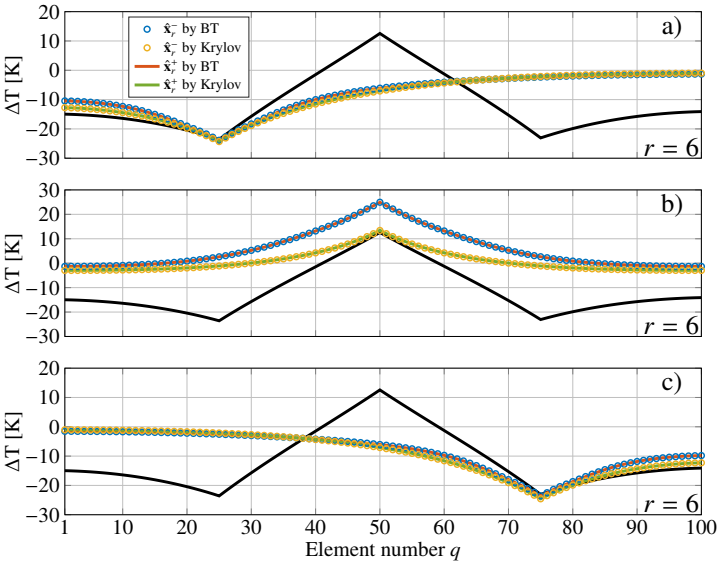


Figure 10: A priori and a posteriori estimated state vectors of local systems reduced by BT (3.1) and *Krylov* (3.2).

Fig. 10 demonstrates that the local SISO systems satisfy the characteristics of the decentralized inputs and outputs of the global system. Hereby, the global temperature profile (50) is shown in black and the a priori estimation $\hat{\mathbf{x}}_r^-$ after the reduction by BT/Krylov of the acutator node, is represented in blue/orange dots as well as the a posteriori estimation $\hat{\mathbf{x}}_r^+$ in red/green lines of the sensor node. In a), b) and c) the estimates based on measurement data of the reduced local systems after the transformation into the original SISO subspace ($n = 100$) are illustrated.

In addition, it is presented that despite the lost interpretability of the states due to the model order reduction as well as the distribution of the estimation algorithms in a interconnected network, the controllability and observability of the global system is approximated by the decentralized reduced local systems.

The input-/output-based distribution from Section 4.1 is also analyzed using the model order reduction method from Section 3.2, the two-sided *Krylov* method. Although the model order reduction with *Krylov* subspace methods is numerically efficient, the stability of the reduced system is not guaranteed. Therefore, the stability must be checked in the preprocessing. Based on the stability analysis of the decentralized local systems, the reduction with *Krylov* subspace methods also yields stable reduced systems for $r = 6$.

This approach also illustrates that the characteristics of the sensors and actuators of the global system is approximated by the decentralized reduced local systems (Fig. 10). Furthermore, the described method for decentralization is used independently of the two model order reduction methods presented.

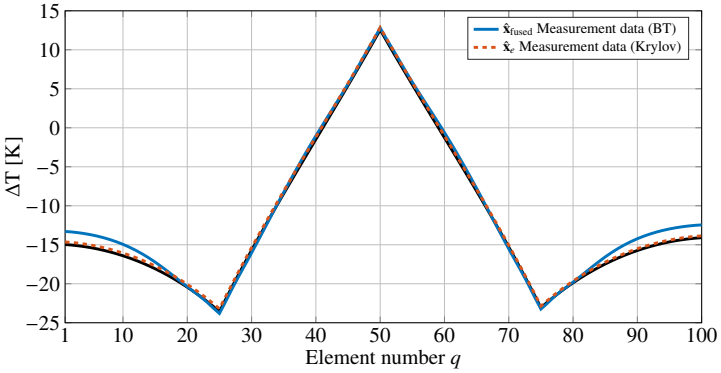


Figure 11: By BT and *Krylov* reduced estimated a posteriori state vectors with (57) and (58) as well as (43) and (44) fused in a center.

If the a posteriori information of the decentralized SISO system reduced by BT is provided to a fusion center, the original $\mathbf{\Sigma}_{Rod}$ is reconstructed. For this purpose, the reduced a posteriori state vector $\hat{\mathbf{x}}_r^+$ and the a posteriori covariance matrix $\mathbf{P}_r^+$ of the $m = 3$ local systems must be transformed into the local model space and fused by

$$\mathbf{P}_{\text{fused}}(k) = \left(\sum_{j=1}^{m}\mathbf{P}_j^{-}(k)\right)^{-1} \quad (57)$$

$$\hat{\mathbf{x}}_{\text{fused}}(k) = \mathbf{P}_{\text{fused}}(k)\sum_{j=1}^{m}\left(\mathbf{P}_j^{-}(k)\right)^{-1}\hat{\mathbf{x}}_j^{-}(k). \quad (58)$$

For the reconstruction of the global system $\mathbf{\Sigma}_{Rod}$ based on the second method of model order reduction, a reference system (45) is derived. The stability and the equivalent system order ($r = 6$) as for the $m = 3$ decentralized reduced local systems have to be considered. Before the decentrally reduced local a posteriori information is used for fusion, a transformation into a known common subspace is required. For this purpose, the decentrally reduced local a posteriori information is transformed into the subspace of the reference system. For further improvement a KF is used , which achieves an increase of the reconstruction $\hat{\mathbf{x}}_e$ of the global system $\mathbf{\Sigma}_{Rod}$ with the reference system and the fusion result as observation.

In contrast to Fig. 10, a fusion center is used in Fig. 11. The global temperature profile according to (50) with the identified parameters is shown in black. The decentralized local a posteriori states, reduced by BT, are shown in blue after fusion by (57) and (58) and transformation ($n = 100$), while the dashed red line represents the decentralized local a posteriori states, reduced by the method of *Krylov* subspaces.

Here, the global system is fully reconstructed despite decentralization and distribution in a sensor and actuator network. Additionally, it is demonstrated that the presented methods are independent of the model order reduction techniques. The decentralization and distribution would be implemented for both model order reduction methods (BT and *Krylov* subspace methods) on BECKHOFF® CX6030 CPUs (Intel® Core™ i7 7700, 3.6GHz) and tested in real-time approaches.

The deviation of the approximation by Balanced Truncation in the ranges $q = 1, \ldots, 25$ and $q = 75, \ldots, 100$ is accounted for by the reduction of the system order. With an increase of the order this

deviation is minimized.
The communication effort is directly dependent on the system order. For clarification, the system (51) was manually distributed into several local systems. The order of the local systems is $n_1 = 35$, $n_2 = 20$ and $n_3 = 45$. If these decentralized local systems are used to reconstruct the global system without further reduction, all local systems must communicate with each other. The communication overhead is significantly higher than in the presented methods with the reduction methods BT and Krylov subspace methods (Fig. 12). Thus, the presented methods reduce not only the computational effort in the network nodes but also the communication effort.

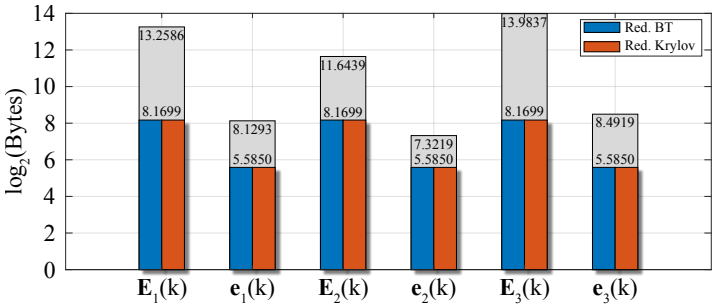


Figure 12: The information of the covariance error and state error sent by manual splitting is illustrated as grey and for DSRKF as blue/red bars (logarithmic scale).

5.3 Subspace based Distribution

Without decentralization, the controllable and observable *Krylov* subspace is derived directly from the global system (51). For this purpose, the repeating of the one-sided method for the *Krylov* input space and the *Krylov* output space is proposed in Secton 4.2. The resulting reduced systems (46) and (47) describe the corresponding subspace. However, to approximate the global input and output behavior, information from both subspaces are required. For state reconstruction in the subspace-based distribution, a SRKF from Section 4.1.1 is used with the matrices (48). By separating the controllable and observable subspace, this distribution can also be used in a network.

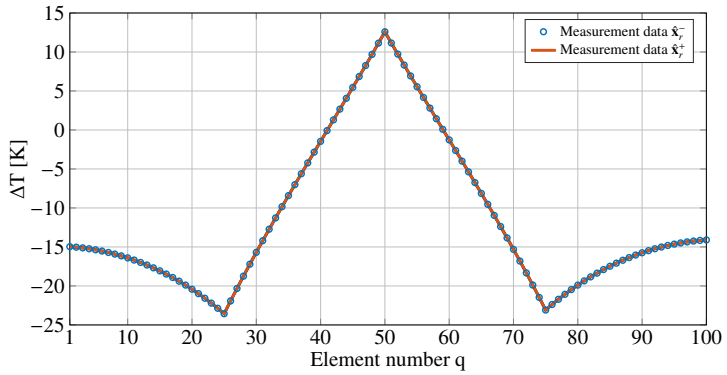


Figure 13: Reconstructing the global temperature profile by communicating the reduced a priori and a posteriori states according to the distribution of the controllability and observability subspace in a sensor and actuator network.

The network nodes transmit the a priori and a posteriori state vectors and covariance matrix, so no fusion steps are required in the algorithms. However, the information received in the subspaces must be transformed (49). Even when *Krylov* methods are used without decentralization, the stability of the reduced system must be analyzed in preprocessing. Fig. 13 demonstrates that also in this case, for $r = 6$, a stable reduced system is present and used.
In contrast to the previous method (Section 5.2), only two network nodes are used here. One network node uses the complete information of the inputs for the prediction of the KF and another network node uses all observations for the correction step of the reconstruction of the global system behaviour.
The minimization of the communication effort is not depicted, as in this distribution only the prediction and correction is separated (SKF). Using the model order reduction (SRKF), there will be a lower communication effort with each reduced system order, since a manual division cannot be performed when separating prediction and correction.

6 Conclusion

In this paper, we have shown the decentralization into SISO systems of large-scale MIMO systems. We have also presented an input/output-based and a subspace-based method for system distribution. Both methods focus on the controllability and observability of the global MIMO system. By applying the two best known model order reduction methods (balanced truncation and *Krylov* subspaces), the system order of the generated decentralized local systems was reduced. By adding a DSRKF to the decentralized reduced local systems, they are used in a network. The reduced decentralized local systems describe different subspaces, so it is assumed that the reduced information is communicated in the network. It could be demonstrated that the transformation matrices resulting from the two presented methods can also be used for communication. Their distribution in a connected network in combination with the model order reduction methods significantly reduced the computational and communication effort required for state estimation of the large-scale linear MIMO system.
In the future, the presented methods will also be applied for decentralized distributed control. In future practical approaches, the methods will also be used in tracking dynamic targets and applied to the decentralized distributed multi-laser tracker system (DDMLTS) [38].

Conflict of Interest The authors declare no conflict of interest.

References

[1] F. Friedrich, J. Mayer, C. Ament, "Reduced and Distributed Estimation in Sensor and Actuator Networks - Automated Design Based on Controllability and Observability," in 2021 IEEE Conference on Control Technology and Applications (CCTA), 666–672, IEEE, San Diego, CA, USA, 2021, doi: 10.1109/CCTA48906.2021.9659129.

[2] F. Friedrich, J. Mayer, C. Ament, "Model-Order Reduction and System Distribution Using Krylov Subspaces - An Approach for Efficient State Estimation in Sensor and Actuator Networks," accepted for CCTA2022.

[3] R. D'Andrea, G. Dullerud, "Distributed control design for spatially interconnected systems," IEEE Transactions on Automatic Control, **48**(9), 1478–1495, 2003, doi:10.1109/TAC.2003.816954.

[4] B. Bamieh, F. Paganini, M. Dahleh, "Distributed control of spatially invariant systems," IEEE Transactions on Automatic Control, **47**(7), 1091–1107, 2002, doi:10.1109/TAC.2002.800646.

[5] L. A. Montestruque, P. J. Antsaklis, "Model-Based Networked Control Systems - Stability," 58.

[6] J. Lunze, Regelungstechnik 2, Springer Berlin Heidelberg, Berlin, Heidelberg, 2016, doi:10.1007/978-3-662-52676-7.

[7] U. A. Khan, J. M. F. Moura, "Model Distribution for Distributed Kalman Filters: A Graph Theoretic Approach," in 2007 Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers, 611–615, IEEE, Pacific Grove, CA, USA, 2007, doi:10.1109/ACSSC.2007.4487286, iSSN: 1058-6393.

[8] B. Noack, J. Sijs, U. D. Hanebeck, "Fusion Strategies for Unequal State Vectors in Distributed Kalman Filtering," IFAC Proceedings Volumes, **47**(3), 3262–3267, 2014, doi:10.3182/20140824-6-ZA-1003.02491.

[9] J. Lunze, Control Theory of Digitally Networked Dynamic Systems, Springer International Publishing, Heidelberg, 2014, doi:10.1007/978-3-319-01131-8.

[10] A. C. Antoulas, Approximation of large-scale dynamical systems, Advances in design and control, Society for Industrial and Applied Mathematics, Philadelphia, 2005, doi:10.1137/978-0-89871-871-3.

[11] P. Benner, A. Schneider, "Balanced Truncation Model Order Reduction for LTI Systems with many Inputs or Outputs," in Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems – MTNS 2010, 4, MTNS, Budapest, Hungary, 2010.

[12] E. F. Yetkin, H. Dag, "Parallel implementation of iterative rational Krylov methods for model order reduction," in 2009 Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control, 1–4, IEEE, Famagusta, 2009, doi:10.1109/ICSCCW.2009.5379421.

[13] Z. Bai, "Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems," Applied Numerical Mathematics, **43**(1-2), 9–44, 2002, doi:10.1016/S0168-9274(02)00116-2.

[14] P. Feldmann, R. Freund, "Efficient linear circuit analysis by Pade approximation via the Lanczos process," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, **14**(5), 639–649, 1995, doi:10.1109/43.384428.

[15] K. Gallivan, E. Grimme, P. Van Dooren, "Asymptotic waveform evaluation via a Lanczos method," Applied Mathematics Letters, **7**(5), 75–80, 1994, doi:10.1016/0893-9659(94)90077-9.

[16] G. A. Baker, P. R. Graves-Morris, Padé approximants, number v. 59 in Encyclopedia of mathematics and its applications, Cambridge University Press, Cambridge [England] ; New York, 2nd edition, 1996.

[17] D. Rafiq, M. A. Bazaz, "Model Order Reduction via Moment-Matching: A State of the Art Review," Archives of Computational Methods in Engineering, **29**(3), 1463–1483, 2022, doi:10.1007/s11831-021-09618-2.

[18] L. Feng, "Review of model order reduction methods for numerical simulation of nonlinear circuits," Applied Mathematics and Computation, **167**(1), 576–591, 2005, doi:10.1016/j.amc.2003.10.066.

[19] H. K. F. Panzer, T. Wolf, B. Lohmann, "$\mathcal{H}_2$ and $\mathcal{H}_\infty$ error bounds for model order reduction of second order systems by Krylov subspace methods," in 2013 European Control Conference (ECC), 4484–4489, IEEE, Zurich, 2013, doi:10.23919/ECC.2013.6669657.

[20] S. Gugercin, Projection Methods for Model Reduction of Large-Scale Dynamical Systems, Dissertation, Rice University, Rice, 2003.

[21] B. Salimbahrami, B. Lohmann, T. Bechtold, "Two-Sided Arnoldi in Order Reduction of Large Scale MIMO Systems," 9.

[22] B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," IEEE Transactions on Automatic Control, **26**(1), 17–32, 1981, doi:10.1109/TAC.1981.1102568.

[23] S. Gugercin, A. C. Antoulas, "A Survey of Model Reduction by Balanced Truncation and Some New Results," International Journal of Control, **77**(8), 748–766, 2004, doi:10.1080/00207170410001713448.

[24] D. Hinrichsen, H.-W. Philippsen, "Modellreduktion mit Hilfe balancierter Realisierungen / Model reduction using balanced realizations," at - Automatisierungstechnik, **38**(1-12), 460–466, 1990, doi:10.1524/auto.1990.38.112.460, publisher: De Gruyter (O).

[25] S. L. Brunton, J. N. Kutz, Data-Driven Science and Engineering, Cambridge University Press, 1st edition, 2019.

[26] A. Laub, M. Heath, C. Paige, R. Ward, "Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms," IEEE Transactions on Automatic Control, **32**(2), 115–122, 1987, doi:10.1109/TAC.1987.1104549, conference Name: IEEE Transactions on Automatic Control.

[27] I. Jaimoukha, "A general minimal residual Krylov subspace method for large-scale model reduction," IEEE Transactions on Automatic Control, **42**(10), 1422–1427, 1997, doi:10.1109/9.633831.

[28] B. Lohmann, B. Salimbahrami, "Introduction to Krylov subspace methods in model order reduction," Methods Applications in Automation, 1–13, 2003.

[29] P. Falb, W. Wolovich, "Decoupling in the design and synthesis of multivariable control systems," IEEE Transactions on Automatic Control, **12**(6), 651–659, 1967, doi:10.1109/TAC.1967.1098737.

[30] D. Simon, Optimal state estimation: Kalman, H [infinity] and nonlinear approaches, Wiley-Interscience, Hoboken, N.J, 2006, doi:10.1002/0470045345.

[31] A. G. Mutambara, Decentralized Estimation and Control for Multisensor Systems, Routledge, 1st edition, 2019, doi:10.1201/9781315140803.

[32] P. Hilgers, C. Ament, "Distributed and decentralised estimation of non-linear systems," in 2010 IEEE International Conference on Control Applications, 328–333, 2010, doi:10.1109/CCA.2010.5611300, iSSN: 1085-1992.

[33] B. S. Rao, H. F. Durrant-Whyte, "Fully decentralised algorithm for multisensor Kalman filtering," IEE Proceedings D (Control Theory and Applications), **138**(5), 413–420, 1991, doi:10.1049/ip-d.1991.0057, publisher: IET Digital Library.

[34] V. Shin, Y. Lee, T.-S. Choi, "Generalized Millman's formula and its application for estimation problems," Signal Processing, **86**(2), 257–266, 2006, doi:10.1016/j.sigpro.2005.05.015.

[35] H. B. Mitchell, Multi-sensor data fusion: an introduction, Springer, Berlin ; New York, 2007, doi:0.1007/978-3-540-71559-7.

[36] J. Ajgl, "Millman's Formula in Data Fusion," in The 10th International PhD Workshop Young Generation Viewpoint, 6.

[37] B. Salimbahrami, R. Eid, B. Lohmann, "On the choice of an optimal interpolation point in Krylov-based order reduction," in 2008 47th IEEE Conference on Decision and Control, 4209–4214, IEEE, Cancun, 2008, doi:10.1109/CDC.2008.4739219.

[38] F. Friedrich, J. Brandl, C. Ament, "Absolute Distance Measurement by a Decentralized and Distributed Multi-Lasertracker-System," in 2022 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), 1249–1255, IEEE, Sapporo, Japan, 2022, doi:10.1109/AIM52237.2022.9863275.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 151-160 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Process Mining in Healthcare: A Systematic Literature Review and A Case Study

Fabrizio Striani[*,1], Chiara Colucci[2], Angelo Corallo[1], Roberto Paiano[1], Claudio Pascarelli[1]

[1]*University of Salento, Department of Engineering for Innovation, Via per Monteroni, Lecce, 73100, Italy*

[2]*National Interuniversity Consortium for Informatics, Via Ariosto 25, Roma, 00185, Italy*

ARTICLE INFO

Article history:
Received: 27 July, 2022
Accepted: 26 October, 2022
Online: 08 December, 2022

Keywords:
Process Mining
Healthcare
Business Process Modeling

ABSTRACT

Process mining is an innovative technique through which inefficiencies in production systems can be eliminated. This technique has therefore become very important internationally in recent years and is also useful for pursuing the improvement of production systems by extrapolating process knowledge from event logs recorded by information systems. Process mining has easy application in production systems as current business processes are integrated with information systems and this makes data available immediately. This makes the complex nature of industrial operations understandable and, for this reason, process mining could also be used in the healthcare field where cost containment and the quality of service increasing offered to the community has become paramount. The problem is that in this sector, it is much more complex to identify useful data in order to extrapolate the relevant log file. The article aim is to examine the state of the art of the application of process mining in the healthcare sector in order to understand the level of diffusion of these techniques. In the light of this analysis, a case study will then be analysed on the application of process mining techniques in the healthcare sector, and in particular in medical teleconsultation in the field of neuroradiology.

1. Introduction

This paper is an extension of the work originally presented in "Application of Process Mining in Teleconsultation Healthcare: Case study of Puglia Hospital" [1].

Process mining is a recently applied research discipline which combines data mining and computational intelligence on the one hand with process analysis and modelling on the other [2], [3].

Despite the relative youth of this discipline, many techniques and methodologies have been implemented. The most widely used are: Fuzzy and Heuristic Miner. The first algorithm allows the process model to be made simpler, as this technique reduces the model to a desired abstraction threshold and is, therefore, particularly suitable for the extraction of less structured processes with a large amount of conflicting behaviour [4]. The second uses frequencies and parameterisation so that the main behaviour can be recorded in an event log. Considering the dependency measure, it takes into account frequencies and causal dependencies, making it suitable for many real logs [5].

In particular, process mining involves the discovery of processes, the monitoring of compliance through the analysis of deviations (which can be either negative or positive) from best practices and the optimisation of the model based on the deviations found.

Indeed, process mining was born with the aim of extracting knowledge from current information systems (using event logs that can be easily accessed), and through this, to be able to discover, monitor and improve real processes [6]. This technique consists of three phases:

1) discovery (process models are extracted from an event log);
2) conformance checking (or check of any deviation between the model and the log by making a comparison);
3) enhancement (using the information in the logs, improves or extends a process model) [7].

Consequently, the application of these techniques is particularly useful in industrial scenarios where there is the production of complex goods, and, therefore, the analysis of a large amount of information becomes necessary for the development management of new applications.

[*]Corresponding Author: Fabrizio Striani, fabrizio.striani@unisalento.it

Another natural application field of process mining analysis is healthcare [8], [9]. Several case studies exist in which process mining techniques and tools are applied in this sector [10], [11]. Among these, the oncology and the hospitalisation sector show the maximum number of successful cases [12]. Hospitals, for example, need to focus on managing their processes to have high quality services while reducing costs, especially when patient demand unexpectedly increases (as was the case with the COVID-19 pandemic crisis) [13]-[15]. The biggest problem in the healthcare sector is that the patient care process ("care flow") very often differs even if the patients have a similar diagnosis. The techniques under consideration offer an interesting possibility through which this problem can be solved: process mining has already been used to a very significant extent in service companies and could therefore also be a good solution in the health sector [16], [17].

Another problem is the existence of several systems which store events and data. An intensive care unit, for example, is very often equipped with its own system to store patients' treatments or examinations, while the radiology department has another system to store the entire patient process. It becomes, therefore, crucial that all information is amalgamated in a single source. However, given the great multiplicity of systems involved, which makes data collection complex, process mining techniques could prove very useful as they would allow, starting from the individual events carried out in each division, to obtain the entire process, which could provide an overview that, for example, could ensure good reporting of the examinations performed [18], [19].

In this paper, the first two sections we introduce, explain and give a systematic literature review which evaluates the process mining tools applications in the healthcare scenario. According to its definition, a systematic literature review is "A review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyse data from the studies that are included in the review. Statistical methods (meta-analysis) may or may not be used to analyse and summarise the results of the included studies" [20].

With this systematic search process, we identify studies addressing a specific research question and we make a systematic presentation and synthesis of the characteristics and outcomes of the research findings. The criteria by which studies are excluded or included are objective, consistently applied and explicitly stated, so that when certain studies are included or excluded from the review, this decision is clear to the readers; or if another researcher decides to use the same criteria, he or she is likely to replicate the same decisions [21].

In the second section we present the Tranfield approach that is used to conduct the systematic literature review, allowing the identification of the most interesting documents in the analysed context. In the third section we show up our case study: an analysis of the log file extracted from the Apulian hospitals' repository. In section 4, we illustrate the use of process mining techniques on the log file (using Disco software [22]) and analyse the results achieved with these techniques. Then, in this section, the AS-IS process is represented with BPMN notation (using the Signavio tool). In the last section, the results of our case study are compared with data related to the Piedmont Region (representing a virtuous Italian region), which are taken from the official source of the Ministry of Health. At this point, we modelled a TO-BE process to optimise the current process.

Finally, there are our conclusions and the references used. Ultimately, our analysis aim is not to investigate new issues but, based on the state of the art, to highlight that Process Mining techniques are useful in healthcare applications. From the above, we can extrapolate the following research question:

"**RQ** - Provided the state of the art obtained from the SLR, and aimed at identifying the scope of application in healthcare, does the use of teleconsultation optimise the performance of the healthcare service in the field of neuroradiology?"

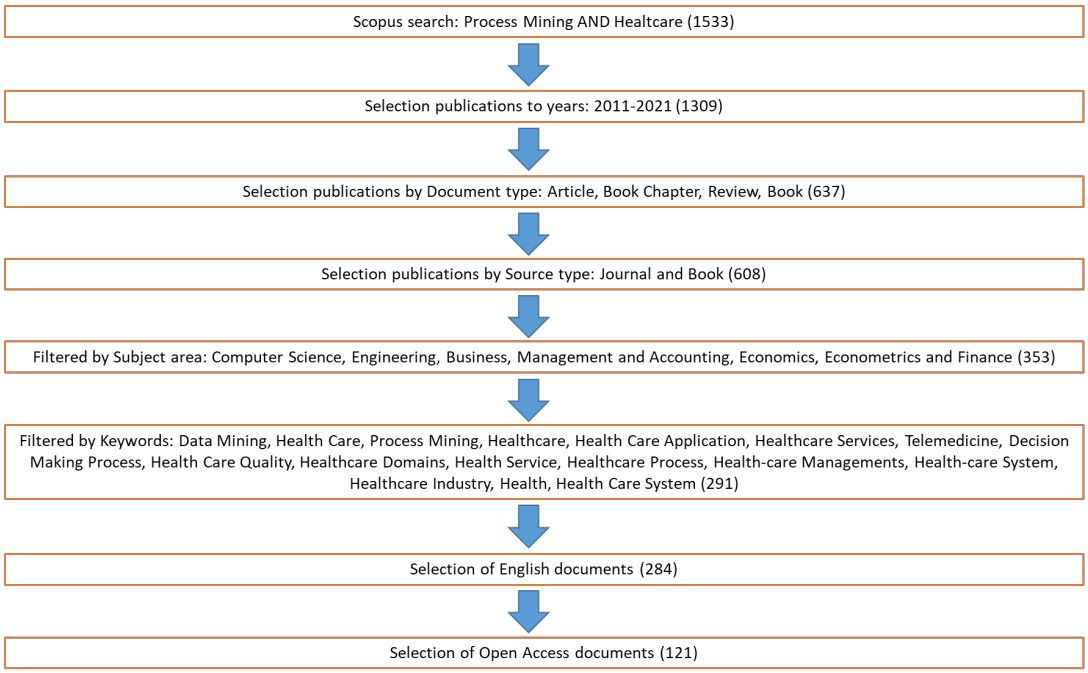


Figure 1: The process of systematic literature review

2. Research Method

Since this is a review article, we will now apply the Tranfield approach to conduct a systematic literature review (SLR), finalised to highlight the usefulness of Process Mining techniques in healthcare applications.

In clinical disciplines, the method theorised by [23] is used to implement a systematic review (SR). We apply the Tranfield approach to develop the "Study Selection" phase of Egger's method, a phase through which the various studies identified are evaluated, so that it is possible to assess whether or not the inclusion criteria are met. The Tranfield approach uses three phases: planning, execution, reporting and dissemination.

In the planning phase, we enter the terms "Process mining" AND "Healthcare" to search for the keywords, abstract and title. Starting to execution, the initial Scopus search, (we are selected the Scopus database to guarantee the review's quality) on the basis of the selected words present in keywords, abstract and title, found a total of 1533 publications. Then, we are limited the analysis to the last 11 years (2011–2021), as our research interest is based on these years, resulting in 1309 publications [21].

By restricting the search to "Article", "Book chapter", "Review", and "Book", we found a total of 637 publications. Considering only the publications in Journals and Books, we identified 608 documents. However, to achieve the aims set initially, we selected only documents included in the fields of Computer Science, Engineering, Business, Management and Accounting, Economics, Econometrics and Finance finding a total of 353 publications.

The selected fields are the one of interest which refer to the case study analysed in the paper. In fact, as mentioned before in the RQ, the aim is to analyse the impact of teleconsultation from a technical point of view (hence the choice of the fields "Computer Science" and "Engineering"), and from a monetary point of view (hence the choice of the fields "Business, Management and Accounting", "Economics, Econometrics and Finance"). At this point, given the important number of articles found, and considering that the ones of greatest interest would be those incorporating the keywords, we applied a new filter (the keywords selected were: "Data Mining", "Health Care", "Process Mining", "Healthcare", "Health Care Application", "Healthcare Services", "Telemedicine", "Decision Making Process", "Health Care Quality", "Healthcare Domains", "Health Service", "Healthcare Process", "Health-care Managements", "Health-care System", "Healthcare Industry", "Health", "Health Care System"), obtaining 291 publications. After choosing the English language (284), and the Open Access documents, we found 121 documents. Figure 1 shows the SLR process. Finally, we faced the last phase of the Tranfield approach, namely reporting and dissemination.

As we show in Figure 2, the number of publications has increased in the last three years: 14 in 2019, 22 in 2020, 26 in 2021. In the previous years, there was an average of 6 publications, with a marked increase in 2018 (21 publications).

This trend highlights the growing attention to these topics in recent years. In addition, 65 journals published the topics analysed, 11 out of which published two or more articles (59), representing 49% of the total. Analysing the Scopus impact evaluation of sources with most publications, 78% were in the first quartile (which represents the highest quality of the publication magazine); 15% in the second; and only 7% in the third, demonstrating the high standard of the publications, and the level of interest in the subjects (Table 1 and Figure 3) [24], [25].

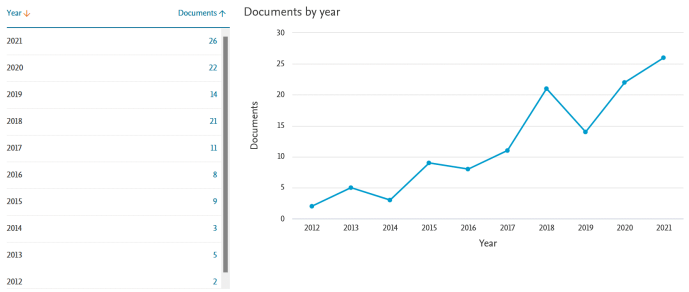


Figure 2 – Documents by year. Source: Scopus

Table 1: Journals with two or more documents published

Source	Documents	Ranking Scopus
"IEEE Access"	21	Q1
"Journal of Biomedical Informatics"	13	Q1
"IEEE Journal of Biomedical and Health Informatics"	4	Q1
"International Journal of Advanced Computer Science and Applications"	4	Q3
Sensors (Switzerland)	4	Q2
Applied Sciences (Switzerland)	3	Q2
Artificial Intelligence in Medicine	2	Q1
Computers Materials and Continua	2	Q1
Expert Systems with Applications	2	Q1
Plos Computational Biology	2	Q1
Sensors	2	Q2
Other	62	
Total	**121**	

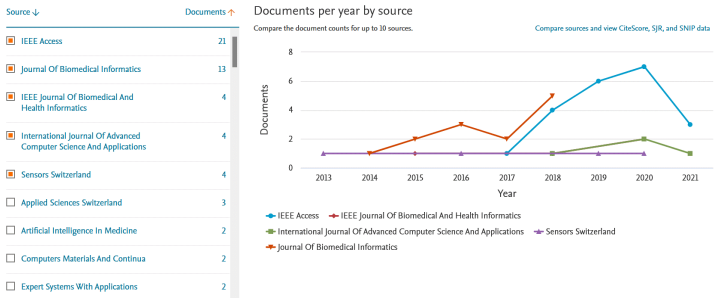


Figure 3: Documents per year by source. Source: Scopus

Moreover, we observed an increase in the number of publications for IEEE Access Journal (Q1) in the years 2018, 2019 and 2020, from 4 in 2018 to 7 in 2020, which again reflects the increasing interest in these themes in recent years.

The Affiliation Identifier allows us to identify and distinguish publications in relation to the reference institutions or organizations. In this case, we noted that the majority of publications comes from the Education area (University, Institute, Laboratory, School, College), followed by other areas: Company, Health Management, Research and Government (Figure 4)

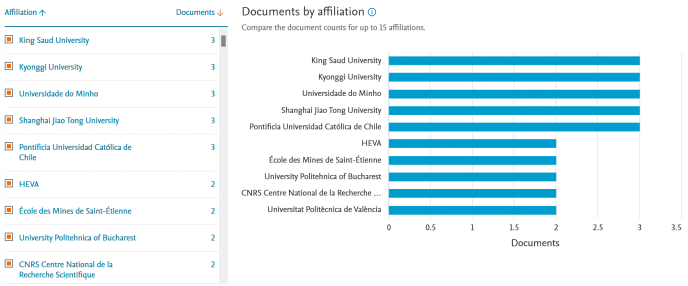


Figure 4: Documents by affiliation. Source: Scopus

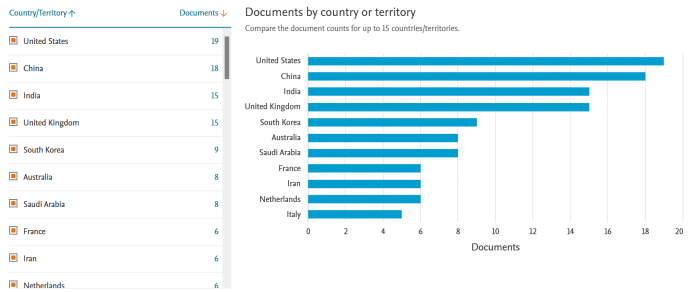


Figure 5: Documents by country or territory. Source: Scopus

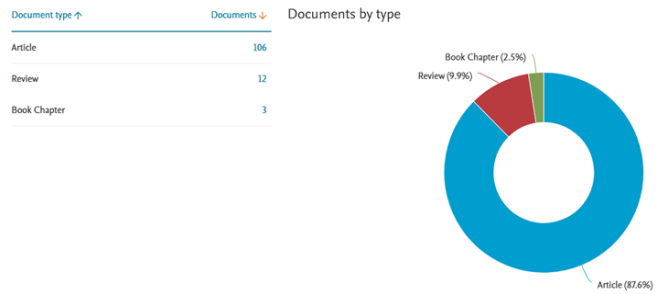


Figure 6: Documents by type. Source: Scopus

Another interesting aspect is the publications on the analysed topics by countries. In particular, most publications were produced in Asian territories (China and Middle East), followed by Europe (United Kingdom, France, Netherlands, Italy), and finally by America (principally United States). This may be further incentive to increase publications in our area (Figure 5).

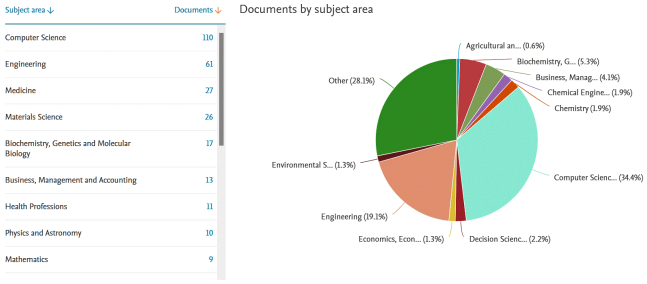


Figure 7: Documents by Subject area (Source: Scopus)

Moreover, as shown in Figure 6, publications consisted mainly of articles (106 – 87,6%), and some reviews (12 – 9,9%). In line with the selected Subject areas for the SLR, the majority of publications were in the Computer Science (110 – 34%) and Engineering (61 – 19%) areas. The areas "Business, Management and Accounting" and "Economics, Econometrics and Finance" counted 13 (4%), and 4 (1%) publications respectively (Figure 7). These findings shows that process mining applied to the healthcare sector has become very useful and relevant [26]. In particular, of the 27 papers about Medicine, the majority (17 – (63%)) concerned research activities in the field of healthcare using innovative approaches, such as: Health Surveillance Software Framework, Time Series Data Mining, Electronic Medical Record Search Engine, Intelligent Personal Health Record. Part of them regarded studies and applications of process mining in the field of neurological pathologies (2 – (7%)), cardiological pathologies (2 – (7%)), and in the radiology sector (2 – (7%)). In the remaining part of this sample, the authors discussed application cases on psychiatric diseases, diabetes and cancer. Since the review process we carried out have not delivered any paper on medical teleconsultation analysis in the neuroradiology field, in this paper we decided to analyse a case study on this subject. This case study is presented in the following section.

3. Case study

Following the considerations in the previous section, in this one we show the case study based on an analysis of process mining about clinical teleconsultation in the neuroradiology domain. We choose the teleconsultation as main topic, since it represents a new and innovative approach to healthcare delivery, also in the light of the pandemic period, therefore representing a very actual topic. Within the scope of this paper, we imagine the process as follows: after the patient is admitted to a medical district, the referring physician requires certain analysis and exams. Once all the necessary documents have been provided, the corresponding doctor sends all the documents (results of tests, images and diagnosis [if expressed]) to the specialist doctor for a second opinion. The specialised doctor returns his/her own diagnosis with the corresponding treatment or further

examinations. At this stage, the patient must undergo the suggested treatment until discharge.

The file log represented in the Table 2 derives from this repository. This log file represents the effective process, knowing that we have anonymised the patient and doctor data. For simplicity, we report only two cases and we add a cost column to the extracted data. The costs presented are the real ones for each activity with the exception of the activity "blood test". For the latter, in fact, we used a flat-rate cost, as a result of ten examinations usually carried out in the medical field of reference, at the cost of 10 euros on average. After acceptance, the patient is assigned to the referring doctor (the corresponding doctor). The physician, after visiting the patient and submitting him or her to the appropriate examinations, analyses the patient's medical history and, if he or she considers it appropriate, may request the specialist support for a second opinion. To request the specialist's support, the physician uses the teleconsultation tool and sends all the documents to the specialised doctor.

When the documentation received has been examined, the specialist can send the diagnosis to the applicant or request further examinations. After the diagnosis has been expressed, the patient must follow the suggested treatment before discharge (patient 1) or is directly discharged if no specific problems occur (patient 2). In next section we will use this log file to obtain the real process.

Table 2: Log file of process

Case id (patient)	Event identification	Timestamp	Activities	Resource	Price (€)
A	1	12/05/2020 h. 08:30	Patient acceptance	Reception	26,40
A	2	14/05/2020 h. 10:30	Referral to the requesting doctor	Support Service	26,51
A	3	17/05/2020 h. 09:30	Visit schedule	Support Service	26,50
A	4	08/06/2020 h. 09:30	Patient official visit	Doctor Silver (Corresponding doctor)	36,16
A	5	10/06/2020 h. 11:00	Angio exam	Doctor Coral	313,19
A	6	11/06/2020 h. 09:30	MRI (Magnetic Resonance Imaging)	Doctor Beige	46,14
A	7	11/06/2020 h. 12:30	Blood analysis	Doctor Blue	80,00
A	8	13/06/2020 h. 10:15	Visit schedule	Support service	26,51
A	9	01/07/2020 h. 15:30	Patient official visit	Doctor Silver	36,15
A	10	04/07/2020 h. 14:00	Request of teleconsultation	Doctor Silver	25,83
A	11	05/07/2020 h. 11:00	Evaluation of the request for consultation	Doctor Golden (Specialised doctor)	41,32
A	12	06/07/2020 h. 12:30	Specialist diagnosis sending	Doctor Golden	33,56
A	13	08/07/2020 h. 13:30	Visit schedule	Support service	26,50
A	14	18/07/2020 h. 15:30	Patient official visit	Doctor Silver	36,17
A	15	29/07/2020 h. 13:15	Radiation therapy I cycle	Doctor Magenta	987,81
A	16	10/08/2020 h. 08:30	Blood analysis	Doctor Blue	80,00
A	17	16/08/2020 h. 10:30	Radiation therapy II cycle	Doctor Magenta	987,80
A	18	18/08/2020 h. 08:30	Blood analysis	Doctor Blue	80,00
A	19	21/08/2020 h. 10:30	Radiation therapy III cycle	Doctor Magenta	987,81
A	20	23/08/2020 h. 08:30	Blood analysis	Doctor Blue	80,00
A	21	25/08/2020 h. 09:15	Visit schedule	Support Service	26,51
A	22	10/09/2020 h. 12:00	Patient official visit	Doctor Silver	36,16
A	23	12/07/2020 h. 09:00	Patient discharge	Reception	26,40
B	24	06/06/2020 h. 10:00	Patient acceptance	Reception	26,40
B	25	10/06/2020 h. 12:15	Referral to the requesting doctor	Support Service	26,51
B	26	12/06/2020 h. 13:00	Visit schedule	Support Service	26,50
B	27	05/07/2020 h. 11:30	Patient official visit	Doctor Brown (Corresponding doctor)	36,16
B	28	08/07/2020 h. 13:30	Eco-doppler exam	Doctor Coral	80,00
B	29	10/07/2020 h. 12:00	Blood analysis	Doctor Blue	80,00
B	30	13/07/2020 h. 13:00	CT (Computed Tomography) scanning	Doctor Blue	56,15
B	31	15/07/2020 h. 12:20	Visit schedule	Support Service	26,50
B	32	28/07/2020 h. 15:00	Patient official visit	Doctor Brown	36,16
B	33	06/08/2020 h. 13:00	Request of teleconsultation	Doctor Brown	25,82
B	34	07/08/2020 h. 07:00	Evaluation of the request for consultation	Doctor Lime (Specialised doctor)	41,32

B	35	09/08/2020 h. 08:15	Sending the request for a new test to the Specialist	Doctor Lime	33,54
B	36	11/08/2020 h. 07:00	Patient communication	Support Service	26,50
B	37	14/08/2020 h. 10:00	PET (Positron Emission Tomography) exam	Doctor Azure	42,21
B	38	16/08/2020 h. 08:40	Visit schedule	Support Service	26,50
B	39	25/08/2020 h. 12:15	Patient official visit	Doctor Brown	36,15
B	40	05/09/2020 h. 09:00	Request of teleconsultation	Doctor Brown	25,82
B	41	06/09/2020 h. 08:15	Evaluation of the request for consultation	Doctor Lime	41,33
B	42	08/09/2020 h. 15:15	Specialist diagnosis sending	Doctor Lime	33,56
B	43	10/09/2020 h. 09:10	Visit schedule	Support Service	26,50
B	44	25/09/2020 h. 17:00	Patient official visit	Doctor Brown	36,17
B	45	26/09/2020 h. 07:30	Patient discharge	Reception	26,40

(Source: own elaborations)

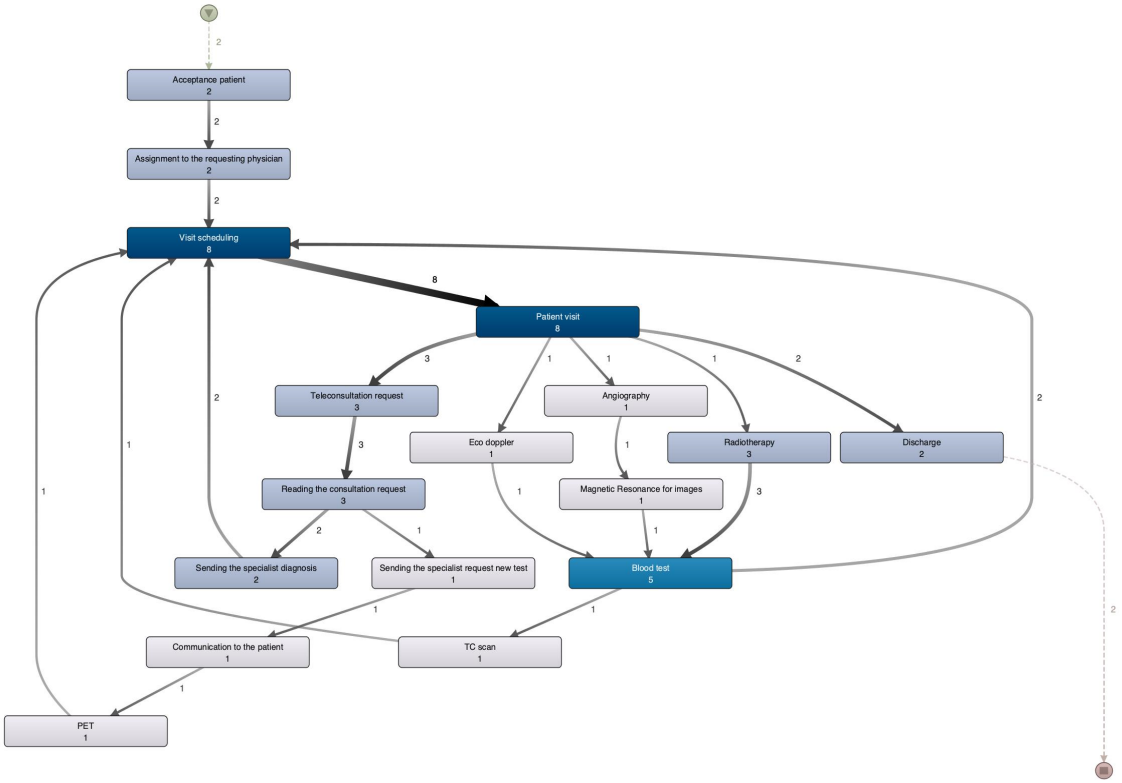


Figure 8: Process discovered (Source: own elaborations)

4. Process mining application and results

In this paragraph, we perform the discovery process as the first step using the Disco software. Elaborating the log file with this software we deliver the process shown in Figure 8.

As we can see, the process obtained from this application does not lead to a good performance. It presents bottlenecks, and even when the teleconsultation tool is used, the case duration is still high.

In particular, the bottlenecks are related to the following resources: Reservation centre (24,44%), Visit scheduling (17,78%), Patient visit (17,78%). Presenting a higher utilisation rate than the other resources, which instead present an underutilised capacity, the above-listed resources led to the entire process slowdown.

Table 3: Statistics for case

Case id	Event	Length	Price	Event price
A	23	118 days	4.073,39	117,10
B	22	104 days 30 minutes	816,22	37,10
Tot.	45		4.889,61	154,20

This aspect is more evident in the following data (Table 3, Table 4 and Table 5).

Now, we use the Business Process Management (BPM) approach in order to model, define, organize and monitor the healthcare processes, analysing and consequently optimising the clinical activities under study [27]. For the graphical representation of process steps, we have adopted the Business Process Model and Notation (BPMN) [28], [29].

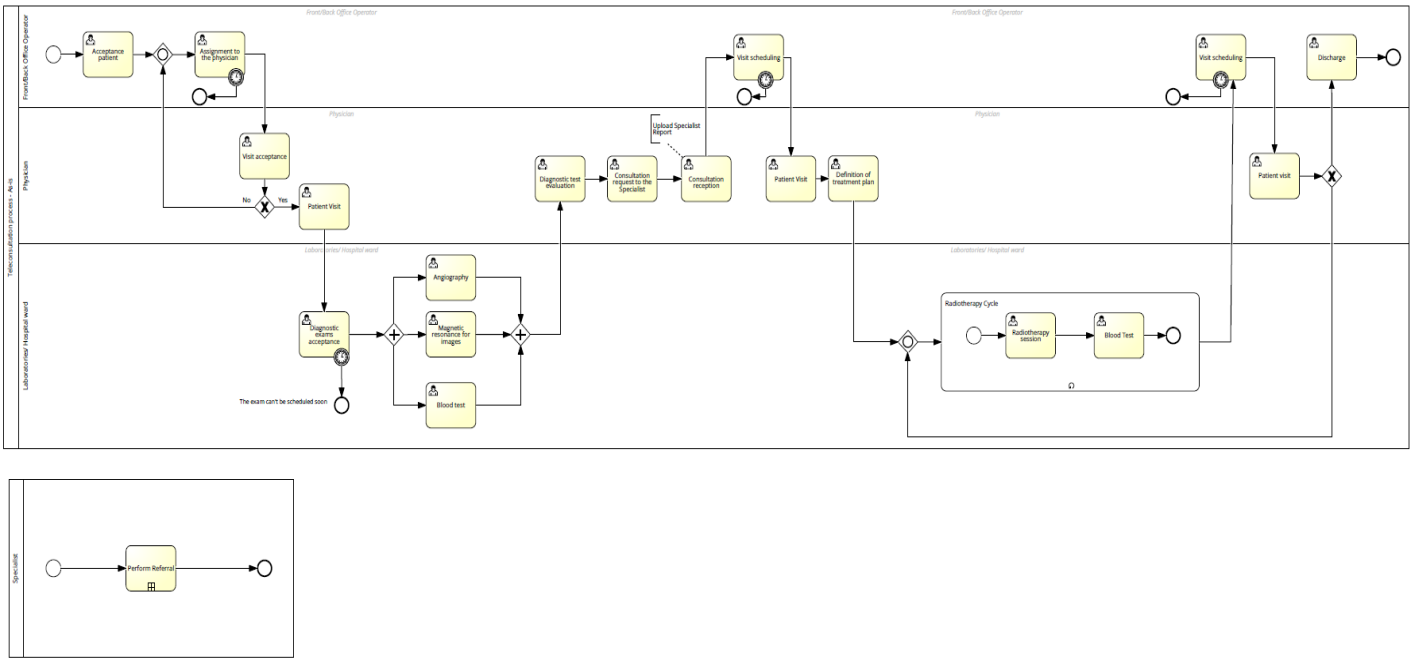


Figure 9: AS-IS process

Table 4: Statistics for activity

n.	Activity	%	Freq.
I	Visit schedule	17,78%	8
II	Patient official visit	17,78%	8
III	Blood analysis	11,11%	5
IV	Request of teleconsultation	6,67%	3
V	Evaluation of the request for consultation	6,67%	3
VI	Radiation therapy	6,67%	3
VII	Patient acceptance	4,44%	2
VIII	Referral to the requesting doctor	4,44%	2
IX	Specialist diagnosis sending	4,44%	2
X	Patient discharge	4,44%	2
XI	Angio exam	2,22%	1
XII	MRI (Magnetic Resonance Imaging)	2,22%	1
XIII	Eco-doppler exam	2,22%	1
XIV	CT (Computed Tomography) scanning	2,22%	1
XV	Sending the request for a new test to the Specialist	2,22%	1
XVI	Patient communication	2,22%	1
XVII	PET (Positron Emission Tomography) exam	2,22%	1
	Tot.	100,00%	45

Table 5: Statistics for resource

n.	Resource	%	Freq.
I	Support Service	24,44%	11
II	Doctor Blue	13,33%	6
III	Doctor Brown	13,33%	6
IV	Doctor Silver	11,11%	5
V	Reception	8,89%	4
VI	Doctor Lime	8,89%	4
VII	Doctor Magenta	6,67%	3
VIII	Doctor Coral	4,44%	2
IX	Doctor Golden	4,44%	2
X	Doctor Beige	2,22%	1
XI	Doctor Azure	2,22%	1
	Tot.	100,00%	45

In particular, the process flow modelled with BPMN notation [30] using Signavio tool is visible in the following figure (Figure 9).

5. Discussion

Starting from the results presented in the previous section, it emerges that our application reveals a fundamentally non-performing process. The analysis actually showed that the process is very slow (almost three months in both cases described), and there are bottlenecks in both resource and activities utilisation.

Then, we match the adopted practices with the best practices implemented in Piedmont, considered a virtuous region. The log file of Piedmont was not available, and therefore we were not possible to perform a compliance between the two processes. From observing data from the Ministry of Health (Table 6), the number of test equipment in Apulia was much lower, and this creates bottlenecks, since many more tests are needed in Apulia than in Piedmont. Furthermore, from the same statistics, we realize that the number of personnel employed in Piedmont was higher than in Apulia, and this affected the unitary personnel performance which, therefore, is higher in Apulia than in Piedmont. This leads to an over-utilisation of the units of work, and consequently bottlenecks also in terms of care pathway duration. These aspects are highlighted in Table 6.

Table 6: Data

Category	Classification	Region	Value
Medical devices number	Angio exam	Region 1	31
		Region 2	12
	MRI (Magnetic Resonance Imaging)	Region 1	33
		Region 2	16
	Eco-doppler exam	Region 1	364
		Region 2	138
	PET (Positron Emission Tomography) exam	Region 1	2
		Region 2	0
	CT (Computed Tomography) scanning	Region 1	80
		Region 2	49
Medical devices performance (per number)	Eco-doppler exam	Region 1	1646,8
		Region 2	2818,2
	CT (Computed Tomography) scanning	Region 1	3960,0
		Region 2	3986,9
	MRI (Magnetic Resonance Imaging)	Region 1	2041,3
		Region 2	2694,4
Age of devices (%)	> 15 years	Region 1	6,5%
		Region 2	18,3%
	10-15 years	Region 1	20,7%
		Region 2	16,6%
	10-7 years	Region 1	13,7%
		Region 2	19,1%
	7-3 years	Region 1	23,5%
		Region 2	18,0%
	< 3 years	Region 1	35,7%
		Region 2	28,0%
Staff workload	Performance per staff unit	Region 1	1362
		Region 2	1647
	Performance by technician	Region 1	2648
		Region 2	3114
	Performance by physician	Region 1	5800
		Region 2	6215
Operating personnel	Facility managers	Region 1	107
		Region 2	60
	Medical managers	Region 1	507
		Region 2	271
	Physical	Region 1	47
		Region 2	12
	Radiology technicians	Region 1	1110
		Region 2	541
	Nurses Professional	Region 1	191
		Region 2	186
	Auxiliaries	Region 1	242
		Region 2	117
	Administrative	Region 1	319
		Region 2	42
	Other personnel	Region 1	92
		Region 2	21

(Note: Piedmont = Region 1, Apulia = Region 2, own elaborations, source: Ministry of Health)

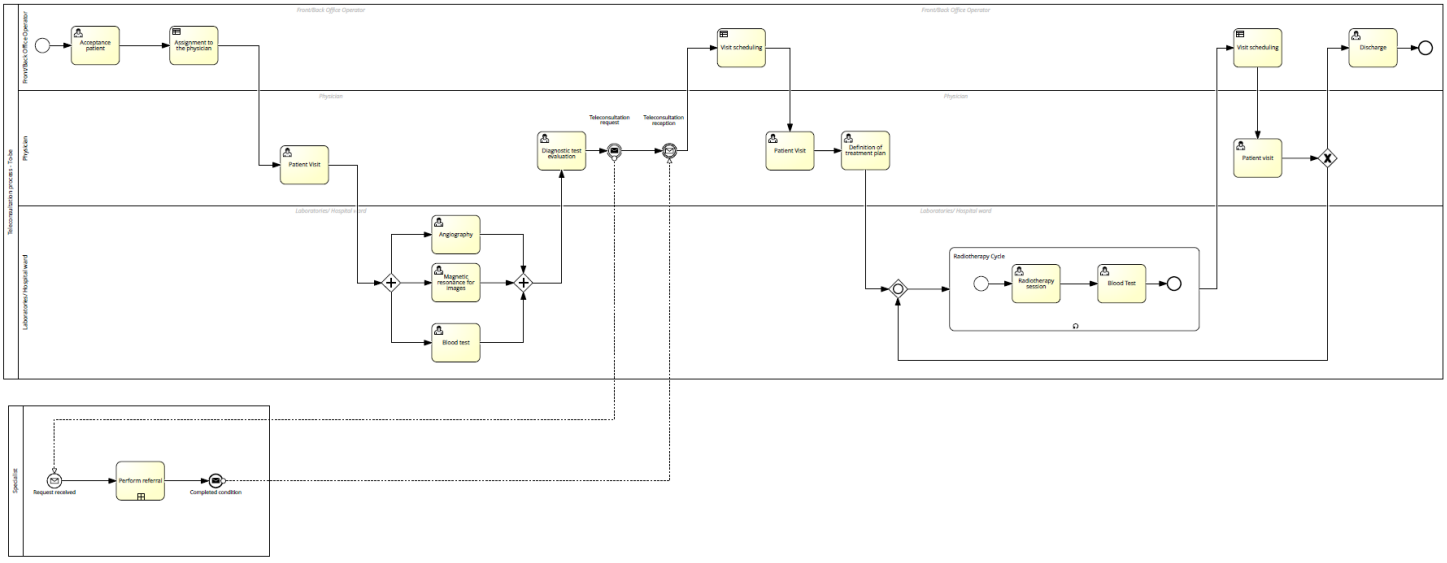


Figure 10: TO-BE process

Given all the problems that we faced in the process, as from the actual application represented by the log file, we designed a new process aimed at improving this process. Everything is represented in Figure 10.

Related to the AS-IS process, we have improved the TO-BE process in terms of time. In fact, in the TO-BE process modelling, the acceptance and scheduling phases of the visit by the doctor rely on certain decision-making procedures well-defined by the system (based on availability and time of doctors).

The TO-BE process included a change in assigning diagnostic examinations to different technicians in order to optimise the use of resources (personnel and healthcare equipment) and waiting times. Furthermore, as far as the teleconsultation is concerned, we have improved the TO-BE process over the AS-IS process. In particular, in the TO-BE process, the teleconsultation is made automatic both in the phase of specialised doctor identification for a second opinion, and in the consultation requesting phase. Even in the case of sending a second opinion, the TO-BE process is more effective, since it uses the same archiving system, which allows the corresponding doctor to examine the diagnosis in a timely manner.

Moreover, with the TO-BE process, we were able to decrease patients waiting duration and to improve the resources utilization, consequently reducing costs and optimising the healthcare service.

6. Conclusions

This paper demonstrates that process mining is a very-often-used technology in business and industrial processes but can also find useful application in the healthcare sector.

In fact, the process mining tools allow to decrease costs, increase employee and customer satisfaction, and to operate more proactively on a daily basis, which is fundamental in the healthcare sector, where it is crucial to improve the services to patients, especially in terms of care duration and cost reduction (primarily on the management side), taking into consideration the considerable financial commitment required by the sector.

However, although (approximate) data are available to create log files, mostly of them are very personal and not-usable. In fact, due to privacy and confidential reasons, hospital organizations are not disposed to grant or share them.

In conclusion, the application of process mining could make available tools finalised to improve the territorial equalisation of the public intervention (expenditure levels homogenization).

In view of the findings above, it would be suitable to use process mining technique in the health sector, and our hope is that this will take place in the future.

References

[1] A. Corallo, M. Lazoi, R. Paiano, F. Striani, “Application of process mining in teleconsultation healthcare: Case study of Puglia hospital. In Proceedings of the 10th International Conference on Information Systems and Technologies”, 1-13, 2020, doi: 10.1145/3447568.3448540.

[2] W.V.D. Aalst, A. Adriansyah, A.K.A.D. Medeiros, F. Arcieri, T. Baier, T. Blickle, M. Wynn, “Process mining manifesto. In: International conference on business process management”, Springer, Berlin, Heidelberg, 169-194, 2011, doi: 10.1007/978-3-642-28108-2_19.

[3] W. Van der Aalst, “Process Mining: Overview and Opportunities” ACM Trans. Manag. Inf. Syst., **3**, 1–17, 2012, doi: 10.1145/2229156.2229157.

[4] C.W. Günther, W.M.P. Van der Aalst, “Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics”, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, **4714**, 2007, doi: 10.1007/978-3-540-75183-0_24.

[5] A. Weijters, W. Van der Aalst, A. de Medeiros, “Process Mining with the Heuristics Miner Algorithm”, Technical Report, WP166 Beta Paper Series. Eindhoven University of Technology, Eindhoven, The Netherlands, 2006.

[6] R. Accorsi, Ernesto D, W. Van Der Aalst, “Unleashing operational process mining”, 2014, doi: 10.4230/DagRep.3.11.154.

[7] W. Van der Aalst, “Process Mining: Discovery, Conformance and Enhancement of Business Processes”, Springer: Berlin, Germany, Heidelberg, Germany, 2011.

[8] G. Geleijnse, H. Aklecha, M. Vroling, R. Verhoeven, F.N. Van Erning, P.A. Vissers, X.A. Verbeek, “Using Process Mining to Evaluate Colon Cancer Guideline Adherence with Cancer Registry Data: a Case Study”, 2018.

[9] C. Fernandez-Llatas, B. Valdivieso, V. Traver, J.M. Benedi, “Using Process Mining for Automatic Support of Clinical Pathways Design”, In Data Mining in Clinical Medicine, Number 1246, Springer: New York, NY, USA, 79-88, 2015, doi: 10.1007/978-1-4939-1985-7_5.

[10] R. Mans, W. Van der Aalst, R. Vanwersch, “Process Mining in Healthcare”, Springer Briefs in Business Process Management, Springer International Publishing, Cham, Germany, 2015.

[11] E. Rojas, J. Munoz-Gama, M. Sepúlveda, D. Capurro, "Process mining in healthcare: A literature review", **61**, 224–236, 2016, doi: 10.1016/j.jbi.2016.04.007.
[12] F. Caron, J. Vanthienen, K. Vanhaecht, E.V. Limbergen, J. de Weerdt, B. Baesens, "Monitoring care processes in the gynecologic oncology department", Comput. Biol. Med., **44**, 88–96, 2014, doi: 10.1016/j.compbiomed.2013.10.015.
[13] V. Tandon, A. Raheja, S. Mishra, K. Garg, C. Dash, S.A. Borkar, S.S. Kale, "Trends in neurosurgical teleconsultation services across the globe during COVID-19 pandemic", World Neurosurgery, **150**, 2021, doi: 10.1016/j.wneu.2021.03.070.
[14] P. Baudier, G. Kondrateva, C. Ammi, V. Chang, F. Schiavone, "Patients' perceptions of teleconsultation during COVID-19: A cross-national study", Technological Forecasting and Social Change, **163**, 120510, 2021, doi: 10.1016/j.techfore.2020.120510.
[15] P. Baudier, G. Kondrateva, C. Ammi, V. Chang, F. Schiavone, "Digital transformation of healthcare during the COVID-19 pandemic: Patients' teleconsultation acceptance and trusting beliefs", Technovation, 102547, 2022, doi: 10.1016/j.technovation.2022.102547.
[16] C. Fernandez-Llatas, J.M. Benedi, J.M. Garcia-Gomez, V. Traver, "Process Mining for Individualized Behavior Modeling Using Wireless Tracking in Nursing Homes", Sensors, **13**, 15434–15451, 2013, doi: 10.3390/s131115434.
[17] R. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini, W. Van der Aalst, "Process mining techniques: An application to stroke care", Stud. Health Technol. Inf., **136**, 573–578, 2008.
[18] R. Dunkl, K.A. Froschl, W. Grossmann, S. Rinderle-Ma, "Assessing Medical Treatment Compliance Based on Formal Process Modeling", In Information Quality in E-Health, Number 7058 in Lecture Notes in Computer Science, Springer: Berlin, Germany; Heidelberg, Germany, 533–546, 2011, doi: 10.1007/978-3-642-25364-5_37.
[19] A. Vathy-Fogarassy, I. Vassányi, I. Kósa, "Multi-level process mining methodology for exploring disease-specific care processes", 2022, doi: 10.1016/j.jbi.2021.103979.
[20] D. Gough, S. Oliver, J. Thomas, "An introduction to systematic reviews", Sage, 2017.
[21] A. Corallo, M. Lazoi, F. Striani, "Process mining and industrial applications: A systematic literature review", Knowledge and Process Management, **27**(3), 225-233, 2020, doi: 10.1002/kpm.1630 .
[22] Fluxicon Laboratories, Disco, Available online: https://fluxicon.com/disco/
[23] M. Egger, G.D. Smith, D.G. Altman, "Systematic reviews in health care: meta-analysis in context", BMJ Books, London, 2008.
[24] F. Striani, "Green and Blue Economy: Definitions, Challenges and Limits", International Journal of Environmental Sustainability and Green Technologies (IJESGT), **11**(2), 16-33, 2020, doi: 10.4018/IJESGT.2020070102.
[25] T. G. Erdogan, A. Tarhan, "Systematic mapping of process mining studies in healthcare", IEEE, 2018, doi: 10.1109/ACCESS.2018.2831244.
[26] Z. Valero Ramón, "Dynamic risk models for characterising chronic diseases' behaviour using process mining techniques", Doctoral dissertation, Universitat Politècnica de València, 2022, doi: 10.4995/Thesis/10251/181652.
[27] M. Weske, "Business process management architectures - Third edition", Springer, Berlin, Heidelberg, 2019.
[28] D. Bano, J. Michael, B. Rumpe, S. Varga, M. Weske, "Process-aware digital twin cockpit synthesis from event logs", Journal of Computer Languages, 101121, 2022, doi: 10.1016/j.cola.2022.101121.
[29] M. Dumas, M. La Rosa, J. Mendling, H.A. Reijers, "Introduction to business process management", In Fundamentals of business process management, Springer, Berlin, Heidelberg, 1-33, 2018, doi: 10.1007/978-3-662-56509-4_1.
[30] M. Ramos-Merino, L.M. Álvarez-Sabucedo, J. M. Santos-Gago, J. Sanz-Valero, "A BPMN based notation for the representation of workflows in hospital protocols", Journal of medical systems, **42**(10), 1-10, 2018, doi: 10.1007/s10916-018-1034-2.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 5, 157-166 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Detection of Event-Related Potential Artifacts of Oddball Paradigm by Unsupervised Machine Learning Algorithm

Rafia Akhter*, Fred Beyette

School of Engineering, Electrical Computer Engineering, University of Georgia, Athens, USA

ARTICLE INFO

Article history:
Received: 26 July, 2022
Accepted: 08 October, 2022
Online: 31 October, 2022

Keywords:
Electroencephalography (EEG)
Event-Related Potential (ERP)
Oddball Paradigm
Artifact Corrupted Epochs
Artifacts
Detection

ABSTRACT

Electroencephalography (EEG) is one of the most common and benign methods for analyzing and identifying abnormalities in the human brain. EEG is an incessant measure of the activities of the human brain. In contrast, when the measurement of EEG is bounded by time and the EEG is synchronized to an exterior stimulus, is known as Event-Related Potential (ERP). ERP has the capability to perceive and explore the human brain's responses to specific sensitive, cognitive, or motor events in real time with high temporal resolution. Among the various techniques, the oddball paradigm is very famous in EEG studies. In an oddball paradigm experiment, brain responses to frequent and infrequent stimuli are measured. However, the success of ERP research is very much dependent on the analysis of clean data sets and unfortunately, EEG is a combination of both neural and non-neural activities which introduce significant sources of noise that are not related to the brain's response to the external stimulus. These unrelated non-EEG components are acknowledged as artifacts and due to these, the quality of the EEG may damage by decreasing SNR (signal-to-noise ratio). In addition, these artifacts may mislead the actual information in the study. Addressing this problem, the purpose of this research is to introduce a machine learning algorithm (ML) that can screen EEG/ERP data to remove data epochs that are disrupted by artifacts and thus produce a clean data set. Overall, three unsupervised ML algorithms are applied to identify noisy epochs and it is found that the DBScan method performs best with 93.43% accuracy. Finally, the success of this study will allow the ERP study to have a cleaner ERP data set in normal laboratory conditions with less complexity in the ERP studies.

1. Introduction

This study is an extension of our previously published [1] paper and was presented at the 7th International Conference on Data Science and Machine Learning Applications (CDMA, 2022). In this paper, we identified and detected the auditory ERP artifacts by unsuper- vised Machine Learning Algorithms (MLAs) and compared the results and features with the visual Event-Related Potential (ERP) artifacts that were presented at the CDMA conference.

Electroencephalography (EEG) is a safe and painless method of measuring the human brain's electrical activities in real-time. In medical research, EEG experiments are very convenient. EEG measures the electric potential of the brain over a continuous time and EEG activities over a bounded time are acknowledged as Event-Related Potential (ERP) [2]. ERPs can explore the brain's response to a specific sensory input with a high temporal resolution. In addition, ERPs are potential for the use of human biomarker, and other cerebral processes [3-5].

The oddball paradigm is one of the common experimental methods in ERP research. In the oddball paradigm, there is a sequence of monotonous stimuli, and it is irregularly interrupted by an uncommon stimulus [6]. In this study, the experimental work is comprised of a typical auditory oddball paradigm. Here, the test subject heard a series of tones with two different pitches. One of the tones is played much more frequently than other. For example, a common tone played 80% of the time with a randomly interspersed

*Corresponding Author: Rafia Akhter, email: Rafia.Akhter@uga.edu

https://dx.doi.org/10.25046/aj070517

uncommon tone making up the remaining 20% of tones. It has been established by different studies that ERP gives a maximum positive peak of around 300 ms- 600 ms and the peak is higher for target/oddball stimuli compared with standard stimuli. This component is known as the P300 component [7].

However, EEG in its raw form is a mixture of neural and non-neural activities. Any non-neural activities are unnecessary in EEG research and recognized as artifacts in EEG/ERP research [8]. These artifacts may produce erroneous results in ERP studies in various ways. For example, may damage the ERP signal quality by reducing the signal-to-noise ratio , there could some arbitrary artifacts which occur infrequently for one certain condition for one test subject while for other test subjects it may occur rarely. As a result, there might be huge differences in the evaluation of two test subjects for the same experiments [9]. There are many more reasons for artifacts that may mislead the conclusion of any ERP study.

There are a huge number of sources of EEG artifacts but the most common artifacts are related to eye and body movements. In this study, our goal is to detect eye-blink artifacts, eye-movement artifacts, and body movement artifacts in normal laboratory conditions. Although much research has been done, there is no standard technique for detecting and eliminating artifacts. Addressing this, the aim of this study is to introduce a method of machine learning (ML), in which we can identify artifact corrupted ERP epochs and by removing those from the dataset, we will have a clean dataset. This experiment is done with the addition of some external effort to create artifacts. So that we could detect these artifacts. As a result, the outcomes from this study will improve the signal quality of the ERP experiment.

We applied the anomaly or outlier detection method of MLA and our data was unlabeled i.e. we applied, unsupervised MLA. Unsupervised MLAs find uncommon data points which have different properties compared to others, in any dataset. We applied three unsupervised MLAs for the detection of artifacts. They are Density-Based Spatial Clustering of Applications with Noise (DB-Scan), Isolation Forest (IsoF), and Local Outlier Factor (LOF). We measured the artifact-mixed ERP epochs detection efficiency of these methods and compared the accuracy of efficiency with the standard EEGLab method. We found the DBScan method is most efficient with 93.43% accuracy while the other methods also showed a good efficiency, ranging from 85% to 87%.

EEG is frequently utilized to analyze epilepsy, which causes variations from the norm in EEG readings. It is additionally utilized to analyze rest clutters, the profundity of anesthesia, coma, encephalopathies, and brain passing[10]. In general, EEG produces amplitude with respect to time and it is a very sensitive measurement. These measures are very important for clinical decisions. But artifacts may change or hides the information in EEG and re- moving these artifacts may take a long time for analyzing the data by extracting features. But MLAs have shown faster learning to process EEG signals [11] by outlier or anomaly detection method. Isolation Forest is an effective MLA for this detection with linear computational complexity. Screening anomaly is useful for the detection of epileptic seizures [12]. LOF can identify artifacts by producing abnormal scores using statistical methods [13]. DBScan uses a clustering-based algorithm to detect artifacts[14]. Overall, MLAs have shown effective importance in analyzing epilepsy and other neurological diseases.

The organization of the paper is as tracks: section 2 consists literature review, the experimental setup and procedure are explained in Section 3, identification of artifacts are in section 4, Section 5 contains practical implementations, and finally, results and conclusions are in Sections 6 & 7.

2. Literature Review

A lot of research has been done to remove artifacts but most methods require labeling the artifacts manually or requiring, additional hardware. For example, Electrooculography electrodes may require to place around the eyes or may necessitate data-sets containing a huge amount of data, and many more [15]. The involvement of humans to label artifacts in EEG data may be not desirable as it might be a tedious and time-consuming process [16].

In [17], the author described an unsupervised EEG artifact detection algorithm. It shows that this algorithm is effective to identify eyeblinks with 98.15% accuracy. They collected their dataset with the OpenBCI system and used EEGLab. In their experiment, sub- jects were instructed to watch a video and read articles, each for 5 minutes. They compared the methods with SVM and k-NN which are learning-based methods. But the accuracy of the performances of these methods was very low, 46.49% 67.82% comparatively.

In [18], the author established a deep learning method using Bayesian and attention modules to improve the performance of the classifier. Here, after the filtering process, to remove line noise, the artifact subspace reform (ASR) technique [19] was revised to remove an artifact that is dispersed throughout the entire scalp with a huge variance. The infomax-ICA technique was then directed to get a set of ICs establishing EEG and artifacts [20]. To end, topographic plots were grown and labeled by EEG experts. The classification accuracy was very high, around 95% but this method needs an attention module, task-dependent, and an EEG expert is required.

In [21], mthe author depicted an EEG noise-reduction scheme that customs representation knowledge to perform patient- and task-specific discovery of artifacts and correction. More specifically, their method is dependent on a given task and extracted 58 features from the signals.

3. Experimental Setup and Procedure

3.1. Dataset

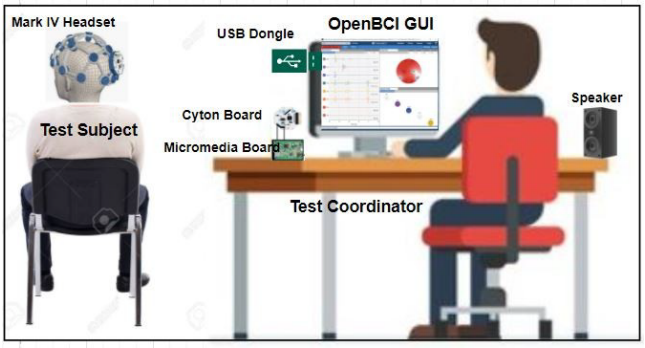
This auditory oddball paradigm EEG data set is approved by the Institutional Review Board (IRB) at the University of Georgia, Athens. All of the participants provided knowledgeable permission. In this study, the number of test subjects was 13 (both male and female). They all were a minimum of 18 years of age. They had no psychiatric conditions and no hearing or eye-sight weakening. The EEG experiment had 3 setups and for every test subject, there were 3 data sets. The first data set is a mind wandering data set and the last two are artifact corrupted/test data sets. For 13 subjects, there is 13 mind wandering data sets and 26 test data sets. However, in 3 of the test data sets, almost all the ERP epochs were characterized by

test subject distraction, and those 3 data sets were not considered in our study. So, there was a total of 26 test data sets.

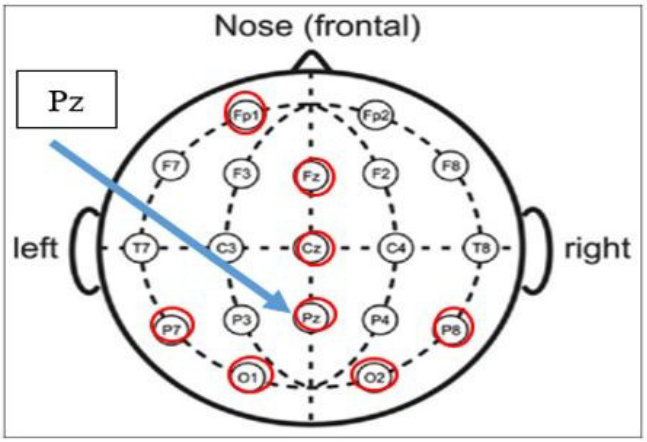
For the recording and collection of data, an OpenBCI EEG capture system with an Ultracortex Mark IV headset was used. Here, the electrodes were connected according to the international 10-20 system along the surface of the scalp [22]. The 8-channel Cyton board was mounted to the headset. We connected electrodes to the frontal (FP1, Fz), central (Cz, C4), parietal (Pz, P7, P8), and occipital portion (O1, O2) in the human scalp (Fig. 1B). The Cyton Board was wirelessly connected via a Bluetooth (4.0 Low Energy BLE) to a data collection computer. To generate the auditory stimuli for the ERP oddball experiments, we used a Mikromedia PIC24EP board, available from MikroElectronika Inc [23].

3.2. Data Collection Procedure

In this auditory oddball paradigm, subjects hear a sequence comprised of 2 tones with a different pitch where one of the tones is played much more frequently than the other (ex. a common tone played 80% of the time with a randomly interspersed uncommon tone making up the remaining 20% of tones).



(a)



(b)

Figure 1: (a) The experimental setup of auditory oddball task. There were a series of tones with two different frequencies of 1000Hz and 2000Hz. The tones were played randomly with the stimuli duration of 200 ms and ISI of 3s 50 ms); (b) The International 10/20 system of electrode positions. The red-colored electrode positions were used to record the data in this experiment. For all the calculations, the measurements of the Pz location (marked in blue) were used.

In this experiment, to record EEG (as shown schematically in Fig. 1A), the Ultracortex Mark IV headset was used. There were 50 tones with two different pitches. One tone with 1000 Hz is known as a " common" tone and the other with 2000 Hz, is "uncommon or oddball ". The subject was instructed to concentrate on the uncommon tones and ignore the frequent tones. The proportion of frequent and uncommon tones was 80:20. The stimuli or tone duration was 200 ms and the Inter-stimulus Interval was 3s 50 ms.

There were 3 experimental setups. In the 1st setup, there was no instruction for the test subjects. The subject heard the tones and data were recorded. The data set from this setup is named a mind-wandering data set. At the 2nd setup, the subjects were instructed to give their special concentration to counting the uncommon/oddball tones and try to overlook the common tones. In the 3rd setup, the experimental condition was the same as in the 2nd, but we made an unexpected disturbance. For example, we made an abrupt flash on another monitor to divert the mind of the test subject, so that he/she would have somebody's movements or blink their eyes.

The data sets collected for the last two setups are named artifact corrupted/test data sets. After completion of each dataset collection procedure, there was a 1–3-minute break of "mental rest" before initiating the next data collection interval.

3.3. Processing of Dataset

There were 3 data collection intervals for every subject and there was a total of 26 test data sets, as described in section 3.1. After recording, the data sets were protected in CSV format and EEGLab (a tool of MATLAB) was used for the primary analysis [24]. The DC offsets were uninvolved, and a band-pass filter (non-causal) of 0.1 Hz-30 Hz was used for the filtering of the EEG signals. For the ERP epochs extraction, a time window of -1000 ms to 2000 ms was bounded and ERP features were extracted in an adequate window of 200 ms to 600 ms after the commencement of the stimulus for the ERP evaluation. For further calculation, all measures (amplitude and latency) were evaluated from the uncommon stimuli at the Pz electrode position.

4. Artifacts Detection

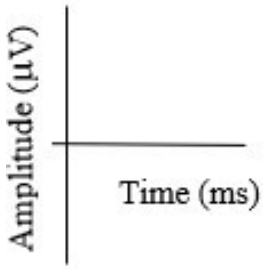
The signal recorded in ERP experiments is a combination of EEG plus the non-neural source of activities. These non-neural sources are mainly from the induced electrical signals of the recording environment (e.g., line noise from lights and computers) and human biological activities. For example, it is very usual that while recording EEG, a test subject may blink his eye, moves his eye, he may move his head, or body, there may be muscle activities, have skin potential, and many more [25]. All of these may create non-neural signals in EEG recording and these are considered artifacts. As a result, EEG signals are often contaminated by several artifacts such as the electrooculogram (EOG), the electromyogram (EMG), electrocardiogram (ECG) and motion artifacts are a result of, for example, an eye-blink activity is an EOG artifact.

In this study, we tried to detect eye-blink, eye-movement, and body-movement artifacts. Due to the spontaneous movement of eyelids, the EOG artifact is normally always present in EEG signals. When a subject blinks his eyes, an eye-blink artifact appears as a high amplitude spike in the EEG signal [26]. Eyeblink responses are opposite in polarities compared to EEG signals and usually consist of a deflection of 50-100 V with a typical duration of 200-400 ms.

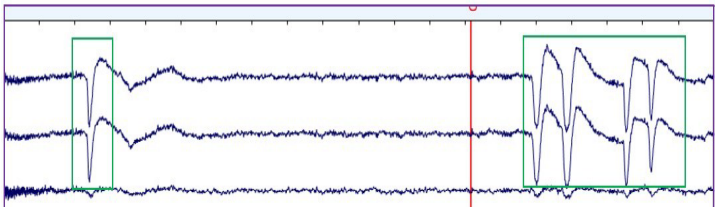
In addition, there are either horizontal eye movements or vertical ones. For a horizontal movement of eyes (HEOG), there is a higher

positive voltage over the side of the head that the eyes now point toward. For a leftward eye movement, a positive-going voltage deflection is shown on the left side of the scalp, and a negative-going voltage on the right. In the case of vertical eye movements (VEOG), higher deflection shows between the electrodes below and above the eyes[27,28].

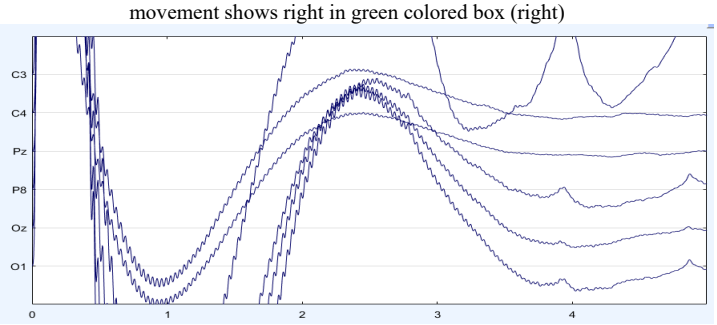
The movement of the body is very obvious for a test subject. Body movements create huge fluctuation of voltage levels and there may be high voltage levels which may shift upward or/and downward drift [29]. If these artifacts are not removed from the dataset, then the measurement or the values of the ERP features may be totally in the wrong format. All of these create huge artifacts in EEG and change the measurement levels of ERP features. Fig.2 shows the artifacts in EEG recording for subject 1.



Axis description of the figures. x-axis: time in millisecond (ms); y-axis : EEG voltage amplitudes in the microvolt



(a) EEG eye blink artifacts shows in green box (left) and artifacts due to eye-movement shows right in green colored box (right)



(b)EEG body movement artifacts

Figure 2: Examples of artifacts in EEG recordings for Subject 1

5. Practical Implementation

At first, we detected the artifact corrupted ERP epochs by a standard method (EEG Lab) and after that, identified by unsupervised Machine Learning Algorithms (MLA). For detecting artifact corrupted ERP epochs , a maximum of 150 V (+/-) was set for peak-amplitude detection and 25 V(+/-) maximum for mean amplitude. The overall process of this study is shown in the following block diagram (Fig. 3).

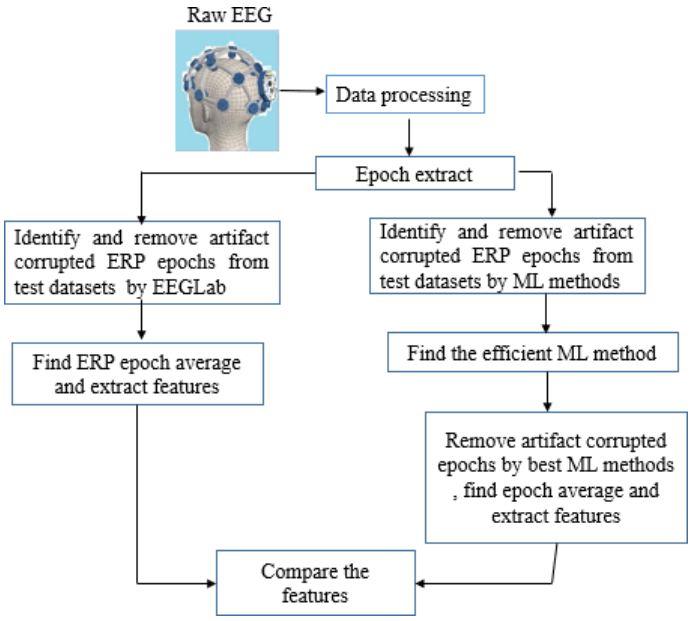
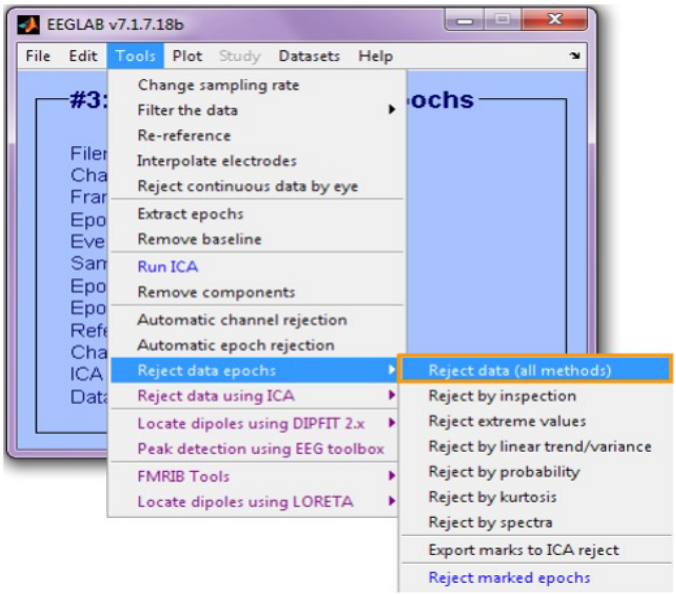


Figure 3: The block diagram of this research shows all the steps of the auditory odd-ball paradigm experiment and the comparison step with the visual oddball paradigm experiment

1.1 EEGLab: Standard method of ERP artifacts detection

EEGLab is one of the most popular EEG software for EEG analysis [30]. It is a freely available open-source toolbox [31] that provides an interactive graphical user interface (GUI), allows users to flexibly and interactively process their high-density EEG, is capable to do the dynamic brain data using independent component analysis (ICA), and able to spectral time/frequency and coherence analysis, as well as standard methods including event-related potentials (ERP) [32] and many more. For all of these reasons, we used EEGLab as a standard method of EEG/ERP analysis. Fig. 4 shows how these parameters for artifact rejection can be set for automatic artifact identification in EEGLab. Beyond these simple approaches, there are many other methods to detect and reject artifacts in the EEG dataset.

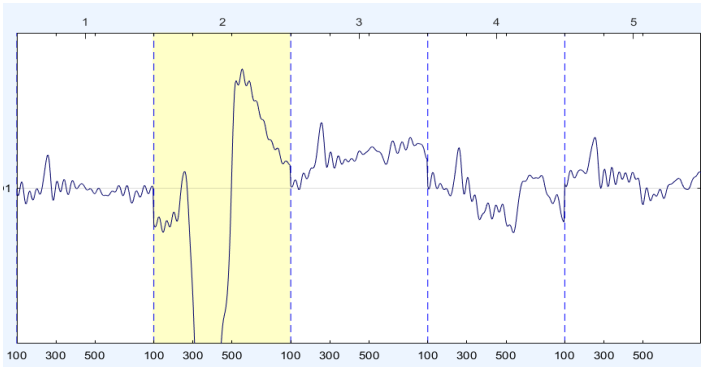


(a) EEG Lab Menu bar shows the methods of detecting and rejecting abnormal ERP epochs

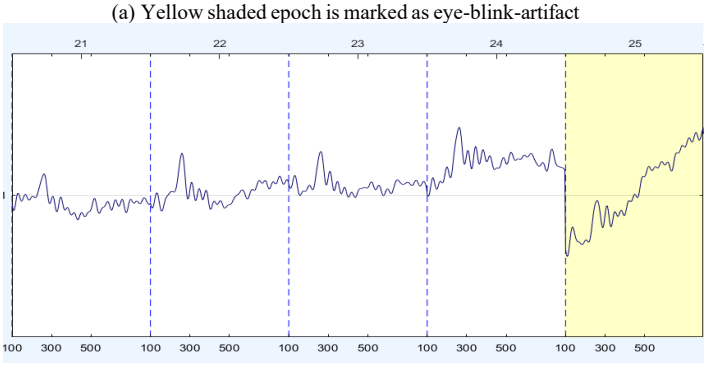


(b) Shows the parameters selection options for rejecting ERP epochs

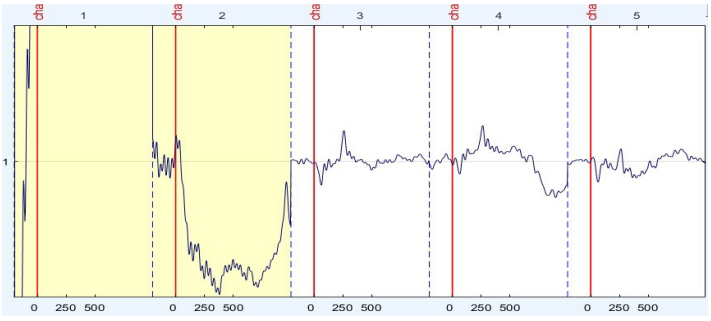
Figure 4: EEGLab processing steps for artifact corrupted ERP epochs detection



(a) Yellow shaded epoch is marked as eye-blink-artifact



(b) Yellow shaded epoch is marked as body movement-artifact



(c) Yellow shaded epoch is considered as an artifact corrupted epoch due to abnormal spectra

Figure 5: Examples of ERP artifacts detected by EEGLab (yellow shaded)

We performed the following steps by EEGLab [33] Toolbox:

- Filter
- Run ICA
- Remove components
- Extract epochs
- Reject Data epochs

In this study, we applied the method "Reject Data Epochs" by:

- Reject extreme values
- Reject by abnormal spectra
- Reject by the linear trend

The artifact corrupted ERPs detected by EEGLab methods, following the a bove steps are shown in Fig. 5. Among the three figures, in every figure, the yellow shaded epochs are not similar to other epochs. These are detected as anomalous ERPs by EEGLab "Reject Data Epochs" methods. For example, in fig 5A, the yellow shaded epoch (2nd from left) are showing a much higher peak compared to others. It is known that, for eye blink, EEG amplitudes give a higher peak. For this reason, it is marked as an eye-blink corrupted epoch.

5.1. Machine Learning Algorithms

Machine Learning Algorithms (MLAs) generally can predict output from the given input values [34-36]. In addition, MLAs can identify the outliers in the data-set in a very fastest ode with good efficiency. Outlier points are significantly different from the majority of the other data points[37] and the process of finding the outliers in the data-set is known as anomaly detection. MLAs have both supervised and unsupervised learning approaches. Supervised learning uses labeled data to help predict outcomes. On the other hand, unsupervised learning does not use labeled data [38]. They analyze and discover hidden patterns and return the data points with abnormal behavior. In this study, to detect the artifact corrupted ERP epochs, the following 3 features were used. i. Mean of ERP amplitude (mean), ii. The peak of ERP amplitude (peak P300), and iii. latency of the peak ERP (known as P300) (peak latency) in the window of 200 to 600 ms. In this study, we applied three unsupervised MLAs for identifying artifact corrupted ERP epochs based on anomaly detection. They are:

1. Isolation Forest (IsoF)
2. Local Outlier Factor (LOF)
3. DBScan

The Local Outlier Factor (LOF) algorithm is an unsupervised anomaly detection method that calculates the region density aberration of a definite data point concerning its neighbors. It considers as deviations from the norm the samples that have a noticeably lesser density compared to the neighbors[39]. In Fig. 6, there are two neighbors, C1 C2. and there are two outliers, P1, and P2. The neighbor's numbers are a typical set of (i)more prominent than the least number of tests a cluster must contain so that other tests can be nearly exceptions relative to another cluster, and (ii) smaller

than the supreme number of close-by tests and these are termed theoretically as local outliers.

Isolation Forest (IsoF) is an unsupervised MLA that separates observations by selecting random features and returns the unusual score of each test point. This algorithm describes that anomalies are data points with unusual behavior and they are few. IsoF could be a tree-based show where segments are formed by randomly selecting a feature and after that picking an arbitrary split value between the minimum and supreme worth of the chosen feature[40]. Fig. 7 shows that red circles are separated from other normal points (blue) due to unusual behavior.

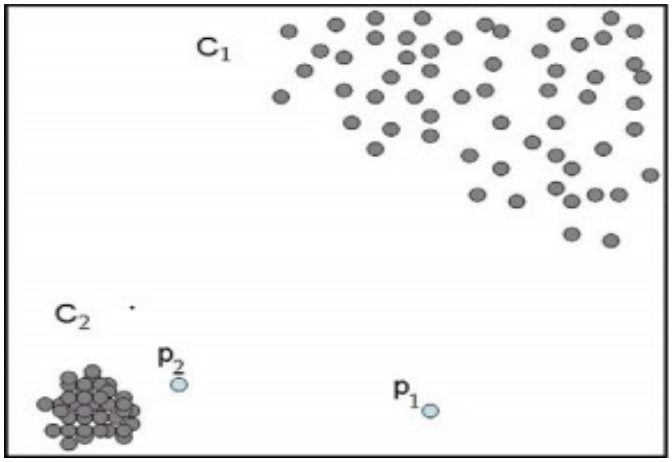


Figure 6: Local Outlier Factor. There are two neighbors, C1 C2. and there are two outliers, P1, and P2. [source: Arun Mohan," Local Outlier Factor", Medium.com, Dec 31, 2008]

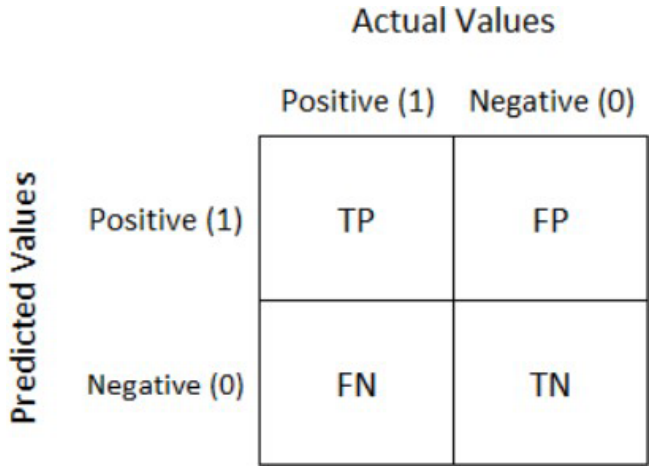


Figure 7: Isolation Forest. In this figure, red circles are separated from other normal points (blue) due to unusual behavior.

DBScan Clustering is an unsupervised MLA that's used to gather data into clusters. It is a famous outlier detection method based on density. This process reveals central samples of high density and expands cluster density. More accurately, this algorithm sees clusters as a region of high density isolated by low density. For this relatively basic view, clusters found by DBSCAN can be in any form[41]. In this procedure, there are two sorts of parameters and three sorts of information focus. One parameter is "eps" (maximum range of the area) and another is "minpts" (least number of facts in the eps-area of a central point).

When any data point contains at least "minpts", are known as "core points" and when the quantity of points is less than "minpts", are known as " border points". In addition, within the "eps" range, when there are any points not surrounded by other points, i.e., absolutely alone in the eps range, those data point is known as an outlier, In Fig.8, there are two clusters, colored as deep-blue and green where outlier is marked as a red circle.

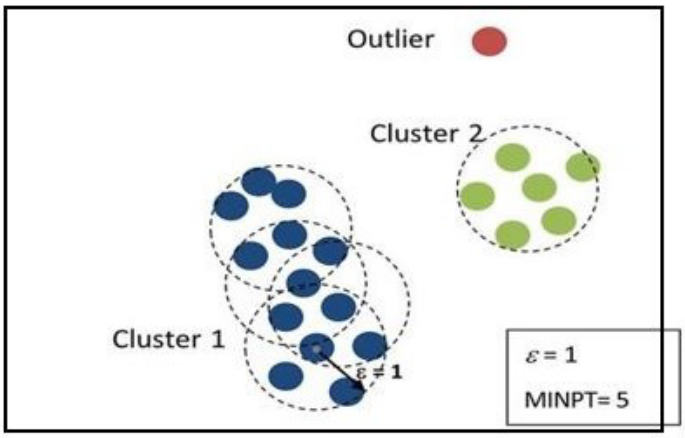


Figure 8: DBScan Cluster Technique . There are two clusters, colored as deep-blue and green where the outlier is marked as red circles. Also, eps are equal to 1 and the minimum number of points is 5.

5.2. *Confusion Matrix*

A confusion matrix is very useful to present the performance of any model [Fig. 9]. It consents to the visualization of the performance of the subsequent procedure. It is a table with two rows and two columns that calculates the number of true positives, true negatives, false positives, and false negatives [42]. For our calculation, we measured the accuracy (Table I) of artifact corrupted epochs for the unsupervised MLAs by the following equation:

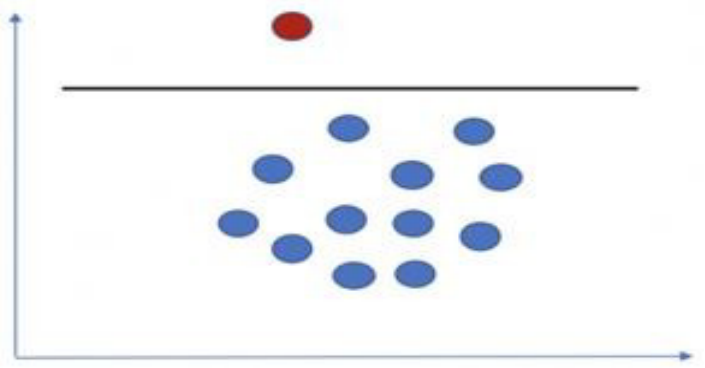


Figure 9: Confusion Matrix

$$\alpha = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

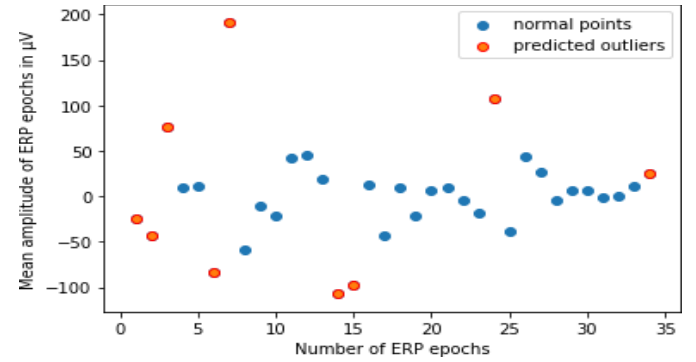
In the above equation, α = *Accuracy,*
TP = True Positive, TN = True Negative,
FP = False Positive, FN = False Negative

6. Results

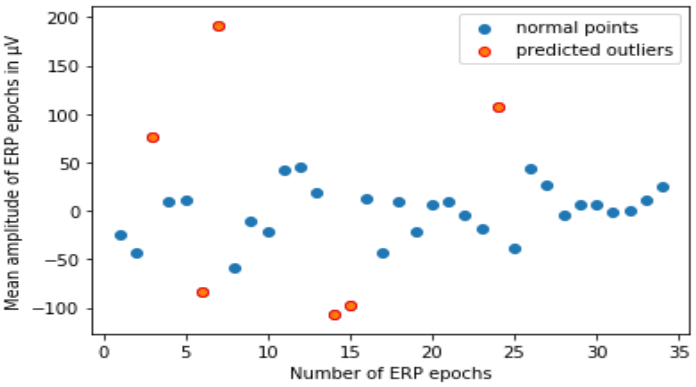
6.1. *Artifact corrupted Epoch Detection Accuracy*

In Fig. 10, the artifact corrupted ERP epochs, detected by Isolation Forest, DBScan, and LOF methods are shown for our test

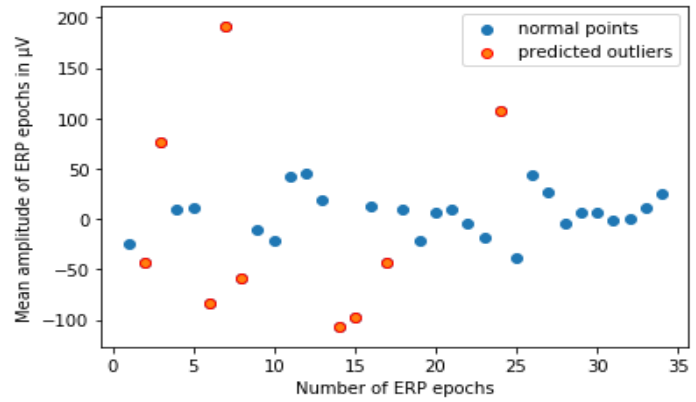
subject 1. The orange colors are indicating the artifact corrupted ERP epochs.



A: Isolation Forest



B: DBScan



C: LOF

Figure 10: Three unsupervised MLAs. Here, orange circles are artifact corrupted ERP epochs and blue circles are normal ERP epochs, for the test subject 1.

Table 1: Comparison of the detection of accuracy of the arti- fact corrupted epoch (in %) with eeglab

Dataset	Isolation Forest	DBScan	LOF
D1	76	91	94
D2	88	91	88
D3	94	97	88
D4	94	91	88
D5	74	88	85
D6	94	97	97
D7	94	97	94
D8	91	97	91
D9	95	88	82
D10	94	97	94
D11	91	97	94
D12	76	94	79
D13	82	94	85
D14	91	94	88
D15	62	94	74
D16	97	94	94
D17	91	94	82
D18	91	91	82
D19	88	91	88
D20	88	97	82
D21	79	88	85
D22	82	95	84
D23	81	92	86
Average	85.54%	93.43%	87.18%

Table-I shows the detection accuracy of the artifact corrupted ERP epoch of the auditory oddball paradigm . There, artifact corrupted ERP epoch detection accuracy of three unsupervised machine learning algorithms is compared with EEGLab's "Reject data epochs" method.

There are 23 test data-sets and named D1, D2, . . . , D23 (as described in section 3.1). Each data-set contains 50 epochs. At first, we detected the artifact corrupted epochs of the D1 data-set by EEGLab (named Data-set E1). After that, we applied 3 unsupervised MLAs (Isolation Forest, DBScan, and LOF) to detect the outlier and named the data-sets as M1, M2 M3, respectively. Then, by confusion matrix, we compared the accuracy of M1, M2 M3 with E1 for the detection of artifact corrupted epochs. We re-peated this procedure for all the remaining 23 data-sets. All of the comparative results are shown in Table-I. All the methods showed good detection accuracy and the DBScan achieved a maximum of 93.43%. The LOF method detection accuracy was 87.18% and the Isolation Forest performed with 85.54%.

Table 2: Comparison of visual and auditory oddball paradigm artifact corrupted epoch detection accuracy

	Isolation Forest	DBScan	LOF
Audi ERP	85.34%	93.43%	87.13%
Visu ERP	79.7%	90.15%	77.95%

Table 2, shows the comparison between the auditory oddball paradigm (Audi ERP) and visual oddball paradigm (Visu ERP) for the detection of artifact corrupted ERP epochs. In our previously published paper [1], there is detail about the procedure and results of the visual oddball paradigm. The dataset is a publicly available IRB-approved dataset by the University of California, Davis. The dataset contains 30 subjects' data. In that experiment, each subject saw five letters (A, B, C, D & E)randomly and one of these letters was assigned as a target letter, and its probability of appearance was 20%. The subject had to identify if the visual stimuli were a target or not, for every block of letters (Fig. 11). The duration of visual stimuli was 200 ms and the gap between each stimulus was 1200 – 1400 ms.

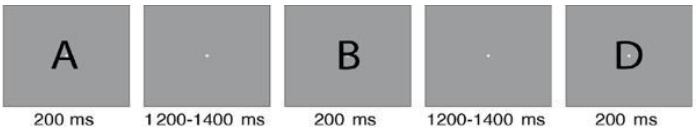


Figure 11: Visual Oddball task (Each subject saw five letters: A, B, C, D & E , randomly ,and one of these letters was assigned as a target letter, and its probability of appearance was 20%. The subject had to identify if the visual stimuli were a target or not, for every block of letters (Fig. 11). The duration of visual stimuli was 200 ms and the gap between each stimulus was 1200 – 1400 ms.);

For this visual oddball paradigm data-set, we detected three artifacts corrupted epochs . They are artifacts due to eye blinks, due to eye-movement, and due to body movement. Both EEGLab and MLAs are applied to detect artifact corrupted epochs and compared the accuracy as we did in this study. In Fig. 12, it is shown that, for both (visual and auditory) oddball paradigms, the DBScan per- formed with maximum accuracy while the other two methods are inconsistent with their positions.

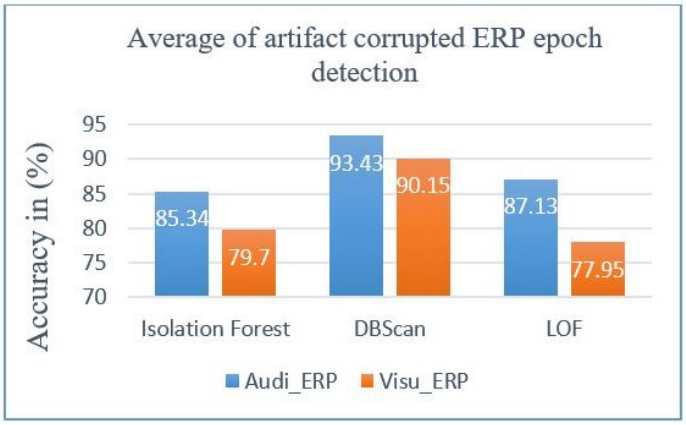


Figure 12: Comparison of artifact corrupted ERP epoch detection accuracy (on average) for both auditory oddball paradigm (Audi ERP) and visual oddball paradigm (Visu ERP) ERP Signal Analysis

6.2. ERP Parameter Analysis

At first, the mean of ERP amplitude, the peak value of ERP amplitude, and the latency of ERP peak were measured in a time window of 200ms to 600ms from the test data sets. Then we identified and removed the artifact corrupted ERP epochs (as described in Section V: A) from the test data sets. After that, we again measured the same values. After comparing the values, before and after removing the artifacts, we found a clear change in values for ERP mean and peak amplitude measures but there were no noticeable changes for the ERP peak latency measures. In Fig. 13 and Fig. 14, it is clearly shown that, for both mean and peak amplitudes, the values become lower after removing the artifact mixed epochs. For mean amplitudes it became 16.14 V from 35.4 V and, for peak amplitudes, it became 142.81 V from 223.9 V.

We also compared the values with our previously published visual oddball paradigm data set [1]. In Tables 3 & 4, values of auditory(Audi ERP) and visual(Visu ERP) ERP mean and peak amplitudes (in micro-volt) are given. In Table 3, mean ampl is the mean amplitude before artifact corrupted ERP epochs removal and mean AR means the mean amplitude after artifact corrupted ERP epochs removal. Same meanings are applicable to Table4. From these tables and Fig. 15 Fig. 16, it is clearly shown that for both mean and, peak amplitude, the value levels were higher before the artifact corrupted ERP epochs were removed. More specifically, there are sharper differences in the auditory oddball paradigm.

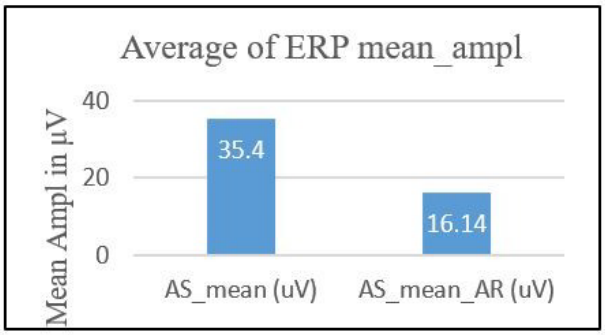


Figure 13: Comparison of the auditory mean amplitude of ERP data before and after (AR) the artifact corrupted EEG epochs are removed (uV: microvolt)

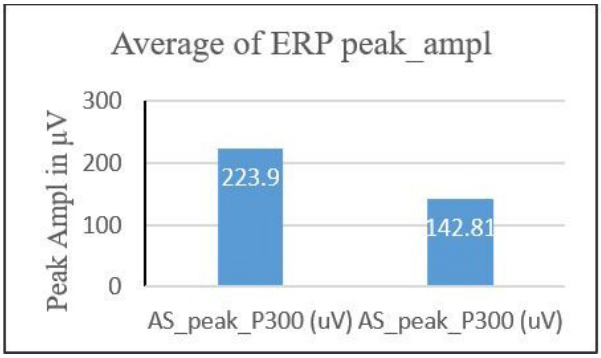


Figure 14: Comparison of the auditory peak amplitude of ERP data before and after (AR) the artifact corrupted EEG epochs are removed (uV: microvolt)

Table 3: Values of auditory and visual erp mean amplitudes in a time window of 200 to 600 ms before (mean) and after (mean ar) artifact corrupted epochs are removed

	Visu ERP	Audi ERP
mean(micro-volt)	9.28	35.4
mean AR(micro-volt)	8.59	6.14

Table 4: Values of auditory and visual erp peak amplitudes in a time window of 200 to 600 ms before (peak ampl) and after (peak ampl ar) artifact corrupted epochs are removed

	Visu ERP	Audi ERP
Peak ampl(micro-volt)	27.75	223.9
Peak ampl AR(micro-volt)	26.33	142.81

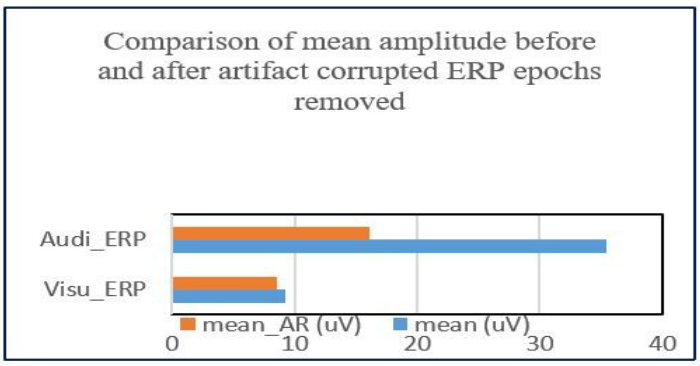


Figure 15: Comparison of the auditory ERP mean amplitude in a time window of 200 to 600 ms before and after the removal of artifact corrupted epochs (V: microvolt)

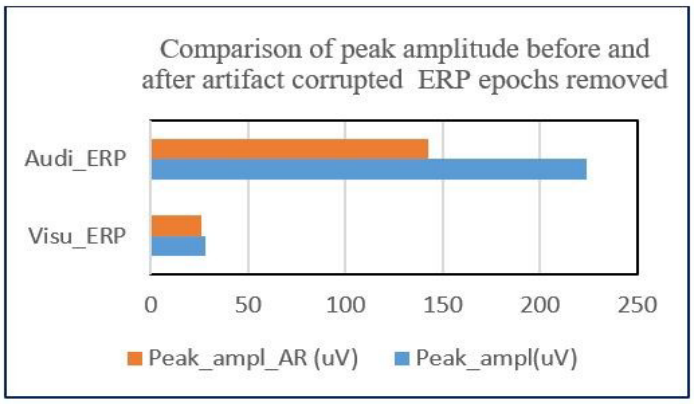


Figure 16: Comparison of the auditory ERP peak amplitude in a time window of 200 to 600 ms before and after the removal of artifact corrupted epochs (V: microvolt)

7. Conclusion

In this study, we detected ERP artifacts of the auditory odd-ball paradigm by unsupervised machine learning algorithms and compared the results with the visual oddball paradigm experiment which is our previously completed experiment. Our data were unlabeled and we found unsupervised machine learning algorithms are fairly efficient to distinguish the artifacts due to aye and body movement. Among the applied unsupervised machine-learning algorithms, the DBScan method performed with the most efficiency for distinguishing artifacts in ERPs for both audio and visual odd-ball paradigms. For the auditory ERP experiment, the accuracy is 93.43%, and for the visual ERP experiment, is 90.15%. In addition, the Isolation Forest and LOF method also showed good efficiency for the audio ERP experiment, 85.34%, and 87% respectively. On the other hand, they showed moderate efficiency for the visual ERP experiment, 79.7% and, 77.95% respectively (Table II). So, with the DBScan algorithm, we will have a cleaner ERP dataset in normal laboratory conditions with less complexity in data processing which may improve the quality of the EEG experiments.

It is very obvious that there is a huge change in amplitude levels of ERPs for the eye and body movement corrupted artifacts and normally amplitude levels lift up or down. From Tables 3 & 4, it is clear that ERP means and amplitude become lower after removing artifact corrupted epochs. In addition, there are specific differences between the audio ERP oddball paradigm experiment. On the other hand, no substantial changes were found for the peak latency, before and after artifact corrupted epoch removal.

For future research, there can be added more complexity to detect artifacts. Also, experiments can be designed to detect muscle artifacts which are one of the other common artifacts. Overall, this study may enable the use of ERPs as a strong bio-marker in EEG research in real-world experiments.

Conflict of Interest

There is not any conflict of interest.

References

[1] R. Akhter, F.R. Beyette, "Machine Learning Algorithms for Detection of Noisy/Artifact-Corrupted Epochs of Visual Oddball Paradigm ERP Data," in Proceedings - 2022 7th International Conference on Data Science and Machine Learning Applications, CDMA 2022, Institute of Electrical and Electronics Engineers Inc.: 169–174, 2022, doi:10.1109/CDMA54072.2022.00033.

[2] S. Luck, "An Introduction to the Event-Related Potential Technique, " Chapter 6, 2nd ed., MIT press, 2014.

[3] A.K.M.A. Siddique, R. Azim, A. Islam, "Analysis of the temperature effect on the P300 component by the left and right-hand movement," **16**(1), 45–49, Oct. 2022, doi:10.9790/1676-1601014549.

[4] P. Kadambi, J.A. Lovelace, F.R. Beyette, "Audio based brain computer interfacing for neurological assessment of fatigue," in International IEEE/EMBS Conference on Neural Engineering, NER, 77–80, 2013, doi:10.1109/NER.2013.6695875.

[5] M.T. Giovanetti, F.R. Beyette, "Physiological health assessment and hazard monitoring patch for firefighters," Midwest Symposium on Circuits and Systems, 2017-August, 1168–1171, 2017, doi:10.1109/MWSCAS.2017.8053136.

[6] Y.A. W de Kort L J M Schlangen Drir K C H J Smolders E Gecer, by Lotte Sap, The Influence of Light on the ERP P300 Waveform Sap, Lotte The Influence of Light on the ERP P300 Waveform THE EFFECT OF LIGHT ON THE ERP P300 WAVEFORM 1 Acknowledgement.

[7] R. Akhter, K. Lawal, M.T. Rahman, S.A. Mazumder, "Classification of Common and Uncommon Tones by P300 Feature Extraction and Identification of Accurate P300 Wave by Machine Learning Algorithms," IJACSA) International Journal of Advanced Computer Science and Applications, **11**(10), 2020.

[8] M.G. Asogbon, W. Samuel, X. Li, K. Dabbakuti, "Methods for removal of artifacts from EEG signal: A review You may also like A linearly extendible multi-artifact removal approach for improved upper extremity EEG-based motor imagery decoding Methods for removal of artifacts from EEG signal: A review 1,2 ShailajaKotte and," 12093, 2020, doi:10.1088/1742-6596/1706/1/012093.

[9] M.K. Islam, A. Rastegarnia, Z. Yang, "Methods for artifact detection and removal from scalp EEG: A review," Neurophysiologie Clinique/Clinical Neurophysiology, **46**(4–5), 287–305, 2016, doi:10.1016/j.neucli.2016.07.002.

[10] J.A. Urigüen, B. Garcia-Zapirain, "EEG artifact removal-state-of-the-art and guidelines," Journal of Neural Engineering, **12**(3), 2015, doi:10.1088/1741-2560/12/3/031001.

[11] R. Akhter, F. Ahmad, F.R. Beyette, "Automated Detection of ERP artifacts of auditory oddball paradigm by Unsupervised Machine Learning Algorithm," in 2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2022, Institute of Electrical and Electronics Engineers Inc., 2022, doi:10.1109/CIBCB55180.2022.9863055.

[12] D. Steyrl, G. Krausz, K. Koschutnig, al -, L. Fiedler, M. Wöstmann, Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T.H. Falk, J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," Journal of Neural Engineering, **16**(5), 051001, 2019, doi:10.1088/1741-2552/AB260C.

[13] Y. Guo, X. Jiang, L. Tao, L. Meng, C. Dai, X. Long, F. Wan, Y. Zhang, J. van Dijk, R.M. Aarts, W. Chen, C. Chen, "Epileptic Seizure Detection by Cascading Isolation Forest-Based Anomaly Screening and EasyEnsemble," IEEE Transactions on Neural Systems and Rehabilitation Engineering, **30**, 915–924, 2022, doi:10.1109/TNSRE.2022.3163503.

[14] Z. Lin, F. Wen, Y. Ding, Y. Xue, "Data-Driven Coherency Identification for Generators Based on Spectral Clustering," IEEE Transactions on Industrial Informatics, **14**(3), 1275–1285, 2018, doi:10.1109/TII.2017.2757842.

[15] M. Piorecký, J. Štrobl, V. Krajca, "Automatic EEG classification using density based algorithms DBSCAN and DENCLUE," Acta Polytechnica, **59**(5), 498–509, 2019, doi:10.14311/AP.2019.59.0498.

[16] N. Bigdely-Shamlo, K. Kreutz-Delgado, C. Kothe, S. Makeig, "EyeCatch: data-mining over half a million EEG independent components to construct a fully-automated eye-component detector," Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2013, 5845–5848, 2013, doi:10.1109/EMBC.2013.6610881.

[17] M. Agarwal, R. Sivakumar, "Blink: A Fully Automated Unsupervised Algorithm for Eye-Blink Detection in EEG Signals," 2019 57th Annual Allerton Conference on Communication, Control, and Computing, Allerton 2019, 1113–1121, 2019, doi:10.1109/ALLERTON.2019.8919795.

[18] S.S. Lee, K. Lee, G. Kang, "EEG Artifact Removal by Bayesian Deep Learning & ICA," in 2020 42nd Annual International Conference of the IEEE

Engineering in Medicine & Biology Society (EMBC), IEEE: 932–935, 2020, doi:10.1109/EMBC44109.2020.9175785.
[19] C.J.T. Kothe, "Artifact removal techniques with signal reconstruction," Google Patents. US Patent App. 14/895,440, 2016.
[20] A.K. Maddirala, K.C. Veluvolu, "Eye-blink artifact removal from single channel EEG with k-means and SSA," Scientific Reports, **11**(1), 2021, doi:10.1038/s41598-021-90437-7.
[21] S. Sadiya, T. Alhanai, M.M. Ghassemi, "Artifact detection and correction in EEG data: A review," International IEEE/EMBS Conference on Neural Engineering, NER, 2021-May, 495–498, 2021, doi:10.1109/NER49283.2021.9441341.
[22] P. Schembri, M. Pelc, J. Ma, "The Effect That Auditory Distractions Have on a Visual P300 Speller While Utilizing Low-Cost Off-the-Shelf Equipment," Computers 2020, **9**(3), 68, 2020, doi:10.3390/COMPUTERS9030068.
[23] LUCIO. di JASIO, "Graphics, touch, sound and usb, user interface design for embedded applications.," 2014.
[24] A. Delorme, S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," Journal of Neuroscience Methods, **134**(1), 9–21, 2004, doi:10.1016/j.jneumeth.2003.10.009.
[25] M. Fatourechi, A. Bashashati, R.K. Ward, G.E. Birch, "EMG and EOG artifacts in brain computer interface systems: A survey," Clinical Neurophysiology, **118**(3), 480–494, 2007, doi:10.1016/j.clinph.2006.10.019.
[26] A.K. Maddirala, K.C. Veluvolu, "Eye-blink artifact removal from single channel EEG with k-means and SSA," Scientific Reports 2021 11:1, **11**(1), 1–14, 2021, doi:10.1038/s41598-021-90437-7.
[27] D.W. Frank, R.B. Yee, J. Polich, "P3a from white noise," International Journal of Psychophysiology, **85**(2), 236–241, 2012, doi:10.1016/J.IJPSYCHO.2012.04.005.
[28] C.J. Ochoa, J. Polich, "P300 and blink instructions," Clinical Neurophysiology, 111(1), 93–98, 2000, doi:10.1016/S1388-2457(99)00209-6.
[29] R. Martínez-Cancino, A. Delorme, D. Truong, F. Artoni, K. Kreutz-Delgado, S. Sivagnanam, K. Yoshimoto, A. Majumdar, S. Makeig, "The open EEGLAB portal Interface: High-Performance computing with EEGLAB," NeuroImage, **224**, 116778, 2021, doi:10.1016/j.neuroimage.2020.116778.
[30] A. Delorme, R. Oostenveld, F. Tadel, A. Gramfort, S. Nagarajan, V. Litvak, "Editorial: From Raw MEG/EEG to Publication: How to Perform MEG/EEG Group Analysis With Free Academic Software," Frontiers in Neuroscience, 16, 359, 2022, doi:10.3389/FNINS.2022.854471/BIBTEX.
[31] C. Brunner, A. Delorme, S. Makeig, "Eeglab - an Open Source Matlab Toolbox for Electrophysiological Research," Biomedizinische Technik. Biomedical Engineering, 58 Suppl 1, 2013, doi:10.1515/BMT-2013-4182.
[32] J. Lopez-Calderon, S.J. Luck, "ERPLAB: an open-source toolbox for the analysis of event-related potentials," Frontiers in Human Neuroscience, **8**(1 APR), 2014, doi:10.3389/FNHUM.2014.00213.
[33] R. Martínez-Cancino, A. Delorme, D. Truong, F. Artoni, K. Kreutz-Delgado, S. Sivagnanam, K. Yoshimoto, A. Majumdar, S. Makeig, "The open EEGLAB portal Interface: High-Performance computing with EEGLAB," NeuroImage, 224, 2021, doi:10.1016/J.NEUROIMAGE.2020.116778.
[34] T. Jiang, J.L. Gradus, A.J. Rosellini, "Supervised Machine Learning: A Brief Primer," Behavior Therapy, **51**(5), 675–687, 2020, doi:10.1016/j.beth.2020.05.002.
[35] M.T. Rahman , R. Akhter, "Forecasting Stock Market Price Using Multiple Ma- chine Learning Technique, " Preprint, 2021.
[36] M.T. Rahman, R. Akhter, "Forecasting and Pattern Analysis of Dhaka Stock Market using LSTM and Prophet Algorithm," Preprint,2021.
[37] S. Sing, "Anomaly Detection Using Isolation Forest Algorithm," Analytics Vidhya Medium, 2020.
[38] M.Y. Pusadan, J.L. Buliali, R.V.H. Ginardi, "Optimum partition in flight route anomaly detection," Indonesian Journal of Electrical Engineering and Computer Science, **14**(3), 1315–1329, 2019, doi:10.11591/IJEECS.V14.I3.PP1315-1329.
[39] M.M. Breunig, H.-P. Kriegel, R.T. Ng, J. Sander, "LOF: Identifying Density-Based Local Outliers," 2000, doi:10.1145/335191.
[40] A. Mavuduru, "How to perform Anomaly Detection with the Isolation Forest Algorithm, Toward Data Science, "2021.
[41] E.E.M. Schubert, "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN," ACM Transactions on Database Systems (TODS), 1–21, 2017.
[42] F. Demir, "Deep autoencoder-based automated brain tumor detection from MRI data, " Elsevier: 317–351, 2022, doi:10.1016/B978-0-323-91197-9.00013-8.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 161-166 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Natural Tsunami Wave Amplitude Reduction by Straits – Seto Inland Sea

Mikhail Lavrentiev[*,1], Andrey Marchuk[1,2], Konstantin Oblaukhov[1], Mikhail Shadrin[1]

[1]*Institute of Automation and Electrometry SB RAS, Novosibirsk, 630090, Russia*

[2]*Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, 630090, Russia*

ARTICLE INFO

Article history:
Received: 24 July, 2022
Accepted: 01 November, 2022
Online: 08 December, 2022

Keywords:
Tsunami wave
Numerical modeling
Hardware code acceleration
Wave period
Passing through the strait

ABSTRACT

Seto Inland Sea is situated between Japanese islands Honshu, Kyushu, and Shikoku. It is separated from the ocean by the Bungo Channel, Kii Channel, and other narrow straits around Shikoku Island. The objective of the article is to draw the attention to the question of how well the coastal population and infrastructure of those locations are protected against a tsunami wave appearing in the area of Nankai through offshore Japan and consequently make informed decision about strengthening the protections if needed. This question is critical because strong underwater earthquakes are expected according to the major earthquake repeatability in this subduction zone. Hence, the influence of strait width and tsunami wave period to the wave amplitude change after passing a strait was studied through the application of numerical modelling. Both models of geometry of computation domain with the flat bottom and the real bathymetry of the area under study were used. Numerical modelling was based on nonlinear shallow water system, commonly used in tsunami related studies. All calculations were made at a personal computer equipped with the hardware code accelerator – FPGA (Field Programmable Gates Array) Calculator, which had been recently introduced by the authors. A series of computational experiments, both with model straits and in the water area around southern part of Japan, have shown that when a tsunami wave passes through the strait, its height is significantly reduced, providing better tsunami safety regulations of the population of the inner seas separated from the ocean by narrow straits. At the same time, longer tsunami waves retain a larger amplitude after passing through the strait compared to shorter ones. Result of the paper consists of numerical study of the influence of narrow straits on the maximal heights of tsunami wave. Qualitative corollaries between length of tsunami and reduction rate of the wave amplitude after passing a strait should help warning services to properly evaluate tsunami danger in water areas separated from tsunami source by a narrow strait.

1. Introduction

This paper is an extension of work originally presented at the conference OCEANS 2021: San Diego – Porto [1].

Tsunami waves, usually created by large underwater earthquakes, are propagated due to the gravity and grow substantially in amplitude as they approach the shore. One of the tasks of a tsunami warning service is to identify coastlines by the level of threat once a tsunami has been confirmed and act accordingly. Namely, it is critical to determine in which areas of the coastline high waves threatening local population and infrastructure are expected and hence evacuation must be announced, and in which areas tsunami damage may be minimal. Observations of tsunamis show that after a wave passes through a narrow strait, its height decreases significantly. Therefore, an area separated by a narrow strait from the open ocean is somewhat protected from tsunami impacts, unless the source of the tsunami is located in that area.

As it's suggested in literature after studies of historical records (see [2]), a major underwater earthquake (following by tsunami wave) is expected at Nankai through offshore Japan. This article evaluates the possible impact of such event to the Seto Inland Sea, bounded by the Honshu, Kyushu, and Shikoku islands of Japan. The bottom relief of the studied water area is given in Figure 1. Tsunami wave can arrive in this water area only through the Bungo

[*]Corresponding Author: Mikhail Lavrentiev, mmlavrentiev@gmail.com

https://dx.doi.org/10.25046/aj070616

Channel in the south and Kii Channel in the north. These channels are narrow enough to reduce valuably of the amplitude of tsunami wave. To study numerically this phenomenon, the wave parameters related to the initial seabed displacement at tsunami source of the event, which took place in this area nearly 300 years ago was used but with a rather lower displacement. Such approximation, shown in Figure 2, has been proposed in [2] on the basis of available historical records.

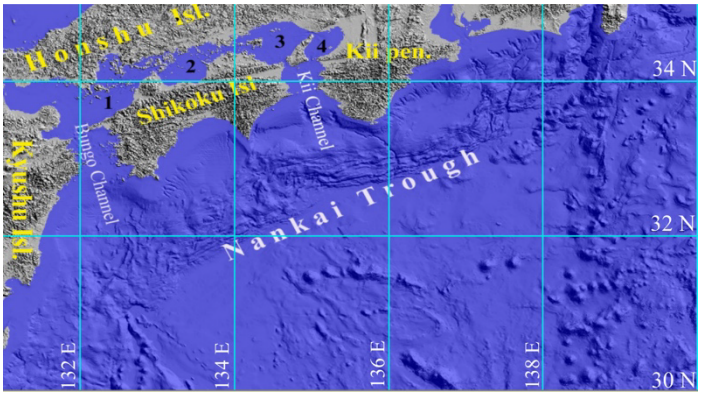


Figure 1: Visualization of the digital bathymetry for the water area under study. Seto Inland Sea surrounded by Shikoku, Honshu and Kyushu Islands is divided in sectors 1 - 4

The goal of the numerical study was to understand the influence of the size of tsunami source to amplitude reduction rate while the wave is passing through a narrow channel. Model tsunami sources considered had time periods from 5 to 15 minutes. Such numbers correspond to sources of an elliptic shape having short axis variation interval between 60 km and 200 km. Note that even trans-oceanic waves with that long period can arrive in the region under the study.

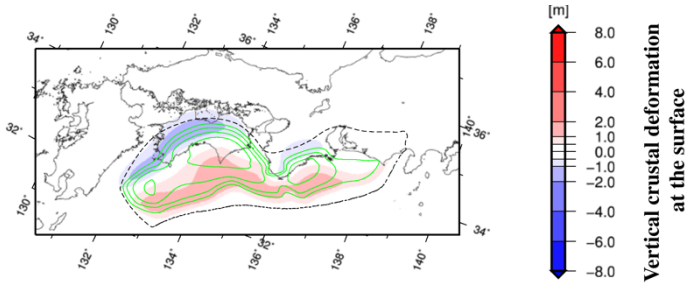


Figure 2: The expected vertical bottom displacement for M_W=8.66 earthquake in Nankai subduction zone [2]

Rest of the paper is arranged as follows. The governing equations (along with the initial shape of the wave) are first introduced and the auxiliary numerical tests concerning straits of model shape are described. In particular, the influence of the strait width and wavelength on the wave amplitude is evaluated. The detected impacts are then confirmed for the real digital bathymetry of the Nankai Through and Seto Inland Sea. The obtained results are then discussed.

2. Design of Numerical Experiments

2.1. Model equations

For numerical modeling of tsunami wave propagation, a nonlinear mathematical shallow water system was chosen without considering external forces (bottom friction, Coriolis force, etc.), except the gravity [3]:

$$\frac{\partial H}{\partial t}+\frac{\partial (uH)}{\partial x}+\frac{\partial (vH)}{\partial y}=0 \quad (1)$$

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}+g\frac{\partial H}{\partial x}=g\frac{\partial D}{\partial x} \quad (2)$$

$$\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}+g\frac{\partial H}{\partial y}=g\frac{\partial D}{\partial y} \quad (3)$$

where $H(x,y,t)=\eta(x,y,t)+D(x,y)$ is the total height of water column, η being the sea surface disturbance (wave height), $D(x,y)$ – depth (which is supposed to be known at all grid points), $u(x,y,t)$ and $v(x,y,t)$ are components of water flow velocity vector, g - acceleration of the gravity. Note that this approximation is widely used for tsunami simulation [4, 5].

2.2. Model Assumptions

The shallow water model (1)-(3) used for numerical modeling of tsunami propagation assumes that the wave length is significantly (at least an order of magnitude) greater than the depth of the water area where the wave is traveling. It follows that the velocity c of tsunami propagation does not depend on its amplitude and is determined only by the water depth by the Lagrange formula [3]: $c=\sqrt{gD}$. Due to the constancy of the tsunami wave period during propagation, its length in the area with constant depth also remains constant. Therefore, in such areas, reasoning in terms of "wave period" and "wave length" is equivalent. In areas with variable depth, such substitution of concepts is incorrect.

2.3. Instrument for Numerical study

In the Method Of Splitting Tsunami (MOST) software package [4,6] (official tsunami modeling instrument of the USA Tsunami Warning Centers) a numerical calculation of shallow water system is based on the splitting method. The system is solved numerically in two steps. During each step the changing of variables only along one space direction is a subject of the study. The essence of the method is that the solution is sought as the result of the variables behavior in the course of the sequential solution of two auxiliary one-dimensional systems. In the present paper a parallel implementation of McCormack's scheme, which has the second order of approximation [7], is used. As the earlier numerical experiments [8] have shown, the proposed software allows finding solutions of the shallow water equation system with the required accuracy. Due to the hardware acceleration, high performance is achieved when using a personal computer [8].

2.4. Auxiliary Numerical Tests

In order to study the process of passage of a long wave through narrow straits and to identify the degree of influence of the strait width and the wave period on the wave parameters beyond the strait, a series of model computational experiments were conducted. In them, the channel (strait) was represented by straight shorelines converging at an angle of 90° until the width of the strait becomes equal to the given “width”. After this narrowest point of the strait ("bottle neck"), the shorelines begin to diverge at the same angle 90° (Figure 3).

Computation domain, given in Figure 3, represents a 1000x1000 km square with a constant depth of 1000 m. The two "water reservoirs" having the shape of triangles are connected by a narrow strait with a width of 40 km (as shown in Figure 3) and also 20 km for alternative numerical tests. The dry land, the boundaries of which reflect tsunami waves, is colored in green (Figure 3). The MacCormack difference scheme [7] was realized on FPGA PC board [8] for numerical modeling of tsunami wave propagation. The use of this particular algorithm is not a determining factor in this study. For calculations, any other correct method of solving the shallow water equations could be used, for example, the basic one in MOST software package [4,6].

The initial tsunami wave was generated using boundary conditions on the left boundary of the region, where during the wave period *tp* the vertical displacement of the water surface η and the horizontal velocity components of the water flow were set by the following formulae:

$$\eta = \frac{\eta_0}{2}\left(1+\cos\left(\frac{t\cdot 2\pi}{tp}\right)\right),\ u = \eta\sqrt{\frac{g}{D}},\ v=0 \quad (4)$$

here the variable t belongs to the interval $0<t<tp$. Period of the tsunami wave is characterized by the quantity tp, and η_0 sets the initial wave height of its crest. Such a way for wave generation was developed and tested in [9]. As a result, a tsunami wave is generated at the left boundary, which has a height $\eta_0=1$ m and a period tp, moving to the right of this boundary (Figure 3). In numerical calculations, the wave period was set to 300, 600 and 900 seconds. The simulation result is the distribution of tsunami height maxima over the entire computational domain. Figure 3 shows the results of calculations of the wave with a period of 300 seconds through the strait with a width of 40 km at the narrowest point. The height of isolines in Figure 3 (and all other figures, showing results for model straits) are drawn with interval of 5 cm..

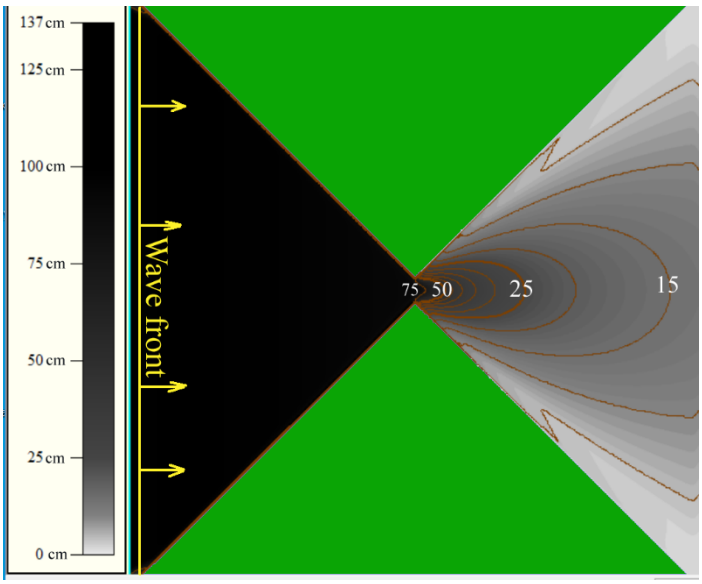


Figure 3: Numerically obtained distribution of the tsunami height maximums. Wave period is 300 sec., strait has 40 km width. The entire area is 1000x1000 km square. Interval (in height) between height isolines is equal to 5 cm

Figure 3 shows that the tsunami height, initially equal to 100 cm, begins to decrease quite rapidly just after the passage of the "bottle neck", which has a width of 40 km. At the same time, the radiation of wave energy in the right half of the calculation region has a narrow directivity (see Figure 3). To study the effect of wavelength (period), the authors conducted two more computational experiments at the same strait for the passage of the tsunami with a period of 600 and 900 seconds. Figure 4 shows the distribution of maximum heights resulting from the passage of a long wave with a period of 600 sec through the same strait.

Another calculation simulating tsunami passage with the period of 900 sec through the same model strait was also conducted. The results are shown in Figure 5.

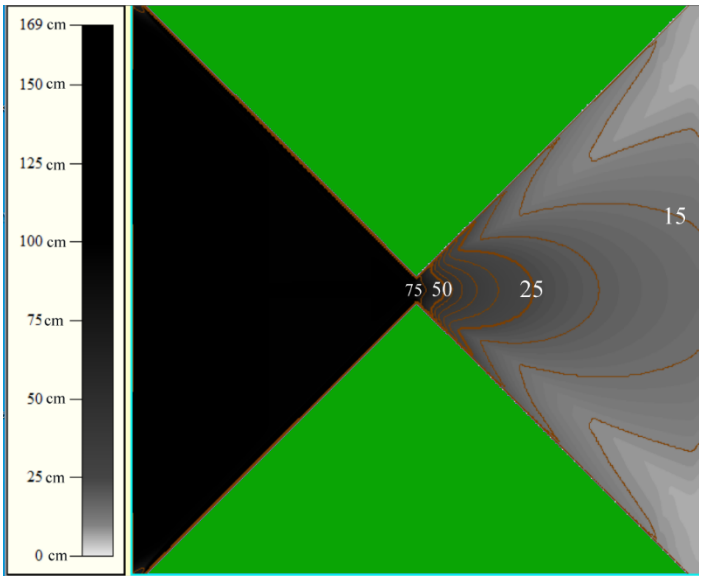


Figure 4: Numerically obtained distribution of the tsunami height maximums. Wave period is 600 sec., strait has 40 km width

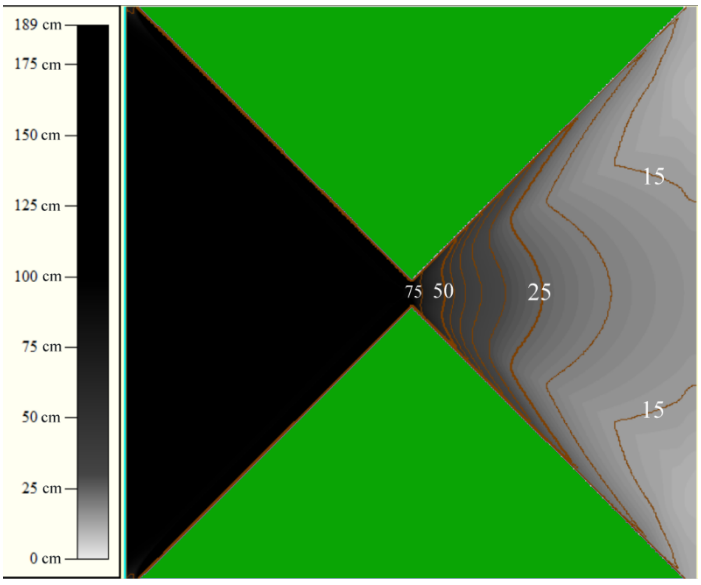


Figure 5: Numerically obtained distribution of the tsunami height maximums. Wave period is 900 sec., strait has 40 km width

Figures 3-5 show that, as the length (period) of the tsunami wave passing through the 40-km wide strait increases, the radiation sector of the wave energy expands in the right subarea (after the wave has passed the "bottle neck"). In case of a longer wave, a

slower decrease in the tsunami amplitude is observed with distance from the narrowest point of the strait. In particular, for a wave with a period of 300 seconds, the tsunami height in the center of the right boundary of the computational domain was about 14 cm. For the wave with the period of 600 sec - 16 cm. And for the longest wave (with a period of 900 sec) - 17 cm (Figures 3 – 5). Note that the increased amplitude of the tsunami at the boundaries of the strait is a consequence of the wave reflection from the shore.

During the next series of computational experiments, the width of the narrowest part of the strait was reduced to 20 km, i.e., half as much as in the first series. At the same time, the angle of convergence of the shorelines remained equal to 90°. The parameters and direction of propagation of the generated wave remained the same. Figure 6 shows the distribution of tsunami height maxima with a period of 300 s as it passed through a 20 km wide strait.

Comparison of Figures 3 and 6 shows that the tsunami height decreases significantly faster after the wave passage of the 20-km wide bottle neck compared to the case of the 40-km wide strait.

Figures 7 and 8 show the distribution of maximum tsunami heights when simulating tsunami waves with periods of 600 and 900 sec for the 20 km wide strait.

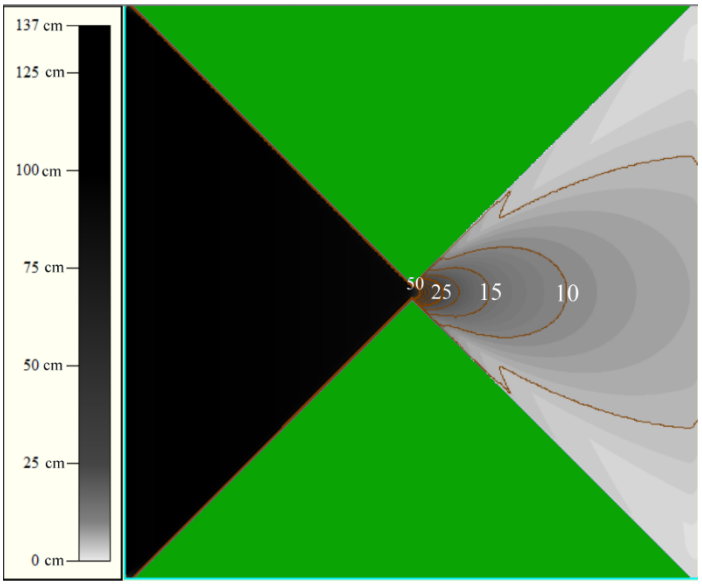


Figure 6: Numerically obtained distribution of the tsunami height maximums. Wave period is 300 sec., strait has 20 km width

Comparing the series of Figures 3-5 with Figures 6-8, when simulating the passage of the wave with the period of 300, 600 and 900 sec through twice as narrow 20 km wide strait, the tsunami heights in the right subarea were expectedly reduced in comparison with the 40 km wide strait. For example, in case of the 40-km wide strait, the height of the wave with the period of 600 seconds in the center of the right boundary of the computational domain equals to 16 cm (Figure 4), while for the 20-km wide strait the height of the initial wave in the same point of the right boundary wave height is detected at 8.5 cm (Figure 7). That is, the height of the wave is almost half as much in the case of a narrower strait. However, when the wave period was increased to 900 sec, this difference decreased and was 17 cm versus 14 cm (Figures 5 and 8).

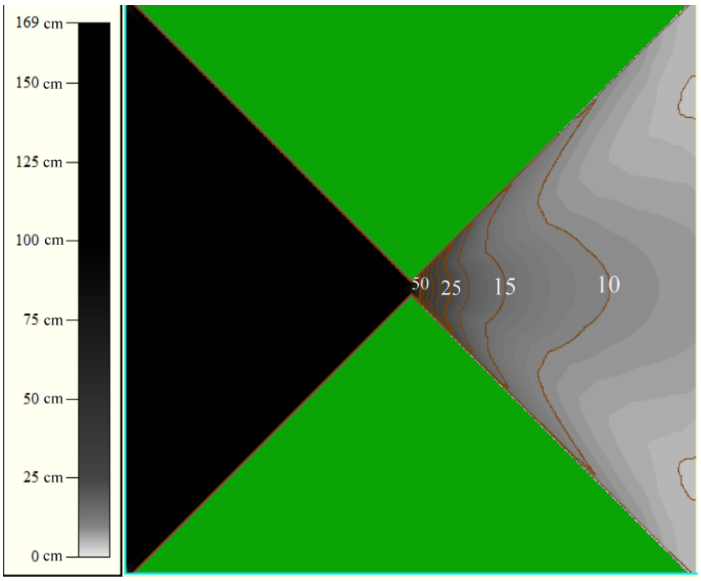


Figure 7: Numerically obtained distribution of the tsunami height maximums. Wave period is 600 sec., strait has 20 km width

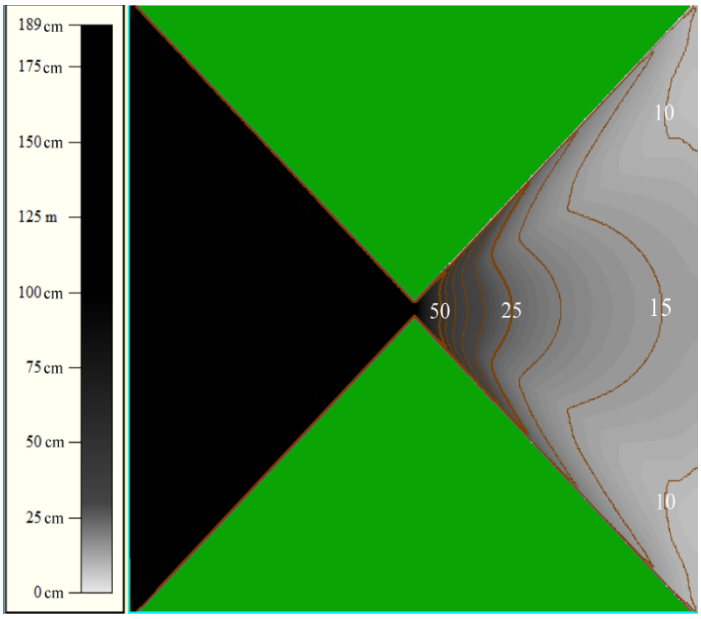


Figure 8: Numerically obtained distribution of the tsunami height maximums. Wave period is 900 sec., strait has 20 km width

3. Numerical Results

3.1. Computational Domain

To investigate the tsunami wave propagation through the straits around Shikoku Island using JODC [10] bathymetric data, a grid bathymetry was prepared, covering the area around the Nankai Trough and Seto Inland Sea from 131° to 140° East Longitude and from 30° up to 35° North Latitude. The depth value array is 3,000 × 2,496 nodes. The spatial grid step in the longitude direction was 0.003 arc degree (282 m), and in the South-North direction the grid step was chosen to be 0.002 arc degree (223 m). Visualization of the computational domain's geography and bathymetry is presented in Figure 1.

3.2. *The influence of the tsunami wave period on the wave height in the Seto Inland Sea.*

Several numerical tests were conducted to simulate tsunami wave propagation through the narrow strait using real bathymetry. Geographic area around the Shikoku Island (Southward of Japan) was considered. Visualization of this computational domain is given in Figure 1.

The tsunami wave entering through the Bungo Channel and Kii Channel into the Seto Inland Sea was generated using the same boundary conditions (2) on the lower boundary of the computational domain as in the calculations of the wave passing through the straits with the model configuration (cf. Figure 3). Values of flow velocity components must be swapped in the formulae (2).

Similar to numerical tests at the bathymetry model, tsunami waves having periods of 300, 600 and 900 seconds were considered. Initial wave height was 1 m. Numerically obtained tsunami heights maximums around Shikoku Island in case of 600 sec. period wave are shown in Figure 9. The wave was initiated at the lower boundary.

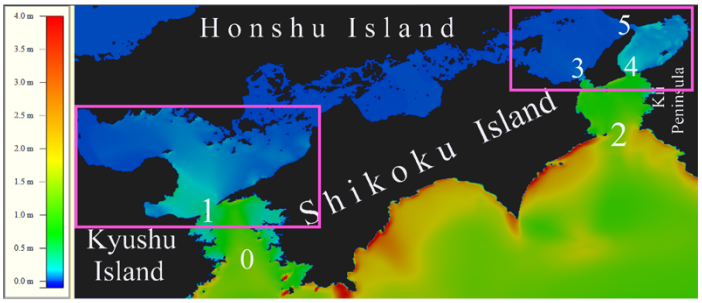


Figure 9: Location and numbering of straits around Shikoku Island. Tsunami wave maxima all around computational area resulted by tsunami modeling of 600 sec. period wave are visualized according to the color legend in the left side of figure

To show the existing narrow straits around Shikoku Island in the considered water area, the white numbers in Figure 9 are used. So, the Bungo Chanel is marked with the number 0. According to numerical experiments, the Kii Channel (having number 2 in Figure 9) has practically no impact on tsunami wave amplitude because it is wide enough. However, the “bottle neck” of the Bungo Channel (indicated by number 1) significantly reduces tsunami height. Numerical results for subareas outlined by pink rectangles (Figure 9) in more detail will be shown later. They contain the straits 1,3,4 and 5.

A series of computational experiments with another tsunami period were carried out in order to study the wavelength influence on the tsunami protection ability of the straits. The influence of the Bungo Chanel on tsunami amplitude is presented in Figure 10 for the 300 sec. period wave.

Numerical results with the same wave period of 300 sec. in the area to the east of Shikoku Island are given in Figure 11. This area, bounded by the pink rectangle in Figure 9, contains the straits number 3-5.

Results of numerical simulation presented in Figures 10 and 11 show high influence of the narrow straits in the Seto Inland Sea on the maximal heights of a tsunami wave. So, the strait 3 (see Figure 9) reduces the wave amplitude by 10 times, while the Straits 1 and 4 – by nearly 4 times. After passing the strait, the wave amplitude is going down because of diffraction and refraction. As a result, the incoming wave of 1 m amplitude becomes almost negligible (with the height of 1-2 cm) to the north of Shikoku Island behind the straits, which is shown in Figures 10 and 11.

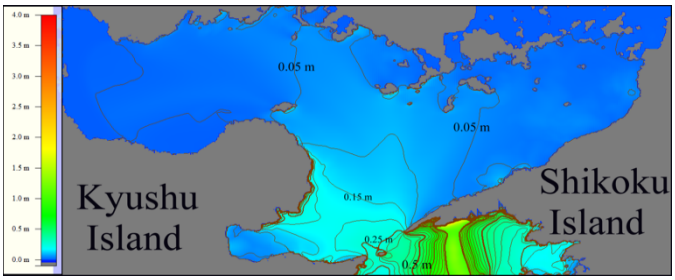


Figure 10: The maximum tsunami heights distribution of the wave having 300 sec period when it passes the Bungo Channel which is indicated by number 1 in Figure 9. Isolines of tsunami height maxima are drawn with an interval of 5 cm

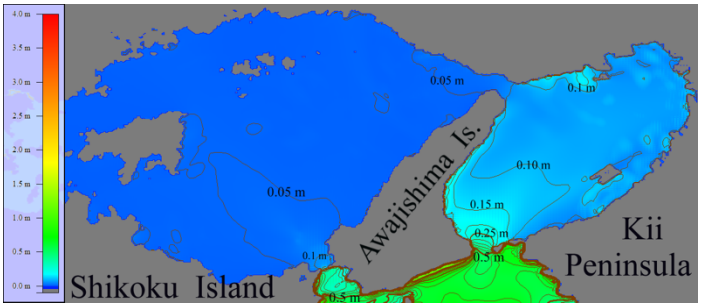


Figure 11: The maximum tsunami heights distribution of the wave having 300 sec period when it passes wide Kii Channel (indicated by number 2 in Figure 9) and narrow straits 3,4 and 5 in the eastern part of Seto Inland Sea

To analyze the dependence of tsunami wave amplitude reduction on the wave period, additional numerical experiments for the Bungo Channel were conducted. The obtained results for the wave periods of 600 and 900 sec. are given in Figures 12 and 13.

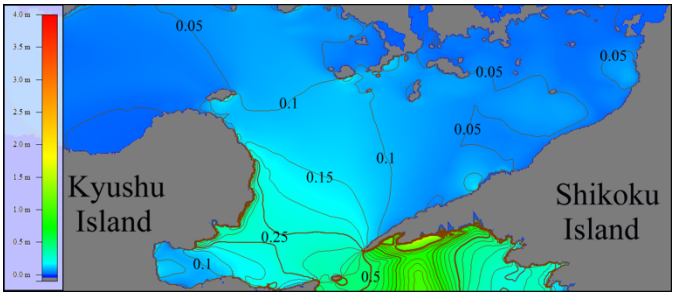


Figure 12: Distribution of tsunami height maxima after the wave having period of 600 sec passed through Bungo Channel. Isolines are drawn with an interval of 5 cm

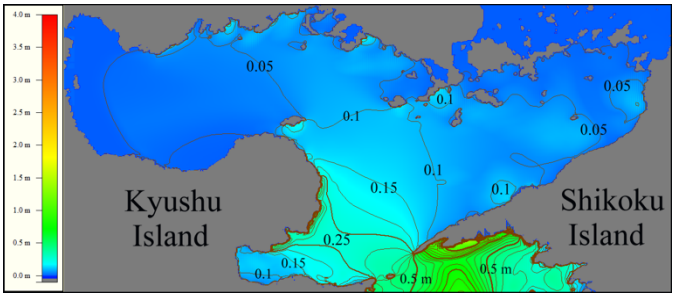


Figure 13: Distribution of tsunami height maxima after the wave having period of 900 sec passed through Bungo Channel

The result of the analysis of the tsunami wave behavior at real bathymetry, shown in Figures 10, 12, 13, is similar to that of the model of numerical experiments, given in Figures 3-8. Namely, the longer is the incoming wave, the larger is the tsunami wave amplitude in the inland sea after the Bungu Channel. The same phenomena are observed for the strait to the right of Shikoku Island studying wave periods of 300, 600 and 900 seconds. Visual results are given in Figures 11, 14 and 15.

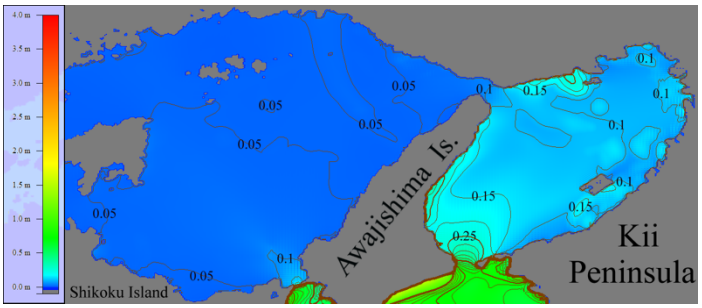


Figure 14: The wave height maxima distribution around Awajishima Island calculated in the course of numerical modeling of tsunami having period 600 sec

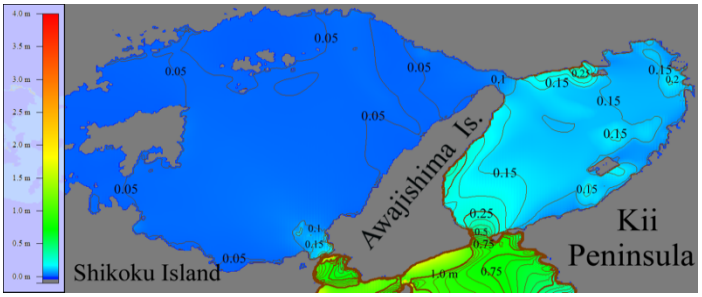


Figure 15: Isolines of the wave height maxima distribution around Awajishima Island calculated in the course of numerical modeling of tsunami having period 900 sec

Considering in detail the water area with the straits number 3, 4 and 5 (Figure 9) one concludes that the rather narrow channels provide better protection against tsunami waves. Indeed, the tsunami height decreases by about three times passing the strait number 4, and by about eight times passing through the strait number 3, Figures 11-15. The inner sea itself, being characterized by a comparatively low depth, reduces tsunami wave amplitude due to diffraction and refraction.

4. Discussion

Numerical calculations for model strait configurations and real straits in the southern part of Japan around Shikoku Island revealed certain patterns of tsunami wave passage through narrow straits. It was found that after passing the narrowest point of the strait, the wave height decreases rapidly in the widening water areas behind the strait due to wave refraction and diffraction. As a result, already in the center of the area, separated from the ocean by the narrow strait, the tsunami height is at an order of magnitude lower than before the wave enters the strait. It is clear from Figures 10-15 that with tsunami heights in front of the narrow straits of the order of 0.5 m to 1 m, the wave heights in the Seto Inland Sea on average are limited by 10 - 15 cm. When the length (period) of the generated wave increases, the maximum heights in the inland waters slightly increase. According to the results of this study, the authors conclude that the population and infrastructure of the Seto Inland Sea coast are sufficiently protected from the impacts of tsunamis generated in the Nankai Trough region. Only in the event of a catastrophic scale with tsunami heights at the entrance to the Straits exceeding 10 m, the population of the coasts of inland waters around Shikoku Island may be severally affected by tsunamis.

5. Conclusion

As a result of a series of numerical calculations of the passage of the tsunami wave with various period through the straits of model configuration, the degree of amplitude reduction after the wave passed the narrowest part of the strait was found. In the same series of numerical experiments, the dependence of the height distribution behind the strait on its width in its narrowest place was studied. The upper limits of tsunami heights in different parts of Seto Inland Sea water area depend on length (period) of the wave coming from Pacific Ocean through the straits around Shikoku Island. The longer the wave period, the higher its amplitude inside this water area. Since narrow straits naturally significantly reduce the height of waves passing through them, tsunamis with a height of more than 1 m can be observed inside Seto Inland Sea only in case of catastrophic events with an extremely large source. Further research on this issue will include model calculations of tsunami behavior in alternative configurations of the straits. Computational experiments simulating the passage of tsunamis through real straits in other places of the Pacific coast will be conducted as well.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

This research was carried out under state contract with IAE SB RAS (121041800012-8) and with ICMMG SB RAS (0315-2019-0004).

References

[1] M. Lavrentiev, An. Marchuk, K. Oblaukhov, M. Shadrin, "The wavelength influence on tsunami height in Seto Inland Sea," in Global Oceans 2021: San Diego - Porto, In Person & Virtual: September 20-23, 2021, Town and County San Diego, doi: 10.23919/OCEANS44145.2021.9705819.

[2] M. Hyodo, T. Hori, K. Ando, and T. Baba, "The possibility of deeper or shallower extent of the source area of Nankai Trough earthquakes based on the 1707 Hoei tsunami heights along the Pacific and Seto Inland Sea coasts, southwest Japan", Earth, Planets and Space, **66**(123), 2014, http://www.earth-planets-space.com/content/66/1/123.

[3] J.J. Stoker, Water Waves. The Mathematical Theory with Applications, Interscience publishers: New York, NY, USA, 1957.

[4] E. Gica, M. Spillane, V. Titov, C. Chamberlin, J. Newman, "Development of the forecast propagation database for NOAA's short-term inundation forecast for tsunamis (SIFT)," *NOAA Technical Memorandum*, 2008. URL: https://nctr.pmel.noaa.gov/Pdf/brochures/sift_Brochure.pdf (accessed on 19 December 2021).

[5] Z. Kowalik, T.S. Murty, Numerical modeling of ocean dynamics, World Scientific, 1993.

[6] V.V. Titov, F.I. Gonzalez, "Implementation and testing of the method of splitting tsunami (MOST) model", NOAA Technical Memorandum ERL PMEL-112, USA, 1977.

[7] R.W. MacCormack, A.J. Paullay, "Computational Efficiency Achieved by Time Splitting of Finite–Difference Operators," in 10th Aerospace Sciences Meeting, San Diego, CA, USA, 17–19 January 1972; AIAA Paper; pp. 72–154.

[8] M.M. Lavrentiev, A.A. Romanenko, K.K. Oblaukhov, An.G. Marchuk, K.F. Lysakov, M.Yu. Shadrin, "FPGA Based Solution for Fast Tsunami Wave Propagation Modeling," in 27th (2017) Int Ocean/Polar Eng Conf, San Francisco, CA, USA, June 25-30, 2017, 924-29.

[9] K. Hayashi, A. Marchuk, A. Vazhenin, "Generating Boundary Conditions for the Calculation of Tsunami Propagation on Nested Grids," Numer. Analys. Appl. 11, 256–267, 2018, https://doi.org/10.1134/S1995423918030060.

[10] J-DOSS|JODC [Electronic resource]. Available online: URL https://jdoss1.jodc.go.jp/vpage/depth500_file.html (accessed 18.07.2022).

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 5, 167-177 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Blockchain Applications in Suning and PingAn

Cong Qi*, Yue Lei, Yuejun Cai

The Hong Kong Polytechnic University, Department of Management and Marketing, Hong Kong, China

ARTICLE INFO

Article history:
Received: 27 May, 2022
Accepted: 15 October, 2022
Online: 31 October, 2022

Keywords:
Blockchain applications
Suning
PingAn

ABSTRACT

The emergence of blockchain technology has facilitated the digital transformation of many businesses and thereby increased the competitiveness of China in the global market. As a cutting-edge technology, Blockchain has significantly influenced the practices across all business sectors. Focusing on the retail and insurance sectors, this paper analyzes the blockchain technology applications in two companies - Suning and PingAn. Success factors in blockchain implementation were discussed by using TOE framework. Similarities and differences in the implementation and development processes are compared. Features from the two corresponding industries - retail and insurance are also summarized. The major results revealed that retail and insurance industries are two of the most important application sectors of blockchain technology, and Suning and PingAn are both pioneers in blockchain technology development and implementation in their specific sectors. However, there are still distances between Chinese blockchain providers and world leading providers of blockchain, even though China has the largest number of blockchain patents in the world. The research results provide meaningful insights and practical implications to the blockchain application fields.

1. Introduction

This paper is an extension of work [1] initially presented at the conference of Artificial Intelligence and Blockchain Technology (2021). Starting from a boom in the crypto-asset speculative movement in 2017, blockchain technology has kept an ascending trend in recent years [2]. Due to the high fever of Bitcoin, blockchain as a novel technology is also gradually gaining popularity in various industries. Blockchain consists of a peer-to-peer (P2P) network, a private key cryptography, and a protocol to make transactions secure [3]. Blockchain shares some unique characters, such as decentralization, persistency, anonymity and suitability [4], and was widely applied in multiple prevalent areas [5].

Although blockchain technology has been studied in various industries, its primary applications exist in the finance industry. Due to the use of blockchain, the labor costs are significantly reduced, and information system security is substantially increased [6]. Retail industry is another industry with a higher penetration rate of blockchain, it can be used to trace veraciously transactional records [7]. Though industry papers have discussed and listed the implementation facts of both industries, little research has been done to systematically summarize the similarities and differences between the two industries, specifically with exemplary cases. With the above, this study examines the blockchain applications in these two industries and thoroughly analyzes two exemplars - Suning holdings group and PingAn insurance group. The Technology-Organization-Environment framework is also used to summarize the success factors of blockchain implementation in both companies. In sum, our paper aims to address the following research questions:

1. How do Suning holdings group and PingAn insurance adopt and provide blockchain technology in their businesses?
2. What are the similarities and differences between both companies regarding to the development and implementation of blockchain technologies? What are the common factors leading to blockchain application success?
3. How was blockchain technology developed in retail and insurance industries on a large scale?

2. Blockchain Technology

Blockchain is a system for recording information, and it is essentially a distributed ledger technology. The decentralized and immutable nature makes it difficult to disrupt. A blockchain is a chained database containing fixed-length blocks in which every

*Corresponding Author: Cong Qi, Email: cong.qi@polyu.edu.hk

https://dx.doi.org/10.25046/aj070518

transaction record is included. Newly generated insuppressible transactions need to be verified before being inserted into the block. Users manage the data stored in the public blockchain autonomously through a P2P network and a distributed timestamp server, they also have free access to the transactions between two parties permanently recorded in the blockchain [8].

From the perspective of information technology, a blockchain is a distributed database in the form of multiple nodes, and a complete copy is available on each node. If more than 50% of the nodes reach consensus, the chain is entirely trustworthy [9]. In addition to the transaction record and timestamp, each block contains a hash value of the previous block and a new random number that can be used to verify the hash. The hash value exists to prevent fraud caused by changes to the content stored in the blockchain, and it follows each block to change [10]. Blockchain also determines whether a block contains a valid transaction. It uses the proof of work algorithm, which empowers data stored in the blockchain to be unaffected by the collapse of one node and thus maintains the authenticity of the data. On an overall basis, such a transparent system helps record the detailed sequence of events [6].

From the perspective of business transactions, blockchain technology offers a shared ledger for all companies involved in transaction. These new ledgers can hardly result in data loss, change or deletion, and can exclude the presence of a trusted third-party company as in traditional transactions [11]. In this study, we discuss the applications of blockchain technology in the retail and insurance industries. Suning holdings group and PingAn insurance are selected in this paper as typical cases due to their pioneer positions in blockchain technology adoption and implementation. Furthermore, we extend our study and discuss the general phenomena of applying blockchain technology in retail and insurance industries. Last, the comparison of successful factors, similarities and differences between the two companies are presented, and the features of the two industries are summarized.

3. Blockchain in Suning holdings group

Suning holdings group (hereafter Suning) has been ranked as the second-largest civilian-run enterprise in mainland China. As a large ecological, platform-based company, Suning's industry chain involves retail, finance, logistics, Internet of things (IoTs), cultural creation, and sports. With an online and offline customer base of over 600 million, Suning is well-positioned to use big data to build consensus collaboration networks. A large amount of data provided blockchain technology a suitable environment. This amount of data also contributed to establish a blockchain lab by Suning Financial Research Institute in 2017, and thereupon, Suning integrated the blockchain technology into the group's business [12]. After that, Suning launched the domestic credit information transmission system, blockchain blocklist sharing platform, "blockchain + IoTs" financing platform for cars, blockchain anti-counterfeiting technology, and blockchain asset securitization service system between 2017 and 2018. After a series of practices of blockchain, Suning released the Suning blockchain white paper in July 2018. The white paper elaborated the development path and application areas of blockchain in Suning and announced Suning's prediction of the future development of blockchain [12].

A year later, Suning launched Blockchain as a Service (BaaS), a cloud-based infrastructure platform to provide developers and users with an ecosystem for creating and managing blockchain applications. The BaaS platform is also committed to provide a series of blockchain financial solutions such as data sharing, IoTs, supply chain finance, and payment settlement. In 2020, Suning launched an external empowerment plan to enhance digital and intelligent financial services, which has accelerated the research and development of blockchain technologies and expanded the blockchain application scenarios to its major partners. It is worth mentioning that Suning's financial blockchain technology has won various industry awards, including being selected for the 2020 FAT Blockchain fintech leading 100 list. These show that Suning's blockchain applications are forward-looking in the industry [13].

3.1. Suning's blockchain in Retail and E-Commerce

Suning's blockchain application in the retail industry originated from the concept of smart retailing [14]. Compared to traditional retailing, smart retailing tends to predict consumers' preferences and customize services by using big data. It involves drawing a user portrait, building a membership ecosystem, and supplying products according to market demand.

3.1.1. User portraits and member ecology

Consumers inevitably reveal their privacy when engaging in consumer behavior. To prevent user information from being leaked to third parties intentionally or unintentionally, Suning introduced blockchain technology to store users' information. Moreover, since the data in the blockchain is available to authorized users only, consumers are able to self-manage the information stored in the blockchain, thus eliminate the possibility of information leakage [15]. Suning's business involves diverse aspects of retail industry, such as clothing, food, housing, and transportation. Building a member ecology can strengthen communication and enhance the sense of user experience. The ecology enables consumers, retailers, and suppliers to join the same blockchain network to avoid the unnecessary cost due to information asymmetry.

3.1.2. Smart logistics

Smart logistics is mainly reflected in supply chain management. Most traditional supply chains use one single approach - bulk goods reach the distributors through the supply chain and are distributed to smaller sales stores [16]. However, Blockchain technology allows factories to supply according to customers' demand, eliminating the need for customers to wait for replenishment and travel to physical stores. The supply chain no longer adjusts passively based on customer demand but proactively predicts consumer demand and market trends [17].

Blockchain technology can also monitor products throughout the logistics process to ensure that goods are traceable. On one hand, real-time logistics records track each transaction and delivery environment, avoiding incidents such as goods lost or replaced in the middle of the process. It strengthens the management of employees and reduces the possibility of

corruption. On the other hand, blockchain technology can implement precise regulation of cargo transportation. All parties involved in the transportation process can exchange real-time data and electronic documents without barriers [18]. Thus, the most optimal transportation route can be filtered out. In 2018, Suning collaborated with Baidu to develop driverless delivery vans. This has made uploading logistics information to the blockchain network easier [19].

According to statistics, Suning's smart logistics network has covered 95% of the areas in mainland China. The company also plans to reach 20 million square meters storage area by 2025. In August 2020, Suning, China Logistics, and other partners jointly released the "5G smart logistics innovation demonstration white paper", in which Suning's experience in exploring smart logistics using blockchain and 5G technology was made public [19].

3.2. Suning's blockchain in financial trade

Since the invention of Bitcoin, the financial industry was the first sector to achieve a breakthrough in blockchain. Suning's blacklist sharing platform is one of the influential application areas of blockchain [12]. Traditional blacklists are used to record individuals or organizations with serious adverse credit problems, and are used mainly by financial institutions. However, most traditional blacklist data are non-public. It is thus difficult to investigate whether the other party commits credit problems before a transaction, which results in inevitable losses for both parties. Suning took the initiative to use blockchain technology to build a decentralized blacklist sharing platform. Blacklists in the industry are pooled and published to authorized organizations. This list is shared with other nodes in the blockchain network through smart contracts [12]. After one year's implementation, the platform gathered 11 million pieces of data, including 7 financial institutions on the blacklist [20]. In September 2020, Suning bank officially joined forces with Citic bank and Minsheng bank to create a blockchain consortium that used the blacklist sharing platform of Suning.

3.3. Suning's blockchain in IoT

Traditional IoTs relies on a centralized structure where all devices are authenticated by accessing a centralized cloud server. When the IoTs applications expand, it is highly prone to a system collapse, and lead to high financial and human resources costs. Meanwhile, the highly centralized nature makes all data stored in the IoTs hub. Once the hub is attacked, all users' private data will be lost or maliciously exploited [12]. Blockchain technology can solve this problem. In the following, we will introduce Suning's two applications of blockchain in IoTs.

3.3.1. Smart home

As a company that promotes smart retail ecology, Suning has taken the lead in the industry to build an ecological intelligence platform to introduce smart home applications. Suning built a Biu+ ecological system based on artificial intelligence technologies, and blockchain technology in the Biu+ smart home system realizes the goal of information interoperability [21].

3.3.2. IoTs financing platform for cars

Suning launched the "Blockchain + IoTs" financing platform for cars in March 2019. Blockchain technology solves the problem of poor credibility supervision of movable assets pledge and inventory pledge financing business. Business process data can be uploaded to the blockchain platform, while sensors can monitor the goods in the warehouse in real-time to ensure the safety of goods during the pledge period. Suning claims that using the platform, the credit volume of the inventory financing business is expected to have a tenfold increase [22].

3.4. Blockchain in culture and creativity

Cultural creation involves film and television, audio-visual, media, and craft design. The degree of copyright protection directly affects the motivation of creators. The major problem with copyright protection is that the certification mechanism is time-consuming and costly. Inspired by the feature of blockchain, Suning uploads the creators' content to the blockchain, and the generated timestamp is a unique proof of work. Subsequent modifications to the work are recorded in the block in chronological order. When the work faces the problem of copyright protection, the data in the blockchain will become a critical source to prove the originality. Meanwhile, smart contracts allow creators and consumers to transact directly without the presence of a third party (e.g., record label and broker). The privacy of the entire sales process is protected, and third-party profit sharing can be largely avoided [12].

The key milestones of the implementation process are summarized in Table 1.

Table 1: Milestones of Blockchain implementation in Suning

Time	Milestones of implementation
April 2017	Suning launched a blockchain-based Letter of the credit system (BCLC), a blockchain financial blacklist data service system.
February 2018	Suning launched the blacklist sharing platform system.
July 2018	• Suning and SAP carry out technological innovation and industry practice and promote the digital economy. • Suning officially published Suning's blockchain white paper and blueprinted different application scenarios of blockchain technology in retail, logistics, finance, culture, and creative industries. • The Suning blockchain commodity traceability platform was launched.
September 2018	Suning Bank successfully granted pledge credit to the coal of Taihe port services through a chattel pledge financing platform based on blockchain and IoTs.
March 2019	Suning finance's "Blockchain + Internet of Things" finance platform was launched.
April 2019	The Suning blockchain forfaiting platform was launched.
June 2019	Suning demonstrated its ambition for blockchain technology in smart homes by deploying Suning's Biu+ ecology.
End of 2019	Suning launched Suning's Blockchain as a Service (BaaS) cloud infrastructure platform.

Beginning of 2020	Suning financial technology launched the blockchain asset securitization service system, cooperating with the initial issuance of RMB 800 million first way asset-backed security loans.
February 2020	The Suning logistics "5G Wolong" unmanned vehicle empowered by blockchain successively completed unmanned terminal delivery deployments in Beijing, Nanjing, and Suzhou.
August 2020	The "5G smart logistics innovation demonstration white paper" that shares the strategic acumen of the combination between blockchain and 5G technology in smart logistics was jointly released by Suning and its alliances.
September 2020	• The Suning blacklist sharing platform buttressed the development and launch of the blockchain consortium, established by Suning bank, Citic bank, and Minsheng bank. • Suning was invited to be a member of the official Jiangsu Internet society blockchain standardization technical committee.
November 2020	The blockchain information sharing platform of anti-money laundering led by Suning bank started its services.

Suning has benefited from the large-scale adoption of blockchain technology. User retention has dramatically increased as a result of blockchain's utilization in user profiling, and Suning is now able to proactively predict customer demand and market trends. Meanwhile, the expenditures brought on by information asymmetry were decreased [15, 17]. Suning was able to quickly upgrade information systems and achieve 95% coverage of the national smart logistics network thanks to blockchain applications in IoT. The presence of blockchain eliminated the volatility of conventional IoT systems and decreased the expenditure on building infrastructure and maintaining systems. Additionally, Suning has been able to grow its business into industries like intellectual property protection due to the application of blockchain in cultural protection, which started in 2018 [12, 23].

As a comparison, Amazon - one of the world's best retailers in US, also has a unique insight into the expansion of its blockchain business. Its Amazon Managed Blockchain (AMB) makes it easy for clients to join public networks or create and manage scalable private networks using the popular open-source frameworks. Many famous companies in the world employed AMB services. Examples include: BWM, Sage, SonyMusic, and Nestle. AMB solutions were also used to facilitate supply chain management, international transactions, digital right protection, record keeping of all kind and more. In sum, Amazon developed a simpler and more efficient blockchain program that significantly reduces operational time and offers more than 70 solutions for customers [24]. Compared with Suning, Amazon provided more international services with a larger capacity across multiple areas to its clients, whereas Suning focuses more on the Chinese business, but still having an advantage in a wide application area of blockchain. It will take Suning a longer time to develop its blockchain technology to reach the maturity level of Amazon.

4. Blockchain in the retail industry

In the post-epidemic era, the challenges faced by retailers, such as supply chain maintenance, product sales, and employee safety, urged the development of blockchain technology in retail industry [25]. The application of blockchain technology is mainly reflected in three aspects. Firstly, it builds consumer confidence in the brand: blockchain has significantly reduced customer concerns about information security leaks. Likewise, supply chain management with blockchain participation ensures the quality of goods. Secondly, it improves the ability of suppliers to meet market demand: the supply chain under blockchain management actively adjusts the production quantity and inventory of goods according to market demand, reducing unnecessary costs. Finally, it creates additional competitive advantages for firms: customer profiling based on customer information, understanding consumer preferences, and designing products to meet market demand.

Alibaba and Jingdong are the industry leaders ahead of Suning in terms of blockchain technology applications. The blockchain competition between these three top companies started with the "618" (an online shopping festival organized by e-commerce platforms in China on June 18th) promotion in 2018. Jingdong generated CNY259.8 billion turnover, and Ali Group's Tmall reached CNY213.5 billion. The high turnover has led to concerns about the quality of goods. Using blockchain technology to trace the originality of goods became a key to solve this concern. Jingdong firstly applied blockchain technology for anti-counterfeit traceability of mother and child, global purchase, and alcohol products. It made the total number of traceable commodities increased by more than 200 times within the 18 days of the "618" campaign. The platform has also attracted more than 400 companies worldwide, which further increased the number of traceable goods to 1 billion pieces [26].

In sum, for the retail industry, the significance of blockchain mainly lies in protecting user information, anti-counterfeit traceability of products, and monitoring the logistics process. A reasonable implementation of blockchain technology can help retail merchants win customers' trust and save unnecessary costs.

5. Blockchain in PingAn insurance

Ping An Insurance Co. of China Ltd (hereafter PingAn) is a Shenzhen-based Chinese holding company engaged in insurance, banking, and financial services, with thirty subsidiaries located domestically and oversea. PingAn was ranked No. 7 on the Forbes Global 2000 list of companies in 2021 and No. 29 on the Fortune Global 500. PingAn has generated $107 billion in gross premiums as China's largest insurer in 2018 alone. In July 2019, PingAn became the second-largest insurer globally, with a market capitalization of $220 billion [27]. PingAn joined the R3 Blockchain Alliance in 2016 to collaborate with a global consortium on how to apply blockchain technology to the financial market, which was PingAn Group's first attempt at blockchain technology [28]. As a traditional insurance and finance company, PingAn aspires to be positioned more as a technology company providing cutting-edge technologies [29].

In the traditional insurance industry, policies are primarily documented in paper-based contracts, and consumers would call or meet the customer representative to finalize the contract. This process relies on manual operations and is subject to the risk of information loss, tampering, and misinterpretation. It is stated that fraud accounts for 10 percent of all property and casualty losses and creates at least $80 billion in losses in the U.S. [30]. Additionally, the lack of security, efficiency, and customer satisfaction in traditional insurance sales are issues that cannot be ignored. Blockchain technology could be an intelligent solution here [30].

Realizing the significant impact of blockchain in insurance business, PingAn increased the investment of research and development in blockchain. 2021's knowledge asset management report shows that PingAn has obtained 1,749 blockchain patent families, implying a significant increase compared to the 291 in 2019 [31]. PingAn's powerful blockchain technologies have created a new financial ecosystem covering five substantial areas: finance, real estate, automotive, healthcare, and smart cities. The blockchain system developed by PingAn has attracted 13.32 million Internet users in 2020 and reached 225 million currently [32].

5.1. PingAn's blockchain in finance

PingAn OneConnect is an associate company of the PingAn insurance conglomerate. It has perfectly solved the challenges of privacy protection and system performance with its self-developed FiMAX S3C cryptographic blockchain structure. In addition to basic blockchain application methods such as trade finance, supply chain finance, and asset-backed securities, PingAn OneConnect also offers customized blockchain solutions for enterprises. Unlike traditional blockchain nodes generated by multiple accounts, the BNaaS (Blockchain Network as a Service) platform developed by PingAn allows users to independently create and publish new blockchain networks and join existing commercial blockchain networks through the BNaaS marketplace [13].

The blockchain-enabled platform, jointly built by OneConnect and UBX Philippines, targets Micro, Small, and Medium Enterprises (MSME) banking needs. It aggregates blockchain, face recognition, micro-expression, big data, and AI technologies and covers many business finance needs. Therefore, it can integrate more MSMEs into the financial system [33]. In addition, OneConnect developed a new smart legal contracting system in 2019, and actively rolled out ALFA smart contracts. The contracts cover seven major financial sectors: banking, funds, securities, trusts, leasing, futures, and insurance across PingAn Group's various divisions. Using this smart contract, managing Asset-Backed Securities (ABS) scheme contracts that would initially take two employees two to three weeks can now be completed in half an hour. After a wide-range use, ABS assets traded through ALFA reached CNY100 billion ($13 billion). The technology has significantly reduced the contracting time by 85% [34].

5.2. Blockchain in real estate, automobile, medical care, and smart city

PingAn, one of China's largest real estate investors, spends CNY50 billion ($7.7 billion) of its annual budget on new physical asset investments, representing 10% of its total investments each year [35]. Lease transactions and property registration in the real estate business involve preserving confidential information such as property rights, management rights, and transaction prices. With the help of blockchain technology, the accuracy and efficiency of the above activities can be improved.

In the automotive industry, where PingAn is involved, blockchain technology enables the traceability of the sources of auto parts and improves operational transparency. Moreover, blockchain allows the automation of processes in car-sharing services. For instance, with the assistance of smart contracts, blockchain will help two parties create an agreement, after all the conditions of the agreement are fulfilled, this agreement will be automatically executed [36].

PingAn's healthcare ecosystem, released in 2020, contains one of the largest healthcare databases in the world. The database holds the second largest number of health technology patent applications globally, covering 30,000 disease profiles, 300 million apps, and over 1 billion medical consultation records [37]. Blockchain technology is mainly used for anti-counterfeit traceability of drugs and storage of patient electronic profiles.

PingAn signed an agreement with the Sanya city government to build a "smart city" based on blockchain, artificial intelligence, big data, and cloud computing technologies. PingAn group invested CNY30 billion（$4 billion) to construct Sanya's smart city [38]. The key milestones of the implementation process in PingAn are summarized in Table 2.

Table 2: Milestones of Blockchain implementation in in PingAn

Time	Milestones of implementation
May 2016	PingAn became the first Chinese financial services institution to join the R3 consortium.
October 2016	PingAn OneConnect successfully registered the patent for "blockchain-based transaction verification method and system", "permission control method and system based on blockchain transaction", and "blockchain-based transaction processing method and system."
January 2017 – May 2018	PingAn OneConnect successfully registered a series of patents for a Blockchain related systems and methods.
August 2018	PingAn unveiled China's first "1+N" smart city integrated platform, supported by blockchain technology.
October 2018	HKMA uses PingAn's FiMAX blockchain to establish an eTradeConnect platform for seven international banks in Hong Kong.
December 2018	PingAn Good Doctor started integrating blockchain technology in their smart severe disease monitoring system, AI-based medical system, PASS and smart medical safety monitoring platform to improve the data security of the healthcare ecosystem.
March 2019	PingAn OneConnect works with IFAB to establish a blockchain-based IFAB trade finance network for SMEs.

April 2019	PingAn OneConnect established the Tianjin port blockchain-based verification pilot project.
May 2019	PingAn OneConnect built and operated a blockchain-enabled integrated Gamma Platform, allowing clients to migrate, manage and enhance their technology infrastructure and simplify the digital transformation process.
June 2019	• PingAn OneConnect started managing the contracts by the ALFA smart contract cloud platform, which supported 1,000 standard contract templates and 80,000 tags stored in the blockchain. • PingAn FiMAX won the Best Blockchain or Distribution Ledger Technology Award from the Asian Banker, the Outstanding FinTech Achievement and the Best Application of Advanced Technology in a Product or Service from bank administration institution.
July 2019	PingAn OneConnect officially launched the Blockchain Network-as-a-Service (BNaaS) for customers to create their blockchain-based networks or take part in existing networks established by others.
December 2019	• PingAn's International Smart City subsidiary launched iShenzhen blockchain electronic license application platform. • PingAn integrated the blockchain witness function and launched the "PingAn Good Lawyer" smart legal compliance assessment systems.
April 2020	PingAn OneConnect and China merchants' port group created a blockchain-based system with Shenzhen customs to serve the Guangdong-Hong Kong-Macao Greater Bay Area.
May 2020	PingAn was one of the pioneers utilizing the blockchain technology to develop a tax-industry alliance chain for the sake of reducing tax costs and trade finance risk.
August 2020	PingAn's International Smart City subsidiary was invited to be the chief compilation team as recorded in the bluebook of blockchain application innovation in the area of Beijing government services.
October 2020	PingAn launched the intelligent transportation integration platform- "155C" to resolve traffic congestion-related issues.
November 2020	• PingAn filed the largest number of blockchain-related patents in 2020 [31]. • PingAn FiMAX was listed on the Top 30 list of the 2020 China Blockchain entities and was rewarded by 5th Golden Gyro Awards for its contribution to the Blockchain industry.
September 2021	• PingAn OneConnect had collaborated with 62 government agencies and regulators and 41 IT service providers to benefit over two million SMEs through its financial ecosystem. • PingAn's FiMax architecture has been widely recognized and won 22 technology awards [39].

In PingAn's insurance business, blockchain technology has significantly advanced the digital transformation, improved traditional insurance paper-based transactions and enhanced employee efficiency. As a result, there has been a corresponding reduction in errors and frauds previously caused by manual labor. PingAn OneConnect is also very good in developing blockchain platforms and offers blockchain services to other companies. The applications of blockchain in other areas has automated procedures and increased operational transparency in various ways. Due to the increased accuracy of confidential information and transaction efficiency, PingAn has won the trust of investors and clients [36]. Statistics show that after blockchain technology was adopted, the total number of users surged to 225 million [32].

As a comparison, United Health Group (UHG) U.S. - the benchmark for blockchain technology adoption in the health insurance sector was chosen. UHG uses the blockchain alliance the same way as PingAn. However, UHG's alliance brings together significant, industry-leading businesses from several sectors and encompasses multiple commercial relations, including suppliers and rivals. In contrast to PingAn's blockchain alliance, which is more homogeneous and is mostly constituted of insurance businesses, UHG's blockchain alliance deploys a multi-company, multi-site, permissioned blockchain. Each alliance member has the flexibility to deploy its nodes in accordance with its corporate demands. To this extend, PingAn's blockchain technology applications are yet to be as innovative and diversified as those of international leaders [40].

6. Blockchain in the insurance industry

Blockchain technology is used in the insurance industry mainly as a decentralized and shared distributed ledger to prevent the risk of fraud in the insurance business by providing immutable and time-stamped records of transactions. In recent years, insurance companies have gradually benefited from blockchain technologies in their insurance business and have significantly increased their investment in developing new models of blockchain technologies.

The insurance market size of blockchain was estimated to be $1,393.8 million by 2023, with an 84.9% compound annual growth rate (CAGR) [41]. The applications of blockchain are reflected in the following aspects: (1) fraud monitoring and risk prevention: insurance companies can keep all transaction records permanently on a distributed ledger and can use public data to predict and analyze fraudulent activities, (2) increasing customer trust in the brand: cryptographic principles in blockchain are used to authenticate customers and ensure transaction security and customer privacy, (3) claim processing: use blockchain technology to analyze data and significantly speed up claim processing, (4) smart contracts: blockchain's smart contracts automatically fulfill contracts when all conditions are met, reducing back-end

paperwork and lowering insurance company management and operating costs.

The current situation shows that insurers are more cautious in blockchain deployment compared with the retailers, and blockchain is primarily applied to prevent fraud risks and manage information. However, some start-up insurers are trying to expand blockchain operations to other areas such as accessing and auditing electronic health records and verifying settlements. These activities are expected to save costs and improve operational efficiency [36].

7. Comparison between the two cases and two industries

So far, we have introduced the implementation of blockchain technology in two companies - Suning and PingAn, and described the trend of blockchain applications in the retail and insurance industries. In this session, we will first use Technology-Organization-Environment (TOE) framework to summarize the factors leading to blockchain implementation success, and then compare the similarities and differences between these two companies. The summarized features of blockchain applications in both industries will be presented at last.

7.1. TOE framework and common factors to blockchain applications success

With the analysis of the two cases, we argue that Suning and PingAn share several commensurate characteristics that lead to their success in blockchain implementations. We draw on the Technology-Organization-Environment (TOE) framework to explicate how and why these characteristics are essential and generalizable rather than domain-specific to all industries and types of organizations.

TOE framework was initiated by Tornatzky and Fleishcer in 1990 [42], and further extended to suit different contextual and/or situational analyses. The main principle of the TOE framework advocates the contextual congruence of technology itself, organization, and external environment to the technology adoption and implementation [43]. In this paper, we draw on the TOE framework extended by [44], which is believed to be more appropriate and deliberately designed for information and communication technology studies. Further, our analysis responds to [44]'s call to examine the external support and user-centric view in different contexts [44].

In TOE, the technological context highlights not only the significant characteristics of technology but also the "compatibility" and "interoperability" elements of the current system [45]. Three sub-factors are conceptualized in the technology context, including system configuration, technology hardware readiness, and technology compatibility. Firstly, system configuration refers to the technical architecture configuration and integration of both software and hardware [44]. Suning and PingAn have done well in this aspect, evidence can be found from the different awards they have received. For example, Suning has been selected for the 2020 FAT blockchain fintech leading 100 lists. PingAn FiMAX won the outstanding Fintech achievement and the best application of advanced technology from a bank administration institution. Secondly, technology readiness refers to the maturity and capability in technology deployment [46]. PingAn has tried to integrate blockchain technology into current business practices and exercises; Suning has pioneered in transforming its businesses with blockchain. Though there might be slight differences, we believe such a difference can be attributed partially to the unique business nature. Thirdly, technology compatibility refers to how technology is perceived as consistent with the current organizational values and practices [47]. We argue that both companies did more than embracing blockchain technology alone; they tried to align it at a strategic level to explore any potential competitive advantages brought by blockchain.

Regarding to the organizational context, the organizational support to the new technologies, the readiness for the technology implementation, resource investments and supportive corporate culture can impact the success of technology implementation to a greater extent [43]. According to the TOE framework, the organizational support can be categorized into two factors: organization fit and user barrier. It is believed that Suning and PingAn's success in blockchain applications is never possible if they have not invested enough time, resources and training for all relevant users, including employees, clients and business partners. Both companies proactively distributed and shared the information with different stakeholders and invited them to get involved in their platform and ecosystem development. For instance, Suning advanced blockchain development by integrating 5G technology with allies in the "5G smart Logistics Innovation Demonstration White Paper", and PingAn developed the tax-industry alliance chain with other partners. On the other hand, we found that both companies spontaneously value users' acceptance of product and service change. They released the news about their new businesses with blockchain and benefits their customers can acquire from these new businesses through different channels such as newspapers, daily news, social media, and even state media, which have significantly cultivated customers' confidence on the new business and technologies in China.

Finally, the environmental context considers the views of external stakeholders seriously, organization should engage in the prosperity of the ecosystem with high commitment [43]. External support is the overarching factor in the organizational context of technology implementation success. In addition to clients and business allies, government plays a critical role in Suning and PingAn's success in blockchain implementation. Specifically, the Chinese central government clearly encourages blockchain industrial ecosystem establishment and blockchain industrial applications. It issued 217 policies, regulations, and program documents pertaining to blockchain technology, clarifying the legal boundary of blockchain application and implementation [48]. For instance, one of the initiators of PingAn's tax-industry alliance chain is the Shenzhen government. In short, based on the TOE framework, we argue that Suning and PingAn share some common characteristics of success in blockchain applications. These commonalities are theoretically and practically generalizable to other fields. Table 3 summarizes these common factors.

Table 3: Common factors to blockchain application success based on TOE framework

Technological context	
System Configuration	Suning and PingAn both have well-developed technical architecture and

	integration systems on both software and hardware.
Technology Readiness	Suning and PingAn both demonstrated maturity and capability in technology deployment. Correspond to different business natures, Suning focused more on blockchain adoption whereas PingAn is more on development and implementation.
Technology Compatibility	Suning and PingAn not only integrated blockchain technology into their current business coherently, but also embedded it into their business strategies.
Organization context	
Organizational Fit	Suning and PingAn's invested much time and resources to train users to process and handle the change brought by blockchain in their business. Both organizations also proactively distributed and shared the information with different stakeholders, and invited them getting involved in their platform and ecosystem development.
User Barrier	Suning and PingAn maximized the benefits from different information channels to promote their blockchain technology and related new business for the sake of aggrandizing customers' acceptance.
Environmental context	
External Support	Suning and PingAn both benefited from blockchain-related national regulations and policies as well as the collaboration with state and/or provincial government.

7.2. Similarities and differences between two companies

First, though Suning and PingAn belong to two different industries, they share some common features. The similarities are summarized in Table 4.

Table 4: Similarities between the two companies

The scale of blockchain application	Large scale covers multiple fields
Application areas	Focus on finance, real estate, supply chain, smart home or smart city, data management, and customer information management
Blockchain features used	Secure, unanimous, distributed, immutable, time-stamped
Government support	Both are based in China and strongly supported by the Chinese government [48]
Patent	Both are obtained patents in blockchain technology
Blockchain provider	Provide blockchain service (BaaS) to other companies [13]
Cooperation with other companies	Both are collaborating with other companies to develop blockchain technology
Common benefits	• Enhance information reliability • Process a large quantity of data automatically • Increase information transparency • Improve security and accountability in transaction • Reduce financial risk • Reduce labor costs
Common challenges	• Solve the relations of production but not productivity itself • Optimize algorithmic capabilities to analyze data • Prevent crime due to the anonymity of the blockchain • Expand potential uses of blockchain • Interoperability between blockchains • System performance • Ecosystem development

Second, the differences between the two companies are presented in Table 5.

Table 5: Differences between the two companies

	Suning	**PingAn**
Application areas	Retailing and retailing-related service	Insurance business and blockchain technology development (blockchain service provider)
Blockchain in IoT	Real-time monitoring of pledged goods in the warehouse and uploading business process data to the blockchain	Collecting insured vehicle data and uploading it to a blockchain-based insurance network
Independently developed cryptographic blockchain structure	FiMAX S3C [13]	No independently developed cryptographic blockchain structure
Vision	Domestic e-commerce business development	International business development
Strategic management	Establish a strategic alliance	Improve product innovation

7.3. Summarized features of retail and insurance industries

The major features of blockchain technology applications in retail and insurance industries are summarized below (Table 6.)

Table 6: Summarized features of blockchain applications in retail and insurance industries

	Retail industry	Insurance industry
Application areas	• Build customer confidence in the brand • Supply chain management • Enhancing the market competitiveness	• Fraud monitoring and risk prevention • Increasing customer trust • Customer retention • Claim processing • Smart contract
Major limitation of blockchain applications	Scalability • Difficult to store the growing amount of data • Decrease transaction processing speed [49]	Security • Inaccurate data cannot be repaired • Private keys are easily lost [50]
Compound annual growth rate	84.6% [51]	84.9% [41]
Investment Return	• Strong revenue growth: including factors that support volume and price • Better operating margins: including factors that support the selling and administrative side as well as cost of goods sold [52]	Reduction of administrative and operations cost: blockchain can deliver reinsurance industry-wide savings up to $10 billion [53]
Influence on industry	• Reduce retail costs and waste of resources • Optimize customer-retailer relationships	• Replace the role of insurance companies as intermediaries • Reduce insurance fraud losses
Outlook on next step	• Reduce the risk of data breaches • Focus blockchain development resources on use-cases with a clear path to commercialization • Push for standardization in technology, business processes and talent skill sets [52]	• Shape a stimulating and regulatory environment • Identify the challenges around blockchain's open and decentralized nature [54]

8. Conclusion

Blockchain technology has gone beyond the early hype of Bitcoin and evolved into an emerging technology commonly adopted by governments and industries with the general goal of improving productivity and reducing costs. China always keeps an open and positive attitude towards developing and using blockchain, especially in the retail, insurance, and finance sectors. Since the year of 2017, China has dominated global blockchain development market.

For the future, the data breach risks will be the main focus of the future blockchain technology development in retail. Regarding to the current cryptocurrency landscape, most retail industry employees expressed their confusion and distrust of blockchain technology. Therefore, developing a trustworthy and secured blockchain information systems will be a key priority to advance in the retail sector [55]. User or customer education is also necessary. Meanwhile, most company would like to embrace below two approaches: focusing blockchain development resources on use‐cases with a clear path to commercialization and pushing for standardization in technology, business processes and talent skill sets [54].

For insurance industry, the way forward for blockchain technology will be to shape a stimulating and regulatory environment. While blockchain technology can provide technical proofs of ownership, consensuses and responsibilities, these proofs still lack legal support. For example, smart contracts can technically bind both parties to the contract, but the law does not explicitly acknowledge the contract's legitimacy. Therefore, the next step for blockchain in the insurance industry should be establishing and implementing specific rules and regulations [56]. Moreover, collaborating with consortia, technologists, startups, and regulators to identify challenges around the open and decentralized nature of blockchain is critical to the insurance industry. The areas include: the limitations of the technology, the development of the market, data protection, and the standardization of operational requirements [54].

This paper firstly introduced the origin and background of blockchain technology and briefly described the technical characteristics of blockchain. It then analyzed the implementation of blockchain in two companies from two industries - Suning in the retail industry and PingAn in the insurance industry. We summarized the success factors of blockchain implementation by

using TOE framework. For technological context, both companies are well developed in system configuration; both of them demonstrated maturity and capability in technology deployment; and both technology deployments are compatible with the companies' long time strategic development. In terms of organizational context, both companies' blockchain developments fit their organizational development and are welcomed by their clients. At last, the blockchain technology implementations are both strongly supported by government. We also compared the similarities and differences between the two companies in blockchain applications. Both companies are pioneers in their specific domain, though sharing slightly different missions or goals in blockchain implementation and development. Compared with US firms, China's leading companies still have some distances in terms of the kind of blockchain services provided and the maturity level. At last, the summarized features of blockchain applications in two industries were presented. The investment of return and future outlook were discussed. In sum, this study provided deep insights in understanding the development of blockchain technology in China, especially in the retail and insurance industries.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Y. Cai, C. Qi, "Blockchain Technology Applications in Retail and Insurance Sectors: Cases from Suning and PingAn," in 2021 International Conference on Artificial Intelligence and Blockchain Technology (AIBT), IEEE: 80–84, 2021, doi:10.1109/AIBT53261.2021.00020.

[2] M. Quiniou, "Blockchain: The Advent of Disintermediation.," Newark: John Wiley & Sons, Inc, 2019, doi: 10.1002/9781119629573.

[3] K. Kulkarni, Learn Bitcoin and Blockchain, Packt Publishing Limited, 2018, available at: https://www.packtpub.com/product/learn-bitcoin-and-blockchain/9781789536133.

[4] Z. Zheng, S. Xie, H. Dai, X. Chen, H. Wang, "An Overview of Blockchain Technology: Architecture, Consensus, and Future Trends," in 2017 IEEE International Congress on Big Data (BigData Congress), IEEE: 557–564, 2017, doi:10.1109/BigDataCongress.2017.85.

[5] F. Casino, T.K. Dasaklis, C. Patsakis, "A systematic literature review of blockchain-based applications: Current status, classification and open issues," Telematics and Informatics, **36**, 55–81, 2019, doi:10.1016/j.tele.2018.11.006.

[6] A. Hayes, Blockchain Definition: What You Need to Know, 2022, available at: https://www.investing.com/academy/crypto/blockchain-definition/.

[7] A.K. Kar, L. Navin, "Diffusion of blockchain in insurance industry: An analysis through the review of academic and trade literature," Telematics and Informatics, **58**, 101532, 2021, doi: 10.1016/j.tele.2020.101532.

[8] M. Belotti, N. Božić, G. Pujolle, S. Secci, "A vademecum on blockchain technologies: When, which, and how," IEEE Communications Surveys & Tutorials, **21**(4), 3796–3838, 2019, doi: 10.1109/COMST.2019.2928178.

[9] J.J. Bambara, P.R. Allen, K. Iyer, R. Madsen, S. Lederer, M. Wuehler, Blockchain: A practical guide to developing business, law, and technology solutions, McGraw Hill Professional, 2018, available at: https://books.google.com.hk/books/about/Blockchain_A_Practical_Guide_to_Developi.html?id=tg7dswEACAAJ&redir_esc=y.

[10] M. Nofer, P. Gomber, O. Hinz, D. Schiereck, "Blockchain," Business & Information Systems Engineering, **59**(3), 183–187, 2017, doi:10.1007/s12599-017-0467-3.

[11] G. Strawn, "Blockchain," IT Professional, **21**(1), 91–92, 2019, doi:10.1109/MITP.2018.2879244.

[12] Birost, "Suning releases blockchain white paper, Suning blockchain will be applied to multiple scenarios of smart retail ecology", 2022, available at: https://blog.birost.com/a?ID=01000-5640385f-a6a0-4b08-9cb7-e9d60c5e5743

[13] A. Frosinini, DLT solutions for trade and finance China focus, 2020, available at: file:///D:/Users/mslaurel/Downloads/DLT%20solutions%20for%20trade%20and%20finance%20-%20China%20focus%20-%20Final%20(2).pdf

[14] English.Wimc.Org.Cn, "Zhang Jindong: To Drive Intelligent Manufacturing with Intelligentitle", 2022, available at: https://english.wimc.org.cn/ennews_show.aspx?id=85.

[15] Xinhua, "Suning releases blockchain white paper", 2018, available at: https://en.imsilkroad.com/p/103916.html

[16] J. Sell, "How Omni-Channel Commerce Is Changing Traditional Supply Chains", Inbound Logistics, 2015, available at: https://www.inboundlogistics.com/articles/how-omni-channel-commerce-is-changing-traditional-supply-chains/.

[17] G.R. Sheffield, "An examination of e-commerce and its influence on the traditional and e-commerce supply chain models", 2019, available at: https://www.proquest.com/docview/2191207183?pq-origsite=gscholar&fromopenview=true.

[18] Y. Cai, C.M.I. Cheung, K.W. Chan, X. Ou, "Be more specific! An exploratory study of 5g and blockchain application on retailing industry with suning limited case," J Huma Soci Scie, 5 (1): 21, 32, 688–2690, 2022, available at: https://opastpublishers.com/open-access/be-more-specific-an-exploratory-study-of-5g-and-blockchain-application-onrnretailing-industry-with-suning-limited-case.pdf.

[19] VisionNav Robotics, "818 Suning Smart Logistics Enlarges The Move! 5G Automated Warehouse Opens to The Media For The First Time", 2020, available at: https://www.visionnav.com/news/index-8.html#news.

[20] Cindy, "Suning Financial Blockchain Blacklist Sharing Platform has collected 11 million blacklist data - , BTC Trade", 2019, available at: https://www.ibtctrade.co/application/2667.html?lang=en.

[21] C.U. Inc, "Suning Unmanned "Biu" Store to make International Debut at CES 2018 to Showcase Smart Retail Worldwide", 2018, https://en.prnasia.com/releases/global/suning-unmanned-biu-store-to-make-international-debut-at-ces-2018-to-showcase-smart-retail-worldwide-199089.shtml.

[22] V. He, "Suning was Praised for Its "Blockchain + Internet of Things" Financing Platform for Cars, 8BTC.COM", 2019, available at: https://news.8btc.com/suning-was-praised-for-its-blockchain-internet-of-things-financing-platform-for-cars.

[23] Suning Holding Group, "Suning Holdings Group Ranks Second on China's Top 500 Non-state Owned Enterprises List in 2020, Cision PR Newswire", 2020, available at: https://www.prnewswire.com/news-releases/suning-holdings-group-ranks-second-on-chinas-top-500-non-state-owned-enterprises-list-in-2020-301129798.html.

[24] Amazon Web Services, "AWS's blockchain", 2022, https://docs.aws.amazon.com/whitepapers/latest/aws-overview/blockchain.html.

[25] H. Anwar, "Blockchain in Retail: Use Cases and Potential Applications, 101 Blockchains", 2020, available at: https://101blockchains.com/blockchain-in-retail/?gclid=Cj0KCQjwteOaBhDuARIsADBqReg36L-4X5UJsJk1ftX-HKaWCwV14QwIJLTRvWwkBbazf7NLw7XOAJEaAuRZEALw_wcB.

[26] Blocking, "Ali, Jingdong, Suning, "chain" battle double eleven Blockchain Network", 2020, available at: https://blocking.net/21987/ali-jingdong-suning-chain-battle-double-eleven/.

[27] F. Lu, "Ep. 144 – A new approach to blockchain – Ping An's Insights , Insureblocks", 2021, available at: https://podcasts.apple.com/in/podcast/ep-144-a-new-approach-to-blockchain-ping-ans-insights/id1361820676?i=1000505557258.

[28] J. Kelly, "Ping An becomes first Chinese member of R3 blockchain consortium", Reuters, 2016, available at https://group.pingan.com/media/news/News-2016/Ping-An-becomes-first-Chinese-member-of-R-blockchain-consortium.html.

[29] Bloomberg News, "China's Ping An Insurance to Spend $22B on Tech, From AI to Blockchain", Insurance Journal, 2019, available at: https://www.insurancejournal.com/news/international/2019/10/24/546474.htm#:~:text=Insurance%20Giving%20Back-,China's%20Ping%20An%20Insurance%20to%20Spend%20%2422,Tech%2C%20From%20AI%20to%20Blockchain&text=Ping%20An%20Insurance%20wants%20to,intelligence%20to%20blockchain%20will%20work..

[30] J. Tropea, "Insurance Disruption: How Blockchain Is Transforming the Industry | Bankrate", Brankrate, 2021, available at: https://www.bankrate.com/insurance/car/blockchain-disruption/.

[31] H. Partz, "Chinese holding firm Ping An overtakes Tencent in blockchain patents race", Cointelegraph, 2021, available at: https://cointelegraph.com/news/chinese-holding-firm-ping-an-overtakes-tencent-in-blockchain-patents-race.

[32] Ping An Group, "Ping An business case series: How integrated finance ecosystems are empowering millions in China", Financial Times, 2021,

available at: https://www.ft.com/partnercontent/ping-an-insurance/how-integrated-finance-ecosystems-are-empowering-millions-in-china.html.
[33] S. Writer, "UBX taps OneConnect for blockchain-enabled platform for MSMEs", Frontier Enterprise, 2019, available at: https://www.frontier-enterprise.com/ubx-taps-oneconnect-for-blockchain-enabled-platform-for-msmes/.
[34] Ledger Insights, "China's Ping An rolls out OneConnect blockchain smart legal contracts", Ledger Insights, 2019, available at: https://www.ledgerinsights.com/china-ping-an-oneconnect-blockchain-smart-legal-contracts/.
[35] S. He, "Ping An to add $7.7bn in real assets despite China property headwind | Asset Owners", AsianInvestor, 2021, available at: https://www.asianinvestor.net/article/ping-an-to-add-7-7bn-in-real-assets-despite-china-property-headwind/472156.
[36] D. Turpitka, "Blockchain In The Automotive Sector: Three Use Cases And Three Challenges", Forbes, 2021, available at: https://www.forbes.com/sites/forbestechcouncil/2021/12/22/blockchain-in-the-automotive-sector-three-use-cases-and-three-challenges/?sh=3df5c9a82508.
[37] Ping An Insurance Group, "Ping An Unveils Health Care Ecosystem Strategy - A Growth Engine for the Group Empowered by World-leading Healthtech", 2020, available at: https://www.prnewswire.com/news-releases/ping-an-unveils-health-care-ecosystem-strategy-301136665.html.
[38] M. Yakubowski, "China: Insurance Giant Ping An, Sanya City Gov't to Build 'Smart City' with Blockchain", COINTELEGRAPH, 2018, available at: https://cointelegraph.com/news/china-insurance-giant-ping-an-sanya-city-govt-to-build-smart-city-with-blockchain.
[39] OneConnect Financial Technology Co., Ltd. "Annual Report 2021", Hong Kong, 2022, available at: https://www.annualreports.com/HostedData/AnnualReports/PDF/NYSE_OCFT_2021.pdf.
[40] United Health Group, "Improving provider data accuracy: a collaborative approach using a permission blockchain", 2018, available at: https://blockchainhealthcaretoday.com/index.php/journal/article/view/62.
[41] MarketsandMarkets, "Blockchain in Insurance Market Size, Share and Global Market Forecast to 2023", 2018, available at: https://www.marketsandmarkets.com/Market-Reports/blockchain-in-insurance-market-9714723.html.
[42] L.G. Tornatzky, M. Fleischer, A.K. Chakrabarti, The processes of technological innovation, Lexington books, 1990, doi: https://doi.org/10.1007/BF02371446
[43] R. Depietro, E. Wiarda, M. Fleischer, "The context for change: Organization, technology and environment," in the Processes of Technological Innovation, 199(0), 151–175, 1990, doi: https://doi.org/10.1007/BF02371446.
[44] T. Masood, J. Egger, "Augmented reality in support of Industry 4.0—Implementation challenges and success factors," Robotics and Computer-Integrated Manufacturing, **58**, 181–195, 2019, doi: 10.1016/j.rcim.2019.02.003.
[45] M. Gattullo, G.W. Scurati, M. Fiorentino, A.E. Uva, F. Ferrise, M. Bordegoni, "Towards augmented reality manuals for industry 4.0: A methodology," Robotics and Computer-Integrated Manufacturing, **56**, 276–286, 2019, doi: 10.1016/j.rcim.2018.10.001.
[46] K. Zhu, K.L. Kraemer, S. Xu, "The process of innovation assimilation by firms in different countries: a technology diffusion perspective on e-business," Management Science, **52**(10), 1557–1576, 2006, doi: 10.1287/mnsc.1050.0487.
[47] Y.-M. Wang, Y.-S. Wang, Y.-F. Yang, "Understanding the determinants of RFID adoption in the manufacturing industry," Technological Forecasting and Social Change, **77**(5), 803–815, 2010, doi: 10.1016/j.techfore.2010.03.006.
[48] L. Cai, Y. Sun, Z. Zheng, J. Xiao, W. Qiu, "Blockchain in China," Communications of the ACM, **64**(11), 88–93, 2021, doi: 10.1145/3481627.
[49] U. Klarman, S. Basu, A. Kuzmanovic, E.G. Sirer, "bloxroute: A scalable trustless blockchain distribution network whitepaper," 2018, available at: https://bloxroute.com/wp-content/uploads/2018/03/bloXroute-whitepaper.pdf.
[50] J. Chen, W., Xu, Z., Shi, S., Zhao, Y., & Zhao, "A survey of blockchain applications in different domains," in 2018 International Conference on Blockchain Technology and Application, 17–21, 2018, doi: 10.1145/3301403.3301407.
[51] B. Jagtap, R. Rake, S. Baul, "Blockchain in Retail Market Statistics - 2026," Applied Market Research, 2019, available at: https://www.alliedmarketresearch.com/blockchain-in-retail-market.
[52] Deloitte, "New tech on the block: Planning for blockchain in the retail and Consumer Packaged Goods industries", 2018, available at: https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/ConsumerIndustrialProducts/deloitte-uk-blockchain-in-retail-and-cpg.pdf.
[53] PwC, Blockchain, "Blockchain, A catalyst for new approaches in insurance", 2018, available at: https://www.pwc.com/gx/en/industries/financial-services/publications/blockchain-a-catalyst.html.
[54] McKinsey&Company, "Blockchain in insurance opportunity or threat", 2016, available at: https://www.mckinsey.com/~/media/mckinsey/industries/financial%20services/our%20insights/blockchain%20in%20insurance%20opportunity%20or%20threat/blockchain-in-insurance-opportunity-or-threat.ashx.
[55] M.T. Nuseir, "Potential impacts of blockchain technology on business practices of bricks and mortar (B&M) grocery stores," Business Process Management Journal, 2020, doi: 10.1108/BPMJ-06-2020-0267.
[56] J. Al-Jaroodi, N. Mohamed, "Blockchain in industries: A survey," IEEE Access, **7**, 36500–36515, 2019, doi: 10.1109/ACCESS.2019.2903554.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 167-173 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Designing the MIMO SDR-based Antenna Array for 5G Telecommunication

Meriem Drissi[1,*], Nabil Benjelloun[2], Philippe Descamps[3], Ali Gharsallah[1]

[1]*MER Laboratory, Department of electronics, FST, University Tunis El Manar, 1068, Tunisia*

[2]*IRSEEM Laboratory, Department of electronics, University Rouen Normandie, Rouen, 76801, France*

[3]*CNRS Laboratory, Department of electronics, University Caen Normandie, Caen, 14000, France*

ARTICLE INFO

Article history:
Received: 31 August, 2022
Accepted: 05 November, 2022
Online: 20 December, 2022

Keywords:
MIMO
AM Modulation
Microstrip antenna
SDR

ABSTRACT

With the significant spread of 5G telecommunication systems, the demand for high quality of service and better coverage is growing rapidly. Multiple-input–multiple-output is planned to be among the key technologies in 5G telecommunication's field. In this article, a new fully digital 2×2 MIMO testbed is implemented, using USRP B210 and 2 pairs of microstrip antennas with 4 radiation elements. In this implementation we cover the reconfigurability of the USRP combined with MATLAB, providing flexible integration between real-time and software host processing.

1. Introduction

Multiple-input–multiple-output becomes the most challenging technology in the 5G and the upcoming 6G telecommunication applications, given the ability of increasing the quality of communication and the rate of data [1], [2].

Multiple input multiple output (MIMO) has proven to be an important technique for communication systems. This technique benefits by using two or more antennas to transmit and receive data simultaneously, in order to increase the data rate and to provide a better reliability. So, MIMO provides the solution to the challenge [3], [4].

To develop a MIMO system, a microstrip antenna is designed, as well as is widely deployed in telecommunication systems due to its low profile, ease of fabrication and miniature size.

Much research such as [2]-[7] focus mostly on simulations and theorical analysis ignoring the system implementation because of its complexity, power consumption and low cost. the software defined radio has been found to avoid such limitations [1], [3]. More specially, by using SDR flexibility, reconfigurability and Rapid prototyping are covered to support new standards and protocols [8]-[10].

On this issue a new fully digital MIMO 2X2 has been implemented, this prototype is combining pairs of microstrip antennas with 4 radiation elements on both terminals.

For conveying and receive one or more types of wave signals simultaneously, An AM modulation is involved using Simulink and the flexibility, reconfigurability and low cost of USRP B210.

In our testbed, we investigate and profile the characteristics of USRP B210 platform including full steps such, transmission, modulation, spectrum, received signal and demodulation.

The rest of this document is arranged as bellow. Section II describes a comparison between rectangular and circular patch. In Section III, in order to perform the gain and assure circular polarization we moved on antenna array. In Section IV, we introduce an overview of our 2×2 MIMO testbed based on SDR. In Section IV, we define our SISO experimental results while analyzing our results. Even we report our novel model MIMO using AM modulation by studying our system's performance, finally, in section V, we come to end the article.

2. Patch Antenna

2.1. Single element design geometry

The antenna is the first component of the MIMO system. This design uses patch antenna. Considering low profile, ease of

*Corresponding Author: Meriem Drissi, drissimeriem1@gmail.com

fabrication, and low cost, patch antennas have proven to be the most frequently antennas [11].

Firstly, we start by a brief comparison between the most common shapes which are rectangular and circular configurations. FR4 epoxy has been chosen. Its relative permittivity is 4.4 [12], in order to provide a balanced comparison.

In fact, both shaped microstrip patch antennas are performed using HFSS. The simulated antennas are resonating at 3.48 GHz.

Starting with the software designed version of rectangular shape which is shown on figure 1, its dimensions are 20 ×27 mm^2.

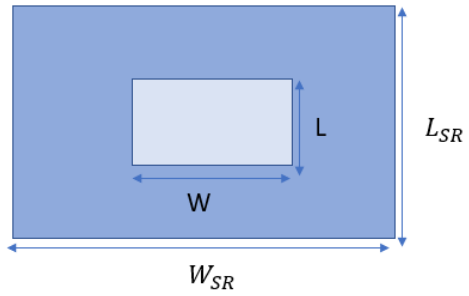


Figure 1: Geometry of the rectangular patch antenna.

In order to determine the microstrip patch antenna's dimensions, the equations below have been used. Based almost entirely on the width and length of the rectangular patch, the design has been conceived [13], [14].

$$W = \frac{C}{2f_r\sqrt{\frac{\varepsilon r+1}{2}}} \quad \text{With} \quad \varepsilon_{eff} = \frac{(\varepsilon_r+1)}{2} + \frac{(\varepsilon_r-1)}{2}\left[1 + 12\frac{h}{W}\right]^{-\frac{1}{2}} \quad (1)$$

$$\text{And} \quad L = \frac{C}{2\times\sqrt{\varepsilon_{eff}}\times f_r} - 0.824*h\frac{(\varepsilon_{eff}+0.3)(\frac{W}{h}+0.264)}{(\varepsilon_{eff}-0.258)(\frac{W}{h}+0.8)} \quad (2)$$

While the figure 2 describes the circular patch's configuration. However, equations [5] and [6] through 2.1 have been used to find out the essential parameter required for the design of this circular patch which is the radius [15].

$$a = \frac{F}{\sqrt{1+\frac{2h}{\Pi\varepsilon_r F}[\ln(\frac{\Pi F}{2h})+1,7726]}} \quad \text{with } F = \frac{8,791X10^9}{f_r\sqrt{\varepsilon_r}} \quad (3)$$

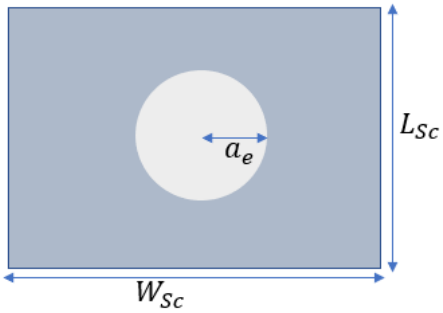


Figure 2: Geometry of circular patch antenna

2.2. Simulations results and discussions

It is clear from the figure below. both the graphs are resonating at the same resonant frequency (3.48GHz). Both shapes present a good adaptation. In fact, the rectangular patch provides a better result(50dB) by 3dB compared to the circular's return loss(47dB).

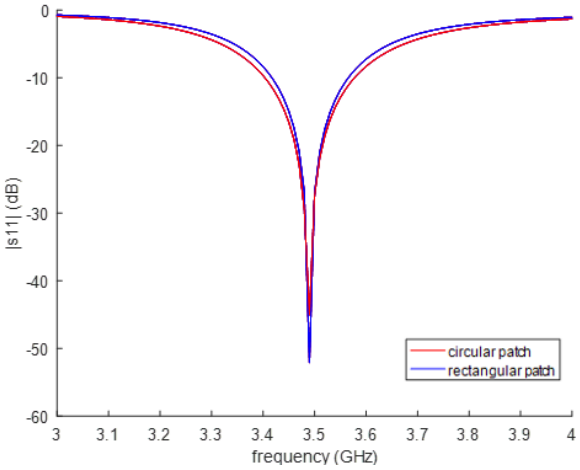


Figure 3: Comparison of return loss for different shapes.

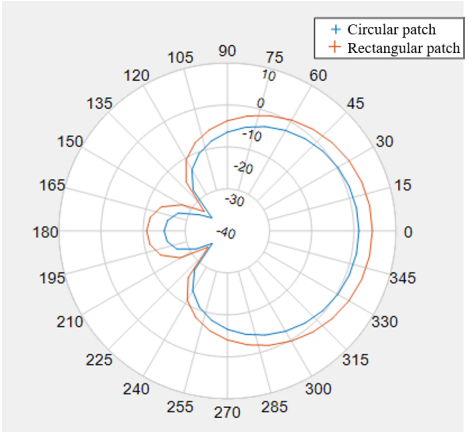


Figure 4: Comparison of 2D radiation pattern for different shapes.

For the antenna suitable for 5G telecommunication systems, a high directivity and a narrow beamwidth are both necessary.

As shown in figure 4, both designs present broadside radiation patterns. Indeed, at 3.48Ghz, the beamwidth shown by the circular patch antenna (90°) is narrower than the rectangular patch (100°). Moreover, the circular patch offers the highest gain value.

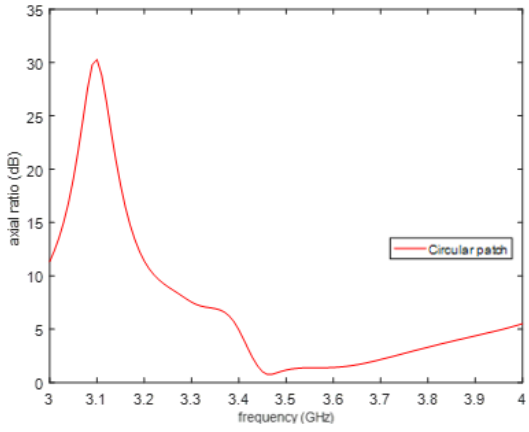


Figure 5: Simulated axial ratio (AR)

A circularly- polarized antenna is defined as having an axial ratio within 3dB at the required frequency [13].

As shown in Figure 5, the axial ratio of the circular configuration presents an axial ratio AR ≤3 dB, with AR bandwidth (less than 3 dB,) is from 3.4 GHz to 3,7 GHz, in fact it reaches 0.7 dB at 3.48 GHz

According to HFSS simulation results, the rectangular shape provides better return loss, whereas the circular shape offers the enhanced gain, the smaller size and the narrower beamwidth.

Occasionally, a unique radiation element can meet the needs of the required antenna. But characteristics such beamforming beam steering, and beam scanning are usually achievable when radiation elements are arranged. So, in order to enhance its performance (gain, beamwidth), a compact 4-Element Microstrip antenna array is designed with the most efficient shape.

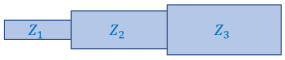
3. Array antenna

3.1. Antenna prototype

To increase the gain and directivity of the antenna, an antenna array interconnect two pairs of circular shaped antenna which is fed by a corporate-feeding network. It was implemented as the cascade of quarter wave impedance transformers with characteristic impedances Z_{100}, $Z_{70,7}$ *and* Z_{50} to equally distribute the power [12].

Indeed, to provide a good impedance adaptation, the quarter wave impedance transformers widths are given by the following equation [16].

$$Z_2=\sqrt{Z_3 \times Z_1}$$



To provide an equal power distribution and to adapt the impedance of patch elements, the quarter wave impedance transformers lines should be tapered.

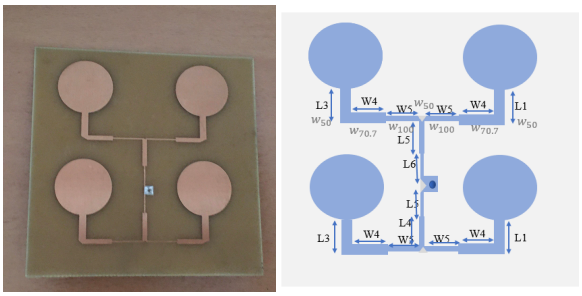


Figure 6: prototype of the proposed antenna array

The figure above describes the geometry of the proposed antenna array. In order to provide reasonable comparison, we kept the same substrate, the same geometry design and feeding network's dimensions [12] with changing the patch's shape. The dimensions of the antenna are illustrated below in millimeters (mm).

3.2. Discussions and results

Figure 7 shows a good agreement between the simulated and measured graphs except minor differences is observed. In fact, both designs deliver the minimum S11 of -30 dB. Indeed, obtain results for the antenna design reveals that the bandwidth of the fabricated antenna is shifted to 200 MHz different to the simulated bandwidth, it reaches 100MHz.

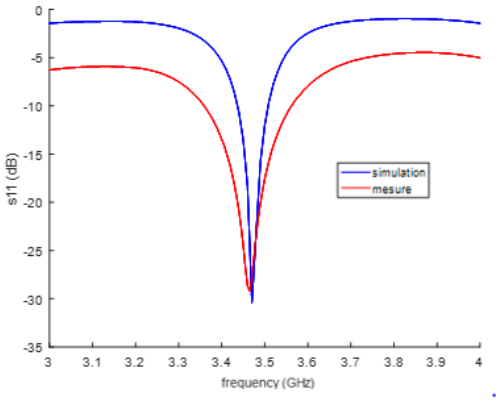


Figure 7: Comparative of simulated and measured S11 plot

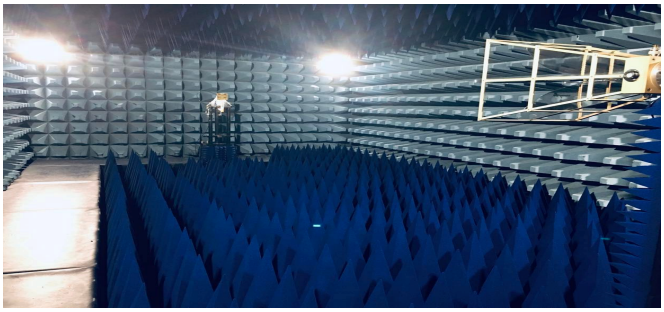


Figure 8: Radiation pattern Measurement

In order to check the electromagnetic field levels emitted by the antenna and to achieve more reproducible measurements, two antennas are fixed in the anechoic chamber. Whereas the first one is the reference antenna which is a QRH18 (QUAD RIDGEDHORN), the other antenna is dedicated to testing antenna in both E-plane and H-plane.

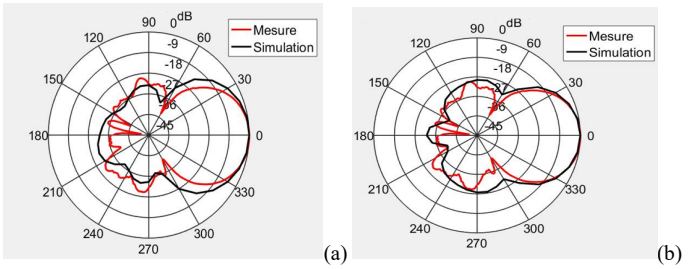


Figure 9: Comparison of the simulated and measured radiation pattern at 3.48GHz; (a) **φ**=90 (b) **φ**=0

The radiation pattern responses for both simulated and measured antenna are shown in figure 9. It can be clearly deduced that the prototype antenna is similar to simulation result. Compared to the single element antenna, the array antenna has the narrower beamwidth. In fact, the antenna array presents a directive radiation pattern. The bandwidth has reached to 40° as predicted in theory. It is also noticed that gain increases as the number of elements goes up. Therefore, the gain is around 5.8114 dB.

Unlike [12] our array antenna with 2 pairs of circular shapes provides a good gain, circular polarization and narrow beamwidth which is suitable for MIMO systems.

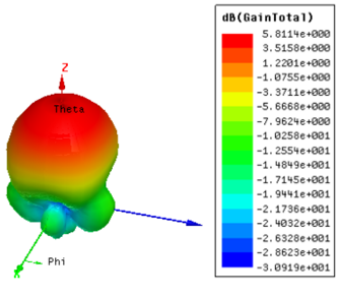


Figure 10: Simulated gain at 3.48GHz.

4. Conception of SISO

The USRP is a radio system conceived for implementation and rapid prototyping systems. [1] Selected because of its flexibility power consumption, low cost compared to other market options, reconfigurability, and high level of customization and high performance [1], [17], the USRP B210 has proven to be among the best choice for our work. Containing an RF front end complete with DACs and ADCs. The SDR attached to computer makes a full SDR system which enables the rest of the signal processing to be within MATLAB/Simulink.

4.1. Experimental setup

For our system prototype, due to the vast variety of supporting platforms, a USRP B210 is used, which provide fully-coherent MIMO 2x2 capability and low cost.

The transmission, reception, and processing of information testbed was performed in IRSEEM laboratory with no supplementary objects between Tx and Rx. The transmitter is located at a distance of around 1m from the receiver. The transmission was done in 3.48 GHz with using the USRP B210 devices which support RF frequencies from 70MHz to 6 GHz frequency. This boards are interconnected to our antennas array prototype via SMA connectors while ensuring a distance of λ/2 between the different Rx and different Rx to provide linear independence between different channels. So, with a carrier of 3.48 GHz, λ is around 85cm, where λ is the wavelength.

On the emission side, we sent a sinusoidal signal of frequency 1Mhz as a data signal to the input of Simulink then with using script of MATLAB via USRP, this data is multiplied by the modulation index, and after adding a constant, the resulting signal is multiplied with a carrier signal of frequency 3.48 GHz.

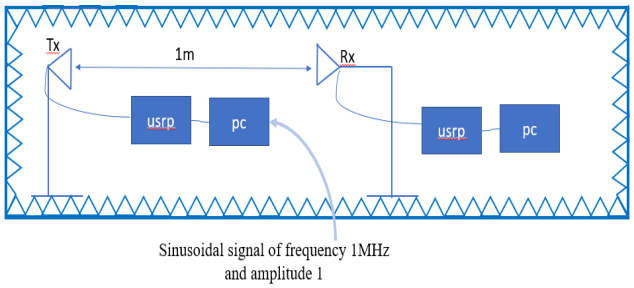


Figure 11: SISO Prototype

To create SDR MIMO applications, USRP B210, host computer and pairs of antenna arrays complete the hardware part of the SDR applications, The host computer is able to configure the USRP using software (MATLAB/Simulink), it enables users to transmit data using AM Modulation. Since, Simulink supports Universal Hardware Driver, it allows to control several parameters, such sample rate, gain, master clock, central frequency, interpolation factor, decimation factor…

So, first, we implement an AM transmitter based on sink blocks and standard blocks of Simulink to configure gain, center frequency, interpolation, filter and signal data [18] Simulink-USRP B210 enables the user to treat data in real time from USRP B210 and Simulink. Once the signal has been modulated, a spectrum analyzer called Agilent 4404B was interconnected to Tx (output of USRP B210) for displaying a signal in the frequency domain and check transmission problems.



Figure 12: Implementation of SISO

4.2. AM transmitter

This model explains the basic concept/model of amplitude modulation using the USRP board with Simulink, in fact, our antenna array is interconnected to USRP to assure transmission of the signal. The USRP is interfaced also with the pc/ Simulink with an interface called Communications Toolbox supports to transmit and receive RF signals in real time, enabling the use of MATLAB and Simulink to configure radio parameters, design algorithms, generate waveforms and measure and analyze signals.

It is known that to transmit a signal by amplitude modulation, it is necessary to use a carrier at a frequency much higher which is in our case $f_p = 3.48\text{GHz}$ than the data signal.

The amplitude modulation is done simply by the equation bellow

$$s(t) = A_c\ \mu m(t)\cos(2\pi f_c t) + A_c\cos(2\pi f_c t) \qquad (4)$$

To build this equation with Simulink, the first part is modelled with Simulink but the second part is done with a block called USRP transmitter which allows to configure the USRP (gain, carrier frequency, interpolation factor).

In our case, we desire transmit a sinusoidal signal with a frequency $f_m = 1MHz$. This signal must be converted into analog form, at the input of the USRP. therefore, the operating frequency of the DAC, or Master Clock Rate must be adjusted, and the interpolation rate operated by the Digital Up Converter above. The following values are possible for these parameters:

* Master Clock Rate: between 5 and 56 MHz

* Interpolation: an integer between 1 and 128, an even integer between 128 and 256, an integer divisible by 4 between 256 and 512.

Master clock rate (Hz)= Baseband sample rate Interpolation factor:

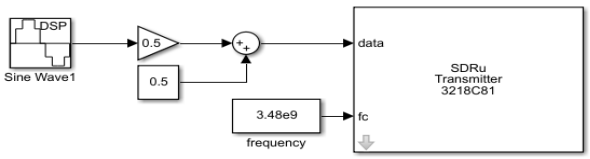


Figure 13: Transmitter model

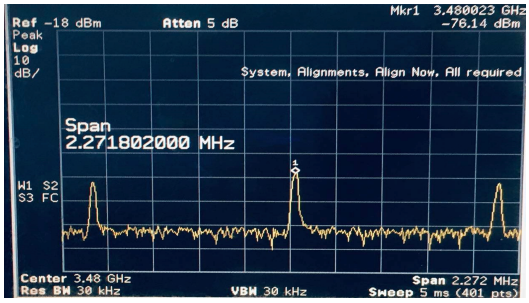


Figure 14: Spectrum

To verify the transmission of this sinusoidal signal, a spectrum analyzer in connected to the output of the USRP B210. When the carrier wave is modulated, both sides of the stable carrier signal sidebands are formed which carry the actual information in the modulation [12]. This can be illustrated by Figure 14 when a strong carrier is on the carrier frequency (3.48Ghz) while the sidebands have a frequency on 3.48Ghz $\pm 1MHZ$.

As expected, the spectrum has the usual presentation: two sidebands are symmetrical on 3.481GHz and 3.479GHz and a powerful carrier signal is in the center. So, the data signal has been successfully transmitted.

4.3. AM Receiver

After finalizing the emission, another patch antenna array 2X2 is used as receiver. For the demodulation part, we use an additional USRP device in order to demodulate the transmitted message using the receiver model.

Next, an overview of the AM demodulator is presented, which starts with the USRP receiver block. Then the demodulated signal can be passed to low-pass filter [17]. In fact, the low-pass filter, is used to remove the high-frequency signals and enable low-frequency signals to pass [15],[17].

As before, the parameters of a block calling SDRu receiver must be programmed in order to recompose the modulating signal. In fact, we have kept the carrier frequency at 3.48. So, we will find two frequencies, the higher frequency will be eliminated by the low pass filter and the other one will allow us to extract the data.

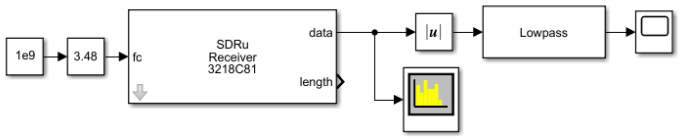


Figure 15: Receiver model

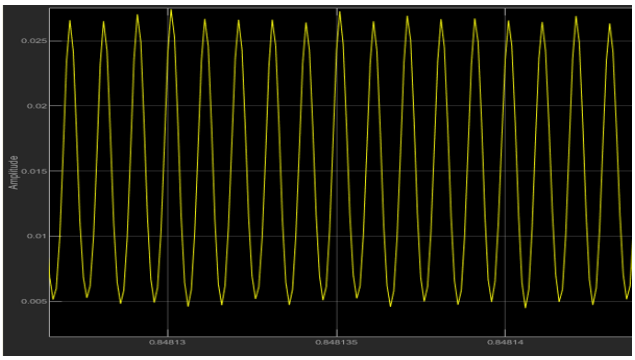


Figure 16: Received signal

The sinusoidal signal has been recuperated but there are fading due to the loss of system.

5. Conception of MIMO 2x2

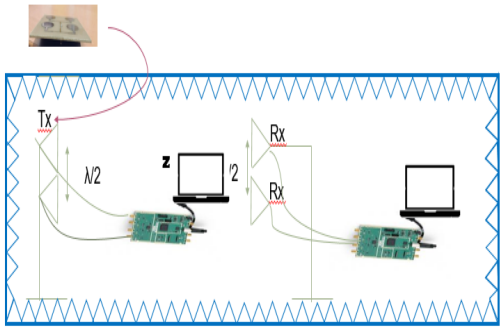
The USRP B210 board is mainly designed using the following components :

- AD9361 transceiver;
- Xilinx Spartan 6 FPGA;
- The USRP toolbox.

The key component of the first section is the AD9361 transceiver. This block is responsible for using two receive and also two transmit channels for RF front-end processing. The transceiver contains zero-IF-to-RF conversion and feedback, ADC/DAC gain, decimation/interpolation and filtering.

To connect the transceiver to the FPGA a digital data interface is used and to transmit data between them the I\Q way is employed.

The second component describes the Spartan 6 and presents software host processing. It uses both signal channels due to combining 2RX and 2TX chains, so, it is definitely feasible to conceive 2X2 MIMO System based on SDR. For this effect, the user could program a specific modulator/demodulator script. While the final part enables processing



(a)



(b)

Figure 17: MIMO 2X2 System (a) illustrative schematic, (b)implementation

As second testbed, we propose a novel Simulink model for coherent 2x2 MIMO, interconnecting each USRP B210 with pairs of antenna array using SMA connectors at both in transmission and reception.

The Full digital MIMO 2x2 system is operating in band C where the digital signal processing is processed on baseband.

5.1. Transmitter

A data matrix received by the SDRu transmitter block will enables the signal waveforms for MIMO mode. Indeed, to use both channels, Channel mapping must be adjusted [1, 2]. And a block called matrix concatenation used as gateway between the block sdru transmitter and the data elements arranged in columns or rows.

The MIMO in this support package receives signals through the RX port and transmits signals through the TX/RX port.

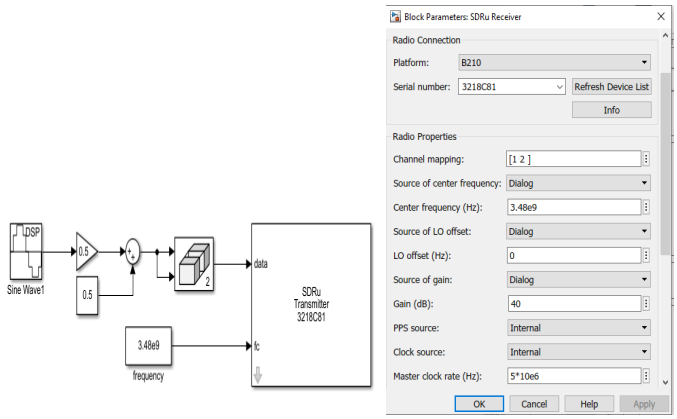


Figure 18: Programming USRP

5.2. AM Receiver

The USRP B210 below describes the hardware implementation framework for MIMO 2x2 model. In fact, for the reception part, the same model of SISO (AM receiver) has been kept, only a slight change with adding a block called select column in order to separate data and to treat each channel separately and simultaneously.

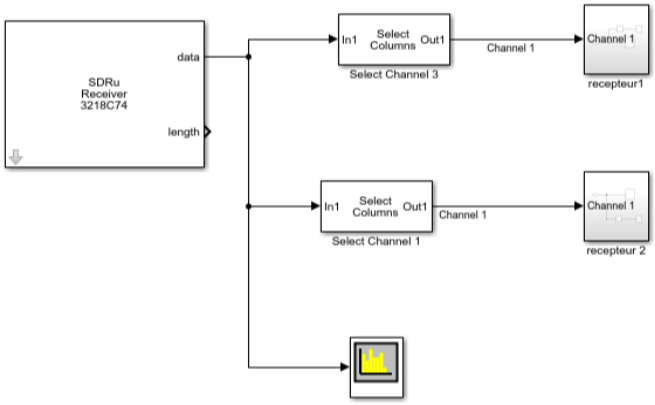


Figure 19: Receiving simultaneously

5.3-Implementation issues

In order to provide synchronization between both Tx antennas array and both Rx antennas array and to achieve interoperability between cards, The 2 Rx should operate synchronously, we send the sinusoidal signal from the first 2x2 antenna array and the same data from the second prototype antenna. It is very important to adjust the different parameters of both boards in a consistent way such as same center frequency value which is 3.48GHz, master clock, channel mapping, interpolation [16].

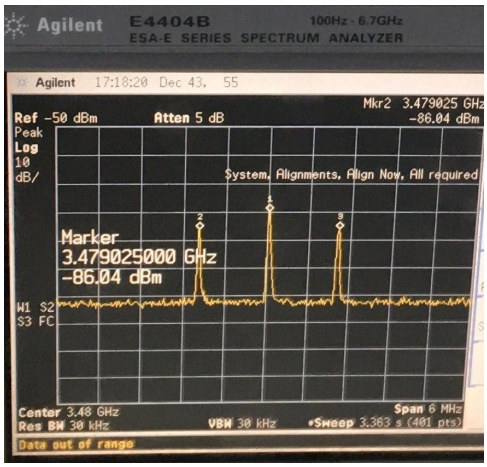


Figure 20: Frequency Spectrum of AM

The spectrum analyzer is connected to the output of the TX/RX port. As before, a strong carrier signal is centered at 3.48GHz, and the two sidebands are symmetrical at 3.481GHZ and 3.479GHz and the information signal is at 1MHz as shown in figure 19. So, MIMO transmission is well adjusted.

The sinusoidal signal with frequency of 1MHz have been usefully recovered. So, MIMO provide better efficiency specially in communications system.

In this work, performance of both MIMO and SISO systems is displayed. In fact, by transmitting the same data over multiple channels, MIMO introduce redundancy and diversity in data transmission that conventional single antenna (SISO) configurations cannot provide. This gives MIMO systems several advantages over conventional SISO configurations such the global throughput can be improved, allowing more quality and quantity data to be sent over the network, and using multiple data flow, problems such as fading caused by lost or dropped data packets can be reduced, this will help to recover all of the resulting data [16].

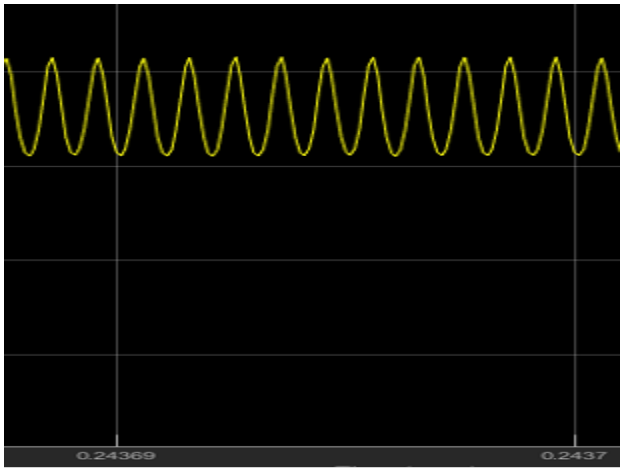


Figure 21: Received signal

6. Conclusion

In this study, a 5G-oriented MIMO prototype system has been designed and implemented equipped with four patch array antenna based on software-defined radios. The model for Amplitude modulation/demodulation has been developed combining USRP and Simulink in both SISO and MIMO Systems, with proving that MIMO is more efficiency than SISO. Moreover, all the analog and digital modules were properly developed and described enabling us to obtain a Full digital 2×2 MIMO system operating in band C. Finally, this system demonstrates great flexibility and scalability to be used in a wide range of applications. In the future MIMO beamforming system is investigated.

References

[1] S.Zhong, H.Feng, P.Zhang, J.Xu , H.Luo, J.Zhang, T.Yuan, L.Huang "Deep Learning Based Antenna Selection for MIMO SDR System," Sensors , 20, 6987, 2020, doi:10.3390/s20236987.

[2] Q. Wang, N. Mu, L. Wang, S. Safavi-Naeini, J. Liu, "5G MIMO conformal microstrip antenna design," Wireless Communications and Mobile Computing. 2017, 1–11, 2017, doi:10.1155/2017/7616825.

[3] S.Zhong, S. H.Feng, P.Zhang, J.Xu, L.Huang, T.Yuan, Y.Huo, "User Oriented Transmit Antenna Selection in Massive Multi-User MIMO SDR Systems," Sensors, 20, 4867, 2020, doi:10.3390/s20174867.

[4] M.Ali and M.Ibrahim, "Design Broadband MIMO Antenna to meet 5G communication systems requirements," in 2021 IEEE 1st International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering MI-STA, 838-842, 2021,doi: 10.1109/MI-STA52233.2021.9464429.

[5] G.K.D. Prasanna Venkatesan, T.Ranjitha, C. Shimona Neethi, S.Swetha, "Optimisation of mimo antenna for 5G applications," International Journal of Advances in Scientific Research and Engineering (ijasre), 4, 133-139, 2018,doi:10.7324/IJASRE.2018.32642.

[6] R. Gandhiraj, Ranjini Ram, K. P. Soman, "Analog and digital modulation toolkit for software defined radio", Procedia Eng, 30, 1155–1162,2012), doi: 10.1016/j.proeng.2012.01.975.

[7] G. N. Kareem, G. A. Gbotoso, S. O. Omogoye, "MATLAB analysis and Simulink model for amplitude modulation technique", World Journal of Advanced Engineering Technology and Sciences, 2, 21-28, 202, doi: 10.30574/wjaets.2021.2.2.0035.

[8] O.Hiari, R.Mesleh, "A Reconfigurable SDR Transmitter Platform Architecture for Space Modulation MIMOTechniques," IEEE Access, 5, 24214–24228, 2017,doi:10.1109/ACCESS.2017.2761859.

[9] D.M. Molla, H. Badis, L. George, M. Berbineau, "Software Defined Radio Platforms for Wireless Technologies", in IEEE Access, 10, 26203-26229, 2022, doi: 10.1109/ACCESS.2022.3154364

[10] M. Danneberg , I.Gaspar, M.Matthé, D.Zhang, L.Leonel Mendes,G. P. Fettweis,"Implementation of a 2 by 2 GFDM transceiver for robust 5G networks", 2015 International Symposium on Wireless Communication Systems (ISWCS), 236-240,2015,doi: 10.1109/ISWCS.2015.7454336.

[11] T. F. A. Nayna, A. K. M. Baki and F. Ahmed, "Comparative study of rectangular and circular microstrip patch antennas in X band", 2014 International Conference on Electrical Engineering and Information & Communication Technology,1-5, 2014, doi: 10.1109/ICEEICT.2014.6919142

[12] W. Chen, Y. Lin, "Design of 2X2 Microstrip Patch Patch Array Antenna for 5G C-Band Access Point Applications", 2018 IEEE International Workshop on Electromagnetics Applications and Student Innovation Competition (iWEM), 1-2, 2018, doi:10.1109/iWEM.2018.8536673.

[13] M. Drissi, N. Benjelloun, P. Descamps, A. Gharsallah, "Multilayer antenna dedicated to MIMO Beamformig antenna for 5G telecommunication applications", 2021 IEEE Conference on Antenna Measurements & Applications (CAMA) ,183-188, 2021, doi:10.1109/CAMA49227.2021.9703463

[14] C.A. Balanis, "Antenna Theory, Analysis and Design", John Wiley & Sons, New York, 1997.

[15] L.B. Laskov, V. M. Georgieva, "Analysis of Amplitude Modulation and Demodulation in MATLAB Simulink Environment". 2021 56th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), 223-226, 2021, doi: 10.1109/ICEST52640.2021.9483470

[16] G. Fokin, D. Volgushev, A. Kireev, D. Bulanov and V. Lavrukhin, "Designing the MIMO SDR-based LPD transceiver for long-range robot control applications",2014 6th International Congress on Ultra-Modern Telecommunications and Control Systems and Workshops (ICUMT), 456-461,2014, doi:10.1109/ICUMT.2014.7002144.

[17] C.Gavrila, V.Popescu, M.Alexandru, M. Murroni, C.Sacchi, "An SDR-Based Satellite Gateway for Internet of Remote Things (IoRT) Applications",IEEE Access, 8, 115423-115436, 2020, Doi: 10.1109/ACCESS.2020.3004480.

[18] T.Liuski," AM- ja DSB-modulaatioiden toteuttaminen Simulink-ohjelmistolla USRP- ohjelmistoradioalustalle ,phd thesis, University of Oulu, 2019.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 3, 174-181 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

A Constrained Intelligent Nonlinear Control Method for Redundant Robotic Manipulators

Dinh Manh Hung, Dang Xuan Ba*

Department of Automatic Control, HCMC University of Technology and Education (HCMUTE), Hochiminh City, 71300, Vietnam

ARTICLE INFO

Article history:
Received: 04 April, 2022
Accepted: 23 June, 2022
Online: 26 June, 2022

Keywords:
Redundant Robot
Intelligent Controller
Inverse Kinematics

ABSTRACT

Redundant robotic systems provide great challenges in solving kinematics and control problems, that are yet also open opportunities for exploring new, diverse and intelligent ideas and methods. In this paper, an advanced control method is proposed for position control problems of redundant robots with output constraints. The controller is structured with two control layers. In the high-level control layer, a cost function is first synthesized from the main control objective under constraint conditions. Virtual control signals are then reckoned to optimize the cost function using a soft Momentum-Levenberg-Marquardt approach. To realize the high-level control command, a nonlinear control signal is employed in the low-level control layer throughout a new nonsingular terminal sliding mode control structure. Comparative simulation results verified on a 7-DOF robotic arm model confirmed the effectiveness of the proposed control algorithm.

1 Introduction

Dealing with the limitations of high-order kinematic redundancy robotic models by advanced control methods has always been focuses of research interests for many years. However, solving the inverse kinematics of these flexible models is not inherently simple, making it even more difficult for models with many degrees of freedom [1], [2]. The same facing issues could be observed in robotic control fields.

To solve the inverse-kinematics problems, many interesting methods have been developed. In [2], a genetic algorithm was employed to minimize both end-effector position errors and joint displacements. Promising control results were obtained, but this algorithm took a long time to search the optimal values. In [3], an intelligent controller was developed for a redundant robot using a null-space approach and Bayesian networks. Mao et al. [4] utilized properties of a multi-layered neural network that could form any continuous nonlinear mapping from one domain to another to design an inverse-kinematics neural network to avoid obstacles and solve different solutions. Implementation time of the biomimetic algorithm [3] was impressively fast while the universal ability was exhibited by the neural approach [4]. The inverse-kinematics problem could be also well treated by various optimization methods such as the stretched simulated annealing (SSA) algorithm [5], and Jacobian pseudo inverse method [6], or the damped least-squares (DLS) law [7]. However, physical constraints are open issues of these advanced controllers. To cope with these constraint problems, quadratic programming (QP) solutions were derived using optimization algorithms [8]–[9]. By using generic QP solver, it leads to high computation cost and limit the real-time applicability [10]. Another direction for such the constraint problem is the use of null-space constraint remedies [11]. Priority-task vectors could be adopted to specify the task sequence and saturation functions were used as barriers of the constraint violations [12]. Outstanding control performances were produced but the physical constraints were not strictly consolidated in a smooth manner by using these advanced controllers.

To realize high-level control commands, a vast of controllers could be employed such as linear controllers [13], [14] and nonlinear controllers [15], [16]. Between them, sliding mode control (SMC) methods are favorite by researchers and developers thanks to their simplicity, robustness and acceptable working performances [17], [18]. Conventional SMC approaches however only result in infinite control errors [15], [19]. To further improve the control performances, terminal sliding mode control (TSMC) schemes were proposed and effectively applied for a plenty of robotic applications [19], [20]. Nevertheless, singularity problems were observed in the normal TSMC frameworks [21]. Such the weak issues were treated by a nonsingular terminal sliding mode control (NTSMC) methodologies in which the terminal power was taken into account for the time-derivative of the error signals instead of original ones [22]–[23]. However, the use of the NTSMC algorithms leads to dependence of the control signal directly on the time-derivative ones and it could activate vibration phenomena in noisy control cases

*Corresponding Author: Dang Xuan Ba, HCMC University of Technology and Education (HCMUTE), Hochiminh City, 71300, Vietnam & badx@hcmute.edu.vn

[24]–[25]. To deal with the NTSMC problem in a comprehensive fashion, new design of the NTSMC is required [22], [26], [27].

This paper is an extension of work originally presented in 2021 International Conference on System Science and Engineering (ICSSE) [1]. In this article, we demonstrate the versatility of the previously mentioned intelligent inverse kinematics solution through its successful application to a 7-DOF redundant robot. Structure of the proposed controller includes two control layers with the following contributions:

1. Robustness of the constrained intelligent inverse-kinematics algorithm in the high-level control layer is improved by the Momentum - Levenberg- Marquardt learning technique.

2. A new nonsingular nonlinear terminal sliding mode control method using a flexible power function is proposed for the low-level control layer to enhance the control performance of the closed-loop system.

3. Effectiveness of the proposed control method is verified by comparative simulation results on a 7-DOF robot model.

The outline of the paper is organized as follows. Section 2 discusses problem statements. The modified design of the intelligent inverse-kinematics algorithm is presented in Section 3. Section 4 shows a new NTSMC controller. The intensive simulation results are presented in Section 5. The paper is then concluded in Section 6.

2 Problem Statements

Forward kinematics of a general *n-dof* robotic manipulator could be obtained using homogenous-transformation computation [5], [6], [7]:

$$P_{ee} = f(\theta_1, \theta_2, ..., \theta_n) \tag{1}$$

where $P_{ee} \in R^{N\times 1}$ is the end-effector position, $\theta = [\theta_1, \theta_2, ..., \theta_n]$ is the vector of joint variables, N is the number of task-space variables, and f denotes the forward-kinematics computation.

Behaviors of the joint angles are presented by the following dynamics using the Euler-Lagrange method [28], [27]:

$$M(\theta)\ddot{\theta} + C(\theta, \dot{\theta}) + G(\theta) + F_r(\dot{\theta}) + d = \tau \tag{2}$$

where $\tau \in R^{n\times 1}$ is the vector of joint torques, $M \in R^{n\times n}$ is a symmetric positive-definite mass matrix, $C \in R^{n\times 1}$ is the Coriolis-centripetal vector, $G \in R^{n\times 1}$ is the gravitational vector, $F_r \in R^{n\times 1}$ is the Coulomb friction vector, and d stands for external disturbances.
Remark 1: Note that with a redundant robot, n is larger than N . We define a position error combining from a desired position P_{eed} and the system output P_{ee}. The main control objective here is to figure out proper control signals (τ) to drive the control error to zero complying with specific constraints. To this end, the main controller with a two-layer control structure is used. The complicated dynamics with redundant characteristics and unpredictable external disturbances are however main obstacles in developing the expected controller.

3 High-level Control Layer

The structure of the high-level control layer includes the inverse-kinematics solution stage, the constraint integration stage, and the optimization stage.

3.1 A Basic Inverse-Kinematics Method

The main control objective is formulated from the current and desired positions, as follows:

$$E_1 = 0.5\|P_{ee} - P_{eed}\|^2 \tag{3}$$

where $\|\bullet\|$ is the Euclidean norm of the term ($\bullet$).

By using the Jacobian-transpose-based method to minimize the high-level objective (2), virtual control signal is selected as follows [6], [29]:

$$\dot{\theta}_v = -\eta J^T(P_{ee} - P_{eed}) \tag{4}$$

where $J \in R^{N\times n}$ is the Jacobian matrix of the considering robot, and η is a positive learning constant.
Remark 2: In fact, there are infinite solutions for the problem (3). Hence, it is possible to shape possible solutions inside the expected regions [30].

3.2 Improvements for Joint Constraints

In this subsection, joint constraints are studied as additional control objectives. The former target (3) is modified as:

$$E_2 = 0.5\|P_{ee} - P_{eed}\|^2 + k_1^T \ln\left(\theta - \underline{\theta}\right) + k_2^T \ln\left(\bar{\theta} - \theta\right) \tag{5}$$

where $k_1, k_2 \in R^{n\times 1}$ are vectors of positive constants, and $(\bar{\theta}, \underline{\theta})$ are the upper and lower bounds of joint variables.

With the new objective (5), the virtual control signal (4) is updated as follows:

$$\dot{\theta}_v = -\eta\left(J^T(P_{ee} - P_{eed}) + \text{diag}\,(k_1)\,\frac{1}{\theta - \underline{\theta}} - \text{diag}\,(k_2)\,\frac{1}{\bar{\theta} - \theta}\right) \tag{6}$$

Remark 3: When the function *ln* is added to the expression (5), it interacts with the angular limit by increasing the value as it gets closer. This means that the virtual speed will also be subtracted a significant amount and has a deceleration effect. The parameter will adjust this interaction amount of the binding component *ln*.

However, the uncontrolled increase and decrease of this constraint component will adversely affect the real error update and cause large errors for the whole system when working near the hardware limit. The requirement here is that the additional constraint must be large enough to keep the robot within the allowable area and small enough to not affect the error update. Therefore, a new type of the parameter with great flexibility is needed to adjust the constraint up and down properly. Hence, we develop an alternative solution that makes the parameters automatically change under a nonlinear way:

$$\begin{cases} k_1 = e^{-0.5b_1|\theta - \underline{\theta}|^2} \\ k_2 = e^{-0.5b_2|\theta - \bar{\theta}|^2} \end{cases} \tag{7}$$

where $b_{i|i=2,3}$ are positive constants.

In this virtue, the control signal (6) is updated again as:

$$\begin{aligned}\dot{\theta}_v = &-\eta J^T(P_{ee} - P_{eed}) \\ &-\eta \text{diag}\,(k_1)\left(\frac{1}{\bar{\theta}-\theta} - b_1\left(\theta - \underline{\theta}\right)\ln\left(\theta - \underline{\theta}\right)\right) \\ &-\eta \text{diag}\,(k_2)\left(\frac{1}{\bar{\theta}-\theta} - b_2\left(\theta - \bar{\theta}\right)\ln\left(\bar{\theta} - \theta\right)\right)\end{aligned} \tag{8}$$

Remark 4: At this time, values of the constraint components will be adjusted up by the Gaussian-type gain when the joint variables approach the upper limits or the lower limits. This is caused by the characteristics of the Gaussian function [30], [31] as the variable values approach their centers. Conversely, when the joints move away from the upper and lower limits, depending on the slope of the Gaussian function, the gain will decrease very quickly and push the influence of the constraint down very small. This will satisfy the requirement set forth above.

3.3 Robust Constraint Inverse-Kinematics Approach

From (8), we have reformed the learning of the virtual control signal by using the Momentum-Levenberg–Marquardt method to make it faster while maintaining accuracy:

$$\begin{cases}\dot{\theta}_{v,i} = m\dot{\theta}_{v,i-1} - \frac{\eta L_b}{\alpha + \|L_b\|^2} \\ L_b = J^T(P_{ee} - P_{eed}) + \text{diag}\,(k_1)\left(\frac{1}{\bar{\theta}-\theta} - b_1\left(\theta - \underline{\theta}\right)\ln\left(\theta - \underline{\theta}\right)\right) \\ \quad +\text{diag}\,(k_2)\left(\frac{1}{\bar{\theta}-\theta} - b_2\left(\theta - \bar{\theta}\right)\ln\left(\bar{\theta} - \theta\right)\right)\end{cases} \tag{9}$$

where $(0 < m < 1)$ and α are positive constants, and $\bullet_i$, $\bullet_{i-1}$ denote current and previous states of $(\bullet)$.

Remark 5: The advantage of this learning algorithm is that it could solve the problems of the normal gradient descent learning methods which are slow to converge and easy to get stuck at the local minimum that is difficult or impossible to reach the global minimum [2], [4], [30]. The Momentum approach could help convergence faster by creating momentum behaviors from the previous velocity, while the shortcoming of instability of Momentum methods would be handled by the Levenberg–Marquardt part due to its ability to damp and ensure all the joint velocities not too different.

4 Finite-time Low-level Control Layer

The low-level control layer is a controller that realize the virtual commands (θ_v) generated by the high-level control one. To control the robot joint (θ) following the desired virtual control signal (θ_v) , we define a low-level tracking control error:

$$e = \theta - \theta_v \tag{10}$$

Inspired by terminal sliding mode control theories [17], [18], [19], we then synthesize an indirect control objective as:

$$\begin{cases} s = \dot{e} + \lambda f \\ f = \text{sgn}(e)|e|^{\frac{e^2+\gamma}{e^2+\beta}} \end{cases} \tag{11}$$

where λ and $(\gamma < \beta)$ are positive constants.

To realize the control mission (11) or (12) for the robotic system (10), we propose the following nonlinear terminal control signal:

$$\tau = (C + G + Fr) + M\left(\ddot{\theta}_v - \lambda\dot{f} - K_{dr}s^{\sigma} - K_{ro}\text{sgn}(s)\right) \tag{12}$$

where K_{dr}, K_{ro} and $(0 < \sigma < 1)$ are positive control gains.

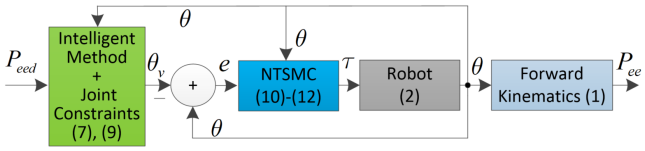


Figure 1: Block diagram of the proposed controller for redundant robots.

Remark 6: The NTSMC method can solve the problem of low settling time and push the error down very small because the larger the error leads to the steeper the selected sliding surface s will be (this is true for both the upper and lower regions of the setpoint); while the steeper the sliding surface provides the faster and more accurate the convergence. This is why the control signal can track the setpoint much better than a Proportional-Integral-Derivative (PID) controller [32], [31]. Block diagram of the proposed controller is presented in **Fig. 1**.

5 Validation Results

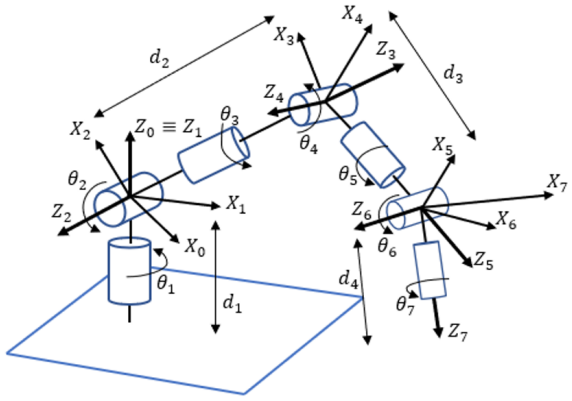


Figure 2: Configuration of a 7DOF robot for investigating.

The proposed controller was verified on a 7DOF redundant robot arm, whose detailed configuration is fully described in **Fig. 2**. The robot had three links (d_1, d_2, d_3) and seven joints $(\theta_{i|i=1..7})$. The desired trajectory of the robot was planned in the Cartesian coordinate system O_{xyz} . The end-effector position $P_{ee} = [x, y, z]^T$ of the robot is expressed as:

$$\begin{cases} x = d_1c_1s_2 - d_2(s_4(s_1s_3 - c_1c_2c_3) - c_1c_4s_2) - d_3(c_6(s_4(s_1s_3 - c_1c_2c_3) \\ \quad -c_1c_4s_2) + s_6(c_5(c_4(s_1s_3 - c_1c_2c_3) + c_1s_2s_4) + s_5(c_3s_1 + c_1c_2s_3))) \\ y = d_3(c_6(s_4(c_1s_3 + c_2c_3s_1) + c_4s_1s_2) + s_6(c_5(c_4(c_1s_3 + c_2c_3s_1) \\ \quad -s_1s_2s_4) + s_5(c_1c_3 - c_2s_1s_3))) \\ \quad +d_2(s_4(c_1s_3 + c_2c_3s_1) + c_4s_1s_2) + d_1s_1s_2 \\ z = d_2(c_2c_4 - c_3s_2s_4) - d_3(s_6(c_5(c_2s_4 + c_3c_4s_2) - s_2s_3s_5) \\ \quad -c_6(c_2c_4 - c_3s_2s_4)) + d_1c_2 \end{cases}$$

where $s_{i|i=1..7} := \sin(\theta_i)$ and $c_{i|i=1..7} := \cos(\theta_i)$.

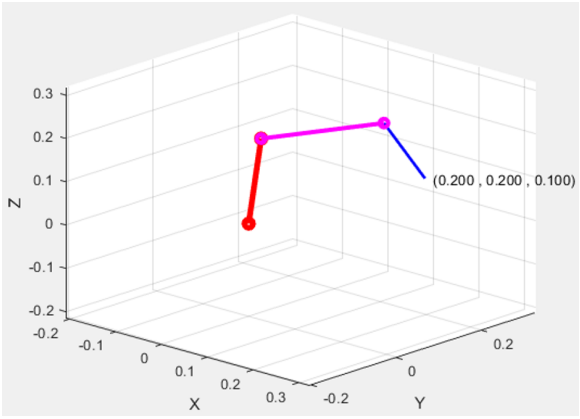


Figure 3: Structure of the 7DOF robot at working position.

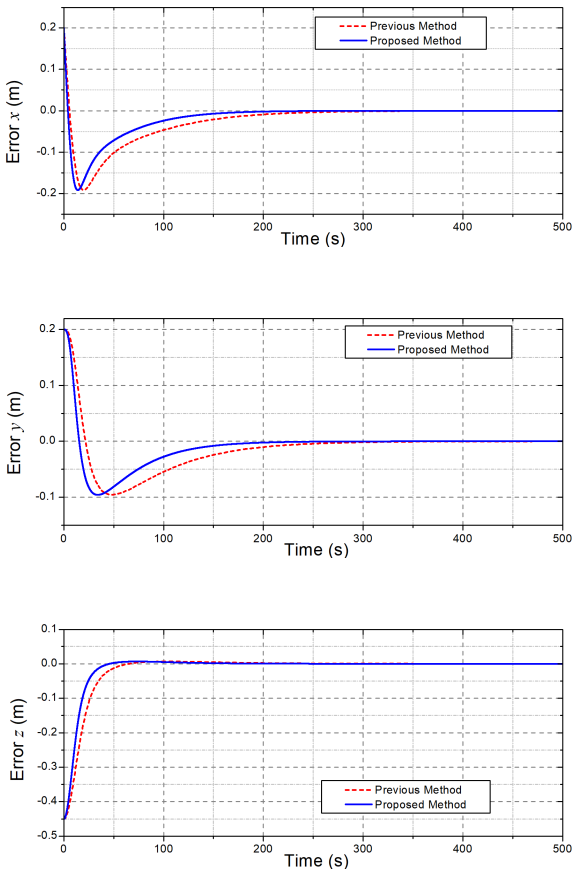


Figure 4: Comparative control errors of the end-effector position obtained by the controllers.

The parameters of the robot model for simulation were selected as follows: $m_1 = 0.5$ kg, $m_2 = 0.5$ kg, $m_3 = 0.5$ kg, $m_4 = 0.3$ kg, $m_5 = 0.3$ kg, $m_6 = 0.2$ kg, $m_7 = 0.2$ kg, $d_1 = d_2 = 0.2$ m, and $d_3 = 0.15$ m . The operating condition was free of interference and had a viscous friction force with a selected friction coefficient $\mu = 20$. The initial values of the joint angles were $\theta_{i|i=1..7} = 0$, and the initial end-effector position of the robot in the Cartesian coordinate was $(0; 0; 0.55)$ (m). The desired position of the end-effector chosen for testing was $(0.2; 0.2; 0.1)$ (m). To clearly evaluate the control performances of the proposed controller, a previous control method [1] was employed to realize the same control mission in the same system under the same testing conditions. Control gains of the previous controller were selected as $b_1 = b_2 = 0.0001, \eta = diag([0.4, 0.4, 0.5, 0.5, 0.6, 0.6, 0.6]), K_P = 200I_7, K_I = 50I_7, K_D = 25I_7$, while those of the proposed controller were manually tuned and obtained as $b_1 = b_2 = 0.0001, \eta = diag([0.4, 0.4, 0.5, 0.5, 0.6, 0.6, 0.6]), \alpha = 1, m = 0.9, \lambda = 10I_7, \gamma_{i|i=1..7} = 0.95, \beta_{i|i=1..7} = 1, \sigma_{i|i=1..7} = 0.9, K_{dr} = 100I_7, K_{ro} = 0.001I_7$.

In the first test, the two controllers were applied to control the robot from the initial position to the desired one with the same low-level control layer but with different high-level control ones. The response posture of the robot working under the proposed controller after the simulation is shown in **Fig. 3**. Control results obtained are shown in **Figs. 4 - 5**.

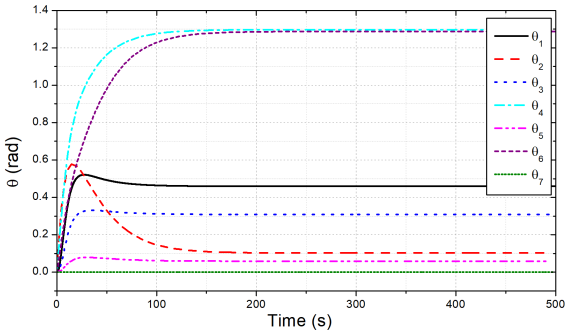


Figure 5: Reference joint angles generated by the proposed controller from the first test.

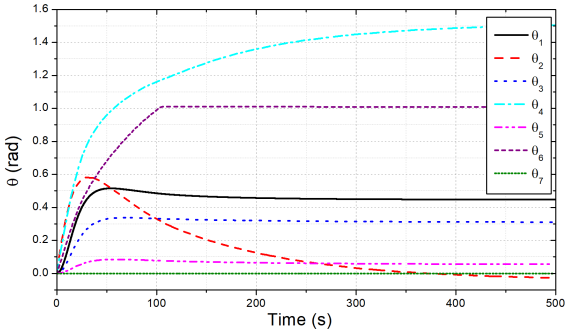


Figure 6: Desired profiles of the joint angles in the joint-constraint test.

The control errors of the end-effector position of the robot accomplished by the two controllers are compared in **Fig. 4**. Generally, the two control systems were stably working with excellent steady-state control errors of about $(6, 7.2, 4.5) \times 10^{-6}$(m) in the x, y, z directions of the end-effector positions, respectively. The figure also shows that the convergence time of the previous controller was about 310 (s) (the red-dot line) while that of the proposed controller was only about 203 (s) (the blue-solid line). The faster results came from the Momentum learning behaviors supported by the Levenberg–Marquardt adaptation scheme (8). Reference joint angles generated by the high-level control layer of the proposed controller are shown in **Fig. 5**. By combining **Figs. 2**, **3** and **5**, it can be easy to understand that to reach the desired end-effector

position, the joints 4 and 6 were hard-working parts with the largest variations (from 0 to 1.29) (rad).

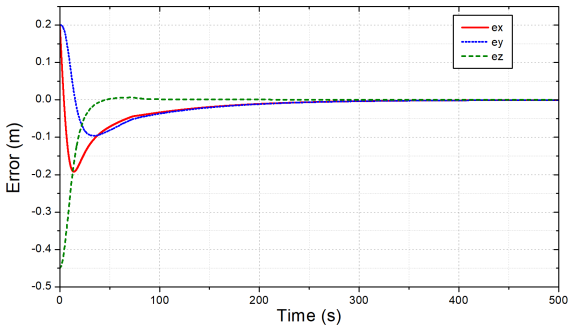


Figure 7: Errors of the end-effector in the joint-constraint test.

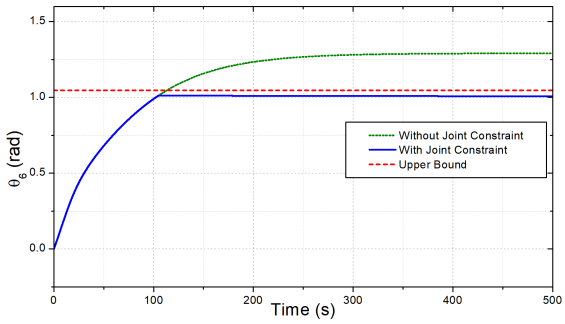


Figure 8: Comparative output at joint 6 in the first and second tests.

To assess the effectiveness of the joint-constraint feature proposed, in the second test, we limited the joint 6 into a range of $(-\pi/3 \leq \theta_6 \leq \pi/3)$. The simulation results achieved by the proposed controller for the old and new testing conditions are presented in **Figs. 6-8**. As seen in **Fig. 7**, the proposed controller still ensured the excellent control quality for the end-effector positions under the constrained working condition. Furthermore, as shown in **Fig. 6**, the profiles of all joint angles would also change to ensure that the data changes of θ_6 was still inside of the given range $(-\pi/3; \pi/3)$. The profiles of the joint variable θ_6 in this test and the last test are compared in **Fig. 8**. It can be seen that, in the case of the constraint defined, the angle θ_6 increased very fast to the upper limit and then stopped increasing and kept a certain distance with the upper limit because the joint velocity was greatly reduced caused by the algorithm. The results would also be similar for the other joints if other joint constraints were applied. Note that, thanks to the redundant robot configuration possessed, even though the joint angle θ_6 was limited, the control burden was shared by the other joint angles, especially by the joint θ_2 and θ_4, that could be observed in this test by comparing the data in **Figs. 5** and **6**.

In the third test, we performed a new simulation to show the importance of the low-level control layer by comparing the response quality between the new controllers with NTSMC and previous PID controllers [1]. The two controllers were used to control the end-effector of the robot to the desired position and their control results are compared in **Figs. 9-12**. For the easier observation, we only took the results of joints 2, 4, 6 because they are folded joints.

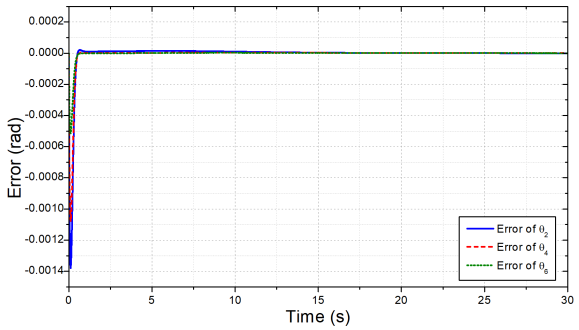


Figure 9: The joint control errors obtained by the new NTSMC controller.

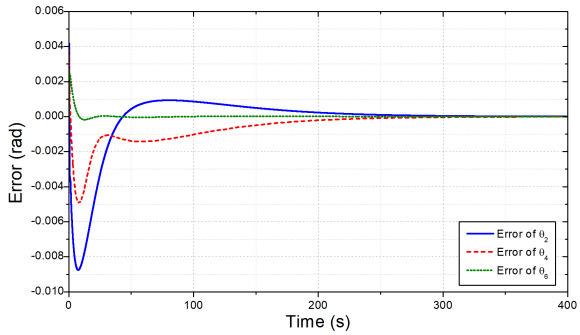


Figure 10: The joint control errors obtained by the previous controller.

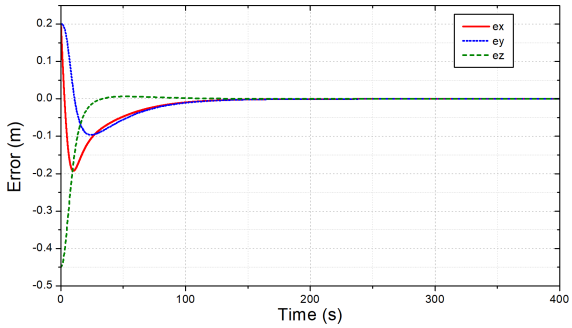


Figure 11: The end-effector control errors obtained the proposed controller.

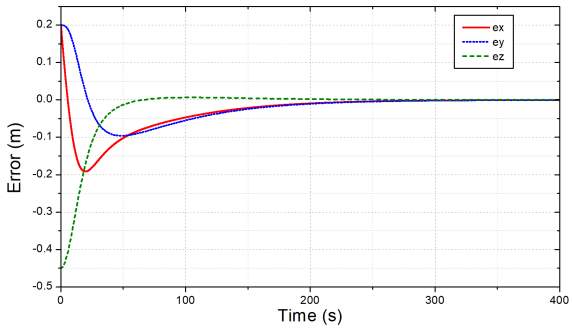


Figure 12: The end-effector control errors obtained the previous controller.

The results in **Figs. 9** and **10** indicate that the settling time and the transient response of the new controller were superior to those of the old one (ten times faster). Achieving a very small error ($\pm 10^{-4}$)

in a very short time has shown the effectiveness of the improved NTSMC method compared to the previous one. The other control results in **Figs. 11** and **12** also imply that with the improvements in the low-level control layer, it is not surprising about the faster response and higher accuracy in the end-effector control space obtained by the proposed controller as comparing to the previous one.

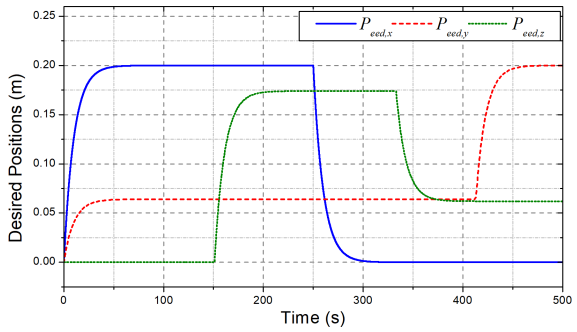


Figure 13: The desired end-effector positions in the fourth test.

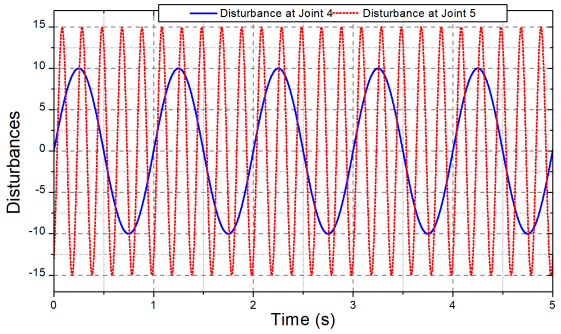
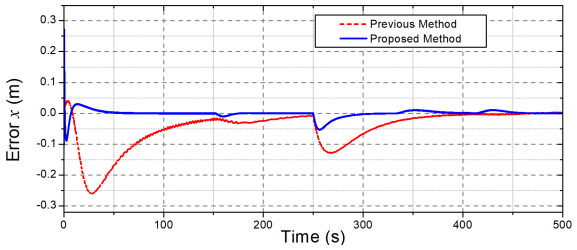
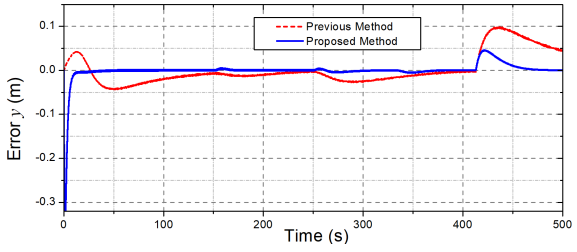


Figure 14: External disturbances in the fourth test.





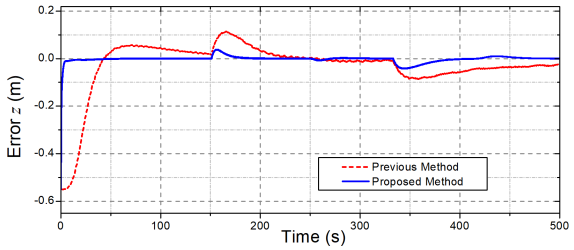
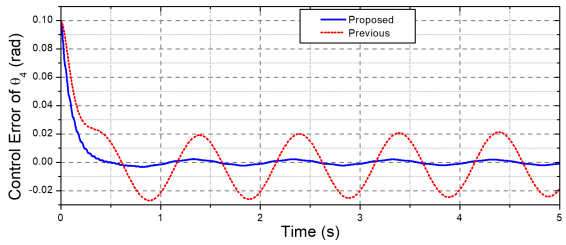


Figure 15: Comparative control errors at the end-effector of the robot.



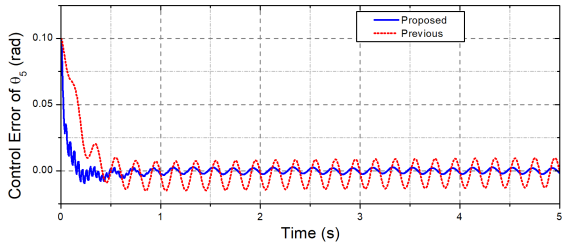


Figure 16: Comparative control errors of the robot joints.

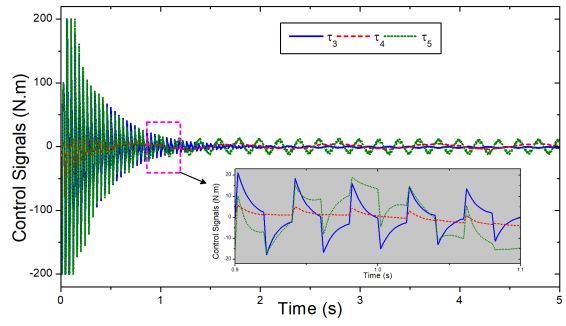


Figure 17: Control signals generated by the proposed controller at the robot joints.

To validate the feasibility of the proposed control approach, in the last simulation, the robot was challenged with new desired end-effector positions of multi-step signals, as depicted in **Fig. 13**, and external disturbances at joints 4 and 5 as presented in **Fig. 14**. The obtained control results of the proposed and previous controllers are illustrated in **Figs. 15 - 17**. Under the new testing conditions, as seen in **Fig. 15**, the designed controller still provided higher control accuracies and faster settling time at the end-effector than the previous one. To this end, the proposed control approach was not only employed the robust intelligent learning control law (7)-(9) but it was also supported by the new NTSMC framework (10)-(12).

Indeed, as shown in **Fig. 16**, the control errors of the new nonlinear low-level controller at joints 4 and 5 in the heavy disturbances were respectively 0.0018 (rad) and 0.002 (rad), while those of the previous one were 0.03 (rad) and 0.015 (rad). **Figure 17** shows the control signals generated by the proposed controller. The feasibility of the proposed control method could be confirmed throughout the data obtained.

6 Conclusions

In this paper, an intelligent two-layer control method for dealing with inverse-kinematics problems of redundant robots has been improved and applied to a 7-DOF robot. In the high-level control layer, the inverse kinematics problem is solved by using the Momentum-Levenberg optimization method. In cases of the joint constraint requirements, an advanced constrained learning feature can be activated to ensure that the robot joints can avoid physical limit collisions. To enhance the control performance of the overall system, a new nonlinear terminal sliding mode control framework is developed in the low-level control layer. The effectiveness of the proposed control algorithm has been consolidated by comparative validation results obtained. In the future, the intelligent method will be integrated more advanced, optimal, flexible working features and verified on a real-time system.

Conflict of Interest The authors declare no conflict of interest.

References

[1] D. M. Hung, D. T. Linh, D. X. Ba, "An Intelligent Control Method for Redundant Robotic Manipulators with Output Constraints," in 2021 International Conference on System Science and Engineering (ICSSE), 116–121, 2021, doi:10.1109/ICSSE52999.2021.9538490.

[2] A. C. Nearchou, "Solving the inverse kinematics problem of redundant robots operating in complex environments via a modified genetic algorithm," Mechanism and Machine Theory, **33**(3), 273–292, 1998, doi:10.1016/S0094-114X(97)00034-7.

[3] P. K. Artemiadis, P. T. Katsiaris, K. J. Kyriakopoulos, "A biomimetic approach to inverse kinematics for a redundant robot arm," Autonomous Robot, **29**, 293–308, 2010, doi:10.1007/s10514-010-9196-x.

[4] Z. Mao, T. C. Hsia, "Obstacle avoidance inverse kinematics solution of redundant robots by neural networks," Robotica, **15**, 3–10, 1997, doi:10.1109/ROBOT.1993.291813.

[5] P. Costa, J. Lima, A. I. Pereira, P. Costa, A. Pinto, "An Optimization Approach for the Inverse Kinematics of a Highly Redundant Robot," in Abraham, A., Wegrzyn-Wolska, K., Hassanien, A., Snasel, V., Alimi, A. (eds) Proceedings of the Second International Afro-European Conference for Industrial Advancement AECIA 2015. Advances in Intelligent Systems and Computing, volume 427, 433–442, 2015, doi:10.1007/978-3-319-29504-6_41.

[6] D. X. Ba, J. B. Bae, "A precise neural-disturbance learning control of constrained robotic manipulators," IEEE Access, **9**, 50381–50390, 2021, doi:10.1109/ACCESS.2021.3069229.

[7] D. S. Binh, K. L. Thanh, D. P. Nhien, D. X. Ba, "An intelligent inverse kinematic solution of universal 6-DOF robots," in Huang, Y. P., Wang, W. J., Quoc, H.A., Giang, L.H., Hung, N.L. (eds) Computational Intelligence Methods for Green Technology and Sustainable Development. GTSD 2020. Advances in Intelligent Systems and Computing, volume 1284, 107–116, 2020, doi:10.1007/978-3-030-62324-1_10.

[8] F. Cheng, T. Chen, Y. Sun, "Resolving manipulator redundancy under inequality constraints," IEEE Transactions on Robotics and Automation, **10**(1), 65–71, 1994, doi:10.1109/70.285587.

[9] A. Escande, N. Mansard, P. B. Wieber, "Fast resolution of hierarchized inverse kinematics with inequality constraints," in 2010 IEEE International Conference on Robotics and Automation, 3733–3738, 2010, doi:10.1109/ROBOT.2010.5509953.

[10] A. Escande, N. Mansard, P. B. Wieber, "Hierarchical quadratic programming: Fast online humanoid-robot motion generation," The International Journal of Robotics Research, **7**(33), 1006–1028, 2014, doi:10.1177/0278364914521306.

[11] F. Flacco, A. D. Luca, O. Khatib, "Prioritized multi-task motion control of redundant robots under hard joint constraints," in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 3970–3977, 2012, doi:10.1109/IROS.2012.6385619.

[12] F. Flacco, A. D. Luca, O. Khatib, "Control of redundant robots under hard joint constraints: saturation in the null space," IEEE Transactions on Robotics, **31**(33), 637–654, 2015, doi:10.1109/TRO.2015.2418582.

[13] V. I. Utkin, "Sliding mode control design principles and applications to electric drives," IEEE Transactions on Industrial Electronics, **40**(1), 23–36, 1993, doi:10.1109/41.184818.

[14] J. Zhang, W. X. Zheng, "Design of adaptive sliding mode controllers for linear systems via output feedback," IEEE Transactions on Industrial Electronics, **61**(7), 3553–3562, 2014, doi:10.1109/TIE.2013.2281161.

[15] D. X. Ba, "A Fast Adaptive Time-delay-estimation Sliding Mode Control for Robot Manipulators," Advances in Science, Technology and Engineering Systems Journal, **5**(6), 904–911, 2020, doi:10.25046/aj0506107.

[16] C. Hu, B. Yao, Q. Wang, "Performance-oriented adaptive robust control of a class of nonlinear systems proceeded by unknown dead zone with comparative experimental results," IEEE/ASME Transactions on Mechatronics, **18**(1), 178–189, 2013, doi:10.1109/TMECH.2011.2162633.

[17] Z. H. Man, A. P. Paplinski, H. R. Wu, "A robust MIMO terminal sliding mode control scheme for rigid robot manipulator," IEEE Transactions on Automatic Control, **39**(12), 2464–2469, 1994, doi:10.1109/9.362847.

[18] L. Fridman, J. Moreno, R. Iriarte, Sliding Modes after the first Decade of the 21st Century, Springer Berlin, Heidelberg, 2012, doi:10.1007/978-3-642-22164-4.

[19] S. T. Venkataraman, S. Gulati, "Control of nonlinear systems using terminal sliding modes," in 1992 American Control Conference, 891–893, 1992, doi:10.23919/ACC.1992.4792209.

[20] Y. Tang, "Terminal sliding mode control for rigid robots," Automatica, **34**(1), 51–56, 1998, doi:10.1109/9.362847.

[21] K. B. Park, J. J. Lee, "Comments on 'A robust MIMO terminal sliding mode control scheme for rigid robotic manipulators," IEEE Transactions on Automatic Control, **41**(5), 761–762, 1996, doi:10.1109/9.362847.

[22] D. X. Ba, H. Yeom, J. B. Bae, "A Direct Robust Nonsingular Terminal Sliding Mode Controller based on an Adaptive Time-delay Estimator for Servomotor Rigid Robots," Mechatronics, **59**, 82–94, 2019, doi:10.1016/j.mechatronics.2019.03.007.

[23] C. K. Lin, "Nonsingular terminal sliding mode control of robot manipulators using fuzzy wavelet networks," IEEE Transactions on Fuzzy Systems, **14**(6), 849–859, 2006, doi:10.1109/TFUZZ.2006.879982.

[24] Z. Ma, G. Sun, "Dual terminal sliding mode control design for rigid robotic manipulator," Journal of the Franklin Institute, **355**(18), 9127–9149, 2018, doi:10.1016/j.jfranklin.2017.01.034.

[25] R. M. Asl, Y. S. Hagh, R. Palm, "Robust control by adaptive nonsingular terminal sliding mode," Engineering Applications of Artificial Intelligence, **59**, 205–217, 2017, doi:10.1016/j.engappai.2017.01.005.

[26] Y. Shtessel, M. Taleb, F. Plestan, "A novel adaptive-gain supertwisting sliding mode controller: Methodology and application," Automatica, **48**(5), 759–769, 2012, doi:10.1016/j.automatica.2012.02.024.

[27] J. Baek, M. Jin, S. Han, "A new adaptive sliding mode control scheme for application to robot manipulators," IEEE Transactions on Industrial Electronics, **63**(6), 3628–3637, 2016, doi:10.1080/00207179308934387.

[28] Y. Feng, X. Yu, Z. Man, "Nonsingular terminal sliding mode control of rigid manipulators," Automatica, **38**(12), 2159–2167, 2002, doi:10.1016/S0005-1098(02)00147-4.

[29] O. Kanoun, F. Lamiraux, P. B. Wieber, "Kinematic control of redundant manipulators: Generalizing the task-priority framework to inequality task," IEEE Transactions on Robotics, **27**(4), 785–792, 2011, doi:10.1109/TRO.2011.2142450.

[30] A. Tringali, S. Cocuzza, "Globally Optimal Inverse Kinematics Method for a Redundant Robot Manipulator with Linear and Nonlinear Constraints," Robotics, **9**(3), 2020, doi:10.3390/robotics9030061.

[31] E. Krastev, "Sliding Mode Control of a Redundant Robot Arm in Motion Planning Subject to Joint Space Constraints," in 2017 International Conference on Control, Artificial Intelligence, Robotics and Optimization (ICCAIRO), 109–114, 2017, doi:10.1109/ICCAIRO.2017.31.

[32] C. M, W. Xu, C. Sun, "On switching manifold design for terminal sliding mode control," Journal of the Franklin Institute, **353**, 1553–1572, 2016, doi:10.1016/j.jfranklin.2016.02.014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 5, 178-192 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Association Rules for Knowledge Discovery From E-News Articles: A Review of Apriori and FP-Growth Algorithms

Thilini Lakshika*, Amitha Caldera

Department of Information Systems Engineering, University of Colombo School of Computing, Colombo, 00700, Sri Lanka

ARTICLE INFO

Article history:
Received: 15 July, 2022
Accepted: 12 October, 2022
Online: 31 October, 2022

Keywords:
Association Rule Mining
Apriori algorithm
FP-Growth algorithm
Frequent Pattern Mining
Knowledge Discovery

ABSTRACT

Owing to technological development, the internet has become the world's largest platform where an unaccountable amount of e-news information is freely available to use. Most of the time, e-newspaper readers have to examine the massive collection of e-news articles to locate necessary information relevant to them. Massive semi-structured and unstructured texts usually mislead the readers when they search and understand data for some knowledge. Furthermore, manually reading a collection of e-news articles for some knowledge is tedious and unproductive. The literature related to Knowledge Discovery from text documents has had a substantial improvement in this regard and Association Rule Extraction using text documents, in particular, has become a more frequent and imperative research approach to finding out the most significant information, patterns, and features in the text documents while diminishing the time for reading all the documents. This study provides a comprehensive review of Association Rule extraction using textual data covering the essential topics; Pre-processing, steps in Association Rule Mining, and rule mining algorithms. Out of the various existing association rule mining algorithms, the two most important algorithms, Apriori and FP Growth, are chosen for the experiment using e-news articles. Based on the experimental results, this study discusses the performance, significant bottlenecks, recent breakthroughs of rule mining algorithms, and finally the perspective directions to facilitate future research.

1. Introduction

Due to the rapid growth of web technologies and data repositories, 80% of the world's data is now available in electronic format [1,2]. The internet is the most common platform where an enormous amount of semi-structured and unstructured texts is freely available in reports, news articles, web pages, log files, data transaction files, and other types of electronic documents to satisfy human information desires. The escalating nature of the internet lets many documents first appear online before printing, which is a major reason for the availability of a valuable source of knowledge. Hence, information in the printed medium is slowly losing its gleam. In contrast, the information in the electronic medium continues to grow with the easy availability of quick and obsolete data to the world with only a few snaps.

Due to the overloaded information on news websites, it is tough and tedious for e-news readers to remember all important information to get some knowledge to satisfy their information desires [1–4]. Even searching the internet for exact details is time-consuming. The resultant information may confuse users to deal with it due to both relevant and irrelevant information. Extracting knowledge from e-news articles in semi-structured and unstructured formats becomes more arduous, but it is worthwhile to explore those because of the necessity of obtaining knowledge. But, manual analysis of bulky information available in e-news articles for effective extraction of useful information is not conceivable [3]. Similarly, manual knowledge discovery has a very little guarantee of generating enough details to support decision-making compared to automated knowledge discovery [3]. This issue leads e-news article readers to make decisions based on applications in knowledge discovery.

An automatic text analysis tool based on techniques such as Knowledge Discovery, called Text Mining or Text Data Mining from textual data sources, is the most demanding solution to address this problem [1, 2]. The higher availability of massive amounts of electronic documents has been a great motivation for

*Corresponding Author: Thilini Lakshika, Department of Information Systems Engineering, University of Colombo School of Computing, Colombo, 00700, Sri Lanka, +94713179029, tlv@ucsc.cmb.ac.lk

automatically introducing new knowledge acquisition approaches. Hence, the automatic extraction of knowledge from a large number of e-news articles available on the web using Text Mining (TM) concepts has become an increasingly important research topic [1-3].

Data Mining (DM) is the process of extracting, exploring, and analysing large blocks of information to gather meaningful patterns and trends that can consider knowledge in the data sources. Applying DM techniques to extract interesting and useful knowledge is challenging because of the information overloading, higher duplication rates, and ambiguities of data [5]. DM in the text is known as Text Mining (TM) and is generally used to extract meaningful patterns in a collection of text data using Natural Language Processing (NLP) techniques such as Information Extraction (IE) to transform unstructured data into a structured format [6, 7]. One of the main operations in TM is frequent itemset mining for identifying frequent patterns, hidden patterns, themes, and the context in large datasets that are very easy to understand and interpret by data analysts and normal users [1–5]. Frequent itemset mining plays an essential role in many data mining tasks. Algorithms such as Apriori and FP-Growth are the more frequently used Association Rule Mining (ARM) algorithms for mining sequential and emerging patterns.

Researchers typically use data mining techniques such as classification, clustering, regression analysis, and ARM to discover knowledge from massive data sets by applying novel experimental approaches. TM is a precious and beneficial process for organizations with large textual datasets as it supports expanding the decision-making process of organizations, leading to better business outcomes. However, the richness and vagueness of natural language is the most complex challenge in applying TM techniques [2].

So, the research questions (RQs) that we aim to cover with the study are:

- RQ1: What tasks are currently being solved with association rules using textual data?
- RQ2: What areas of applications currently have been addressed with association rules using textual data?
- RQ3: Which pre-processing steps need to be applied for ARM using textual data?
- RQ4: Which ARM algorithm: Apriori or FP Growth has fast execution in association rule generation using e-news articles?
- RQ5: What are the current trends and future problems to be faced by Apriori and FP Growth algorithms in textual data?

The main contributions of this study are,

- Discuss the diverse areas of ARM along with its applications and motivate research on TM.
- A detailed discussion of the main phases in ARM including the most important text pre-processing steps.
- An experiment conducted using the major rule generation algorithms including Apriori and FP-Growth, their usage, performance, and limitations in TM using a collection of e-news articles.
- Discuss the Apriori and FP-Growth algorithm perspective directions to facilitate future research in TM.

The organization of this paper is as follows. Overview in section 2 explores the fundamental ideas related to ARM. Further elaborations on the role of Association Rules, including Apriori and FP-Growth algorithms in Text applications, are presented in section 3. The significant phases in ARM are discussed in section 4. Our experimental application of Apriori and FP Growth algorithms for association rule generation using e-news articles is described in section 5. Section 6 includes the results and discussion, and the directions for future research are presented in section 7. Finally, section 8 concludes the research work.

2. Overview

2.1. Introduction to Data Mining

The automatic knowledge discovery from text documents using TM concepts has become an emerging research topic [1-3]. Extracting, exploring, and analyzing information to discover meaningful patterns and trends in data is the main focus of DM approaches. Even though the DM supports different techniques such as ARM, frequent itemset mining, and pattern mining, applying those techniques for the extraction of interesting and useful knowledge is challenging because of the information overloading, higher duplication rates, and ambiguities of data [5]. The most common issues in DM are handling many patterns, finding relationships between keywords, and generating knowledge using keyword relationships. Typically, DM techniques discover a rich set of patterns that can consider knowledge in the data sources. However, effective analysis of such discovered patterns for making decisions becomes more complicated due to the meaningless patterns and noises present in the discovered patterns. Even though data mining techniques directly deal with structured data, they may be adapted to mine unstructured text data that presents unique characteristics [2].

2.2. Introduction to Text Mining

Text Mining or Document Mining is an emerging field of research due to the need to retrieve important non-trivial information or knowledge and features by transforming a collection of unstructured text data into a structured format [3, 5, 6]. Data Mining using textual data, also known as Text Mining and is generally used to identify meaningful patterns in a collection of text data using Natural Language Processing (NLP) techniques such as Information Extraction (IE) to transform unstructured data into a structured format [6,7]. TM is generally used to extract semantic logic in the textual data while DM is used to discover novel insights and hidden patterns in text data. Even though many DM tools and applications are designed to work on structured data from databases, many TM tools and applications are designed to work with both unstructured and semi-structured databases such as full-text documents, emails, and HTML files. The lack of explicit structure in the unstructured text data

increases the difficulties of discovering the documents' implicit knowledge. Natural language may express the same concept in multiple ways, making it harder to extract and represent the conceptual concepts in the natural text [1].

Hence, text data mining approaches are still attracting more attention in designing novel TM models to retrieve the users' expected data [6]. Association rule discovery is one of the main techniques in TM, and existing rule discovery approaches are categorized as sequential patterns, maximal patterns, and closed patterns [5]. The various applications such as information extraction, summarization, document classification, and document clustering are the more commonly used applications in the research field of text data analysis.

2.3. Basic Text Pre-processing steps

Generally, e-news articles comprise information in an unstructured format that cannot be straightly used for further processing and cause inaccurate results in text mining [3]. Thus, mining extensive e-news article collections in unstructured text format to extract valuable patterns requires applying essential pre-processing steps and converting the information into a more appropriate data format than keeping the information in a plain text file [3]. The required pre-processing steps for the text data are more important before- applying any DM techniques. Text pre-processing steps typically compromise transformation, tokenization, filtration of keywords, and stemming and indexing of the keywords [2,3].

Transformation: Information on the web presents in different document formats and structures, which requires converting them into Extensible Markup Language (XML) format that is amenable to further processing [1,2].

Tokenization: The unstructured text formats in e-news articles cause inaccurate results in text mining [3] and are required to convert them into structured formats and proceed with only valid tokens by removing inappropriate information such as special characters, parentheses, commas, etc. Tokenization plays a significant role in increasing accuracy in knowledge discovery.

Filtration: Finding out the relationship between all the keywords in the e-news articles is required to generate accurate knowledge from a collection of e-news articles. When the document is not filtered well, filtration becomes complex. The filtration of stop words and suffixes is essential to find the relationship between the keywords. Manual filtration of e-news articles will take much time, and users may miss track of what they are searching for.

Stemming: Stemming reduces the ambiguous association rule generation due to suffixes in the natural text and further reduces the complexity and extra memory consumption while refining the effectiveness and performance of knowledge discovery systems [3]. Many works of literature [3] use a well-known rule-based stemming algorithm; Porter stemmer since it meets both the satisfaction and effectiveness in attaining more relevant information in the text.

Indexing: It is required to index the text to use the unordered words in the text documents [1,2]. Manually or automatically generated indexes using textual data can be considered the foundation for the knowledge discovery process [2]. However, it is not easy to apply manual indexing to large-scale textual data in e-news articles due to its limitations such as heavy time consumption [2]. To address these issues, automated indexing procedures have been examined as a technique that allows association extraction techniques on large-scale textual data [2,3]. Many pieces of research follow the most widely used frequency-based automated indexing weighting scheme which is known as TF-IDF (Term Frequency - Inverse Document Frequency) to calculate and assign higher weights to differentiate terms in a source document [2,3]. Extracting association rules using text documents are based on keyword features in the text, and correlations between those keyword features are based on the weights calculated using the TF-IDF method [1–8].

3. Role of Association Rules in Text Applications

The continued growth in large databases emphasizes the necessity of improving mining performance and precision. Hence, many studies develop novel mining algorithms, theories, and improvements to existing methods. ARM is an exciting DM research domain and raises the interest of many researchers to develop highly efficient algorithms to mine association rules from text documents. ARM concepts span across many domains as per the nature of the problem and demand.

The medical field is a universal domain that requires excessive effort in terms of knowledge management. Most medical applications are daily producing a bulk of medical data in free text formats, which is a very time-consuming task for medical professionals to read and interpret. Hence, applications of TM are widely used in the biomedical field to automatically extract medical findings in these free-text reports based on functional keywords [6]. In [6], the author integrated the Apriori algorithm to find interesting medical data patterns and easily understand the results.

The application of ARM for Knowledge discovery in the newspaper domain is very useful since newspaper readers can easily discover knowledge from a collection of documents without reading all the documents manually [3]. For example, in recent studies [1,2], the authors describe their experimental results of applying Extracting Association Rules from Text (EART) using a set of medical documents related to the epidemic of H5N1 avian influenza virus taken from the MEDLINE bibliographic database. EART uses XML technology with TF-IDF Information Retrieval (IR) scheme for feature selection and the DM techniques for association rules extraction using web documents. As a solution to the time-consuming manual itemsets generation, EART identifies the most essential word features such as the name of the disease, reported outbreak location, current status, and type of the victim, etc. for use in association rules extraction. EART consists of four phases; (1) structural phase (2) indexing phase (3) TM phase and (4) visualization phase.

The main focus of the EART system [1,2] is on the keywords that appear in the extracted association rules and their statistical

distributions without considering the order in which the words occur in a sentence. The ERAT system uses extracted association rules to identify the relations between features in the text document collection. The EART system analyses the keywords in the extracted association rules from two perspectives. (1) The co-occurrence of a keyword in one sentence in the original text document. (2) The existence of a keyword in one sentence without co-occurrence. The EART system outperforms other systems that extract and evaluate the association rules using the Apriori algorithm [1,2]. Experimentation results in [1,2] accepting the challenge of extracting association rules from multidimensional information presented in web news documents.

Another study in [3] discovers association rules and thereby enables the user to generate knowledge from a collection of web news articles related to diseases such as Cholera and Dengue. However, the process of association rule extraction has an expanding growth and a massive number of rules can be extracted from the database [2]. To enhance the efficiency in discovering knowledge from the large number of association rules based on user queries, this study includes an association rule training system that contains pre-defined details about each disease such as impacts of those diseases, reasons for spreading diseases, locations, and victims. This system supports reliable decisions makings due to the knowledge discovered by identifying significant association rules depending on the keywords in the query supplied.

Due to the comprehensive, up-to-date, and valuable information, which contains many broad and collective types of concepts or features, many pieces of literature [1–3] use web news documents as their experimental datasets. Furthermore, web news can more easily be adapted to TM as it does not require the involvement of a subject matter expert to understand and interpret the text features and concepts present in news documents.

In [7], the author deviated from traditional ARM techniques and presented a fuzzy framework for identifying association rules in a text document. This fuzzy framework consists of a fuzzy extended Boolean model, and generated fuzzy association rules are applied to query refinement in Information Retrieval. The generated fuzzy association rules let users query the text document and refine it by showing them a list of candidate terms. Furthermore, this study presented different procedures to apply fuzzy association rules automatically and semi-automatically.

A recent study in [5] accepts the challenge of handling duplications and ambiguities in text data by explaining the relationship between rough set-based decision rules and association rules. The rough association rule-based approach in their study improves ARM effectiveness. The specific information included in rough association rules distinguishes them from standard association rules. The rough association rule comprises a set of terms and their frequency distributions. It is also feasible to dynamically update these rough association rules to increase the effectiveness of the results [5].

The demand for text classification is increasing with the increasing number of online texts. In [9], the author presents a novel fast algorithm for classifying textual data using the concept of ARM. This algorithm is capable of deriving feature sets from pre-classified text documents. Later, the Naïve Bayes classifier applies to the derived features to get the final output of the classification. The algorithm's accuracy in classifying a new document directly depends on the associated word sets generated from pre-classified text documents and performs better with the nonoverlapping text categories.

Detecting and displaying topics from a group of documents is another ARM application. For example, in [4], the author presented a novel approach that uses the visualization of generated association rules to identify the topics from a collection of documents. This approach [4] extends the Apriori algorithm-based approach described in [10]. The experimental results in [4] have shown a reliable match between extracted topics to those present in the data set.

3.1. Role of Apriori Algorithm in Text Applications

The preliminary concept of the Apriori algorithm was the supermarket shopping cart transactions with a set of frequently purchased items [11]. Apriori is one of the well-recognized ARM algorithms that find the frequent items in transactional data sources [1–12]. Association rule mining or discovery using the traditional Apriori algorithm is the most common approach in retrieving hidden rules in the data sources [1–12] but has to pay with immense resources and time.

In [10], the author implements two extending algorithms to discover association rules that are primarily different from the traditional Apriori algorithm. The evaluation results in [10] show that these algorithms perform better for problems with small itemsets than large ones compared to the Apriori algorithm. Furthermore, their study elaborates on another algorithm; Apriori-hybrid [10] by combining the best features in the initially proposed two algorithms to increase the performance in ARM-based applications. Experiments on the Apriori-hybrid algorithm show that it has a linear relationship with the number of items in an association rule and the number of transactions in the database.

Another study in [11] implements Apriori MSG-P, a modified association rule mining technique to demonstrate rare association rules in an operational databank taken from a hospital. During the evaluation process, they examine different characteristics in support values by applying multiple minimum support value approaches for discovering rare itemsets. Experimental results prove that these rare itemsets are more effective than the traditional Apriori algorithms in identifying widespread knowledge and patient behaviors in their databank.

In [2,3], the authors used GARW (Generating Association Rule based on Weighting Scheme) algorithm to overcome the multiple passes over the chosen data, which is another major problem in the Apriori algorithm. The GARW algorithm scans only an XML file containing all the keywords above the given threshold values and their frequencies in each document to generate large frequent keywords. The execution time of the GARW algorithm is less than the Apriori algorithm [1,2].

In recent research, the algorithms for ARM implement numerous optimization techniques to minimize space utilization

and save time in computing frequent itemsets and their support values. The novel scalable algorithm presented in [8] proposes a solution for the existing problems in frequent itemset mining by partitioning the search space. This [8] algorithm discovers closed frequent itemsets which are the lossless and concise presentation of all the frequent itemsets that can be mined from a transactional database. The adaptation of the bitwise vertical representation of the database and the divide-and-conquer approach makes this a fast and memory-efficient scalable algorithm. Moreover, this algorithm addresses the main problems in generating the same closed itemset multiple times using an effective and memory-efficient pruning technique. Mainly, this algorithm does not require keeping the extracted closed patterns in the main memory and enables each visited partition of the search space to be mined independently in any sequence or parallelly.

3.2. Role of FP-Growth Algorithm in Text Applications

Frequent Pattern Growth (FP-Growth) is a frequently used alternative algorithm for generating Frequent Item Sets in a data set. This algorithm enhances the Apriori algorithm that overcomes many of the major problems associated with the Apriori algorithm [13]. Theoretical research in [14] has proven that the FP-growth algorithm resolves two significant complications of the traditional Apriori algorithm. The experimental results in [14] demonstrate that the FP-growth algorithm has higher frequent pattern mining efficiency, less memory usage, and less CPU utilization than the Apriori algorithm. The FP-Growth algorithm accompanies the divide-and-conquer strategy and does not generate candidates because it uses FP-Tree. This dense tree data structure can compress the original transaction database to extract Frequent itemsets [14]. Frequent pattern mining using the FP-Growth algorithm works very well in many broad applications [13-29], including software bug detection, clustering, classification, and recommender systems, and supports selecting the best patterns while reducing time and cost.

The FP-Growth algorithm is mainly used for mining complicated patterns from graph databases [15]. This algorithm is designed for itemset mining, not for graph mining and it does not mine frequent subgraphs well [15]. Hence, finding the frequent subgraphs that support is greater than the given minimum support using a collection of graphs is one of the main problems of frequent subgraph mining [15]. Thus, it is required to make necessary variations in the algorithm so it can be efficiently used for graph mining.

Many of the existing document mining methods for document clustering are based on the frequency of keywords presented in the document [15]. Such methods identify a document as a vector and keywords with their frequency as its elements. However, a document has no way to possess relationships among the keywords. Hence, these methods are not adequate in representing the document's ideas. Clustering documents using the FP-Growth algorithm for association rule mining is a novel approach that overcomes the issues in existing approaches [15]. In [15], the author used the FP-Growth algorithm to mine text documents, find association rules, check the similarity between generated association rules using clustering, and use those results to determine frequent subgraphs. This study modified the FP-Growth algorithm to find the frequent subgraph with clustering affinity propagation in a graph. The ability to evaluate the large-scale graph paths improves the performance of this modified algorithm.

The performance of frequent pattern mining algorithms is affected by many factors, including the databases' characteristics. The FEM (FP-growth & Eclat Mining) algorithm proposed in [26] takes advantage of both FP-tree and TID-list (transaction ID list) data structures to efficiently mine short and long frequent patterns from datasets. The experimental results of the FEM algorithm show a significant improvement in the performance of mining frequent patterns compared to the FP-Growth algorithm. Another experiment presented in [25] concluded that the Dynamic FP-tree construction/reordering algorithm (DynFP-growth algorithm) presented by them behaves better than the FP-growth algorithm.

This algorithm modified the original structure of the FP-Growth algorithm by replacing the single linked list which was used to link the tree nodes to the header with a doubly-linked list and further adding a master table to the same header. Even though the existing database is getting updated, the generated FP-tree in the DynFP-growth algorithm doesn't rebuild the tree. Instead, the algorithm has to be performed considering both new transactions in the database and the initially generated FP tree. Hence, this approach can provide speedy responses to any queries even on the databases that are being repeatedly updated.

The N Painting-Growth and Painting-Growth algorithms developed in [28] are two different improved versions of the FP-Growth algorithm which use two-item permutation sets to extract association rules. The N Painting-Growth algorithm builds two-item permutation sets while the Painting-Growth algorithm builds an association picture based on the two-item permutation sets to find out association sets of all frequent items and then extract all the frequent itemsets according to the association sets. These algorithms reduce the overhead of scanning the database twice in traditional FP-Growth algorithms up to once while increasing the time efficiency. Furthermore, using two-item permutation sets gives several advantages such as lesser memory consumption, running faster, low complexity, and being easy to maintain.

4. Association Rule Mining (ARM)

Association Rule Mining (ARM) is an imperative research area to discover frequent patterns by highlighting correlations between word features in the texts [13–18]. In general, ARM is a process with three major phases; (1) the Text Pre-processing phase, (2) the Association Rule Mining phase, and finally (3) the Visualization phase.

4.1. Text Pre-processing phase

The massive amount of information available in unstructured text format cannot be directly used for further processing. Mining

large document collections to extract useful patterns requires applying essential pre-processing steps to the source documents and storing the information in source documents in a more appropriate data structure for further processing than a plain text file [3]. Text pre-processing steps typically compromise tokenization, filtration of keywords, and stemming and indexing of the keywords [3] to improve the performance in the ARM phase.

4.1.1. Text Transformation

Information on the web presents in different document formats and structures which requires converting them into a machine-processable format that can be easily processed by computers [1,2]. The Extensible Markup Language (XML) is a metalanguage that is used to display documents on the internet based on a user-defined markup language that is amenable to further processing. Supporting techniques in XML help automate the indexing of documents and thus make machines processable [1]. The ERAT system [2] initially saved the Web news pages as text documents and later transformed them into XML format for further processing.

4.1.2. Tokenization

Generally, unstructured data in web pages cause inaccurate results in TM [3]. Thus, it is required to convert unstructured data formats into structured formats by splitting the text in documents into words or terms and selecting only valid tokens by removing inappropriate information such as parentheses, commas, special characters, etc., using tokenization. Tokenization is crucial in removing irrelevant tokens, as it can reduce unnecessary memory and time consumption in the knowledge discovery process. MySQL data structures [3] and matrixes developed using MATLAB software [6] can be used to store tokenization operation results, making it easier to perform further operations such as filtration and stemming, etc.

4.1.3. Filtration

Finding out the relationship between all the keywords present in the text is required for generating accurate knowledge from the collection of text documents. But finding relationships becomes complex if the document is not filtered well. Due to the morphological richness of language, text documents contain many suffixes and stop words [3] which occur frequently likewise "a ", "the", "so" and so on. The careful disregard of stop words without losing the sentence's meaning is required as they do not add much meaning to a sentence and the suffixes may attach to the same word but with different forms. Without filtering such suffixes and stop words in the input documents, finding relationships in text documents and generating accurate knowledge using keywords may become an inefficient and complex process [3]. Unimportant words such as articles, determiners, pronouns, prepositions, common adverbs, non-informative verbs, and conjunctions get discarded. Only the more significant words from the document content are used for further processing [2,3].

Manual filtration will take much time and may miss the track quickly. EART [2] used a list of stop words that are frequently used in English document scripts. Moreover, the EART [2] system replaces parentheses, commas, and special characters with whitespaces.

4.1.4. Stemming

Stemming is reducing the variant forms of the same word to its word stem or root which is known as a lemma. This process happens by removing a word's prefixes and suffixes [2]. Suffixes are single or sets of letters added at the end of a word to generate different forms of the initial word while prefixes are single or sets of letters added at the beginning of a word to generate different forms of the same word. Applying the process of stemming is vital in applications with Natural Language Understanding (NLU), Natural Language Processing (NLP), and Information Retrieval (IR). Applying stemming operation before retrieving information from the documents is much important as it reduces the size of the dataset and the ambiguous association rules generated due to the presence of suffixes and prefixes, and further reduces the complexity and extra memory consumption while refining the effectiveness and performance of knowledge discovery systems [3].

Several types of stemming algorithms attempt to convert a word into its stem or root form. Still, they differ concerning performance and accuracy. Many works of literature [3] use a well-known rule-based stemming algorithm; Porter stemmer since it meets both the satisfaction and effectiveness in attaining more relevant information in the text. One of the common problems in porter's stemmer is that it does not give precise root words at all times. But still, the Porter algorithm was found to be the best algorithm to perform stemming among diverse stemming algorithms such as Paise and Krovtz [3] that are described in the literature [3]. The studies in [1,2] designed a novel stemming dictionary (lexicon) which is suitable for use in the medical domain.

4.1.5. Indexing

Indexing using a weighting scheme comes after the filtration and stems in the text documents. Manually or automatically generated indexes using textual data can be considered the foundation for the knowledge discovery process [2]. However, it is not easy to apply manual indexing to large-scale textual data due to its limitations such as heavy time consumption [2]. Hence, automated indexing procedures have to be considered in applications that use ARM techniques on a large scale [2,3]. In automated indexing techniques, each text document is represented by a set of keywords known as index terms where each index term is a single word or a phrase whose semantics support representing the central theme in the document [2]. Thus, each index term has varying relevance in representing and expressing the document's

theme in a document collection. A numerical weight is assigned to each index to capture its effectiveness.

Definition and Notations in TF-IDF

Many pieces of research [2,3] on extracting association rules follow the most widely used frequency-based weighting scheme; TF-IDF (Term Frequency - Inverse Document Frequency) to identify and filter the terms assigning higher weights to differentiate terms in a source document.

The Term Frequency (TF) is the count that denotes the frequency of a term (keyword) in a collection of documents. TF can be calculated using (1),

$$(tf)_{i,j} = \sum_{i=j=1}^{n} (Nt_i, d_j) \quad (1)$$

Where Nt_i, d_j represents the number of times the term t_j occurs in a document d_j.

Inverse Document Frequency (IDF) is the count that denotes the frequency of source documents that contains the term (keyword) at least once. IDF can be calculated using (2),

$$(idf)_{i,j} = \sum_{i=j=1}^{n} log_2 \left(\frac{|D|}{Nt_j}\right) \quad (2)$$

Where Ntj denotes the document frequency(df) of the term t_j which is the number of documents included in the document collection D in which t_j occurs at least once. The total number of documents in the document collection is denoted using |D|.

TF-IDF value for each term (keyword) can be calculated using (3),

$$w(i,j) = tfidf(d_i, t_j) = Nt_i, d_j * \sum_{i=j=1}^{n} log_2 \left(\frac{|D|}{Nt_j}\right) \quad (3)$$

$$w(i,j) = tfidf(d_i, t_j) = (tf)_{i,j} * log_2 \left(\frac{|D|}{df_j}\right) \quad (4)$$

The simplified version of (3) is in (4) and it calculates the w(i,j) only if $Nt_i, d_j \geq 1$. When the words that do not appear in D, the $Nt_i, d_j = 0$ and thereby w(i,j) = 0.

The value of the Document Frequency (df) scales the d logarithmically. Hence, the formula $log_2 \left(\frac{|D|}{Nt_j}\right)$ gives the full weight for *tfidf* if a term occurs in only one document and if a term occurs in all the documents, then the value of *tfidf* would get zero.

The two important aspects reveal from the TF-IDF weighting scheme are,

1. By considering the *tf* values, the more often occurring terms in a document can be identified as the representative terms of the content of the document.
2. By considering the *idf* values, a term that occurs in more documents can be identified as less discriminative.

Many applications based on Information Retrieval (IR), Text summarization, and association rule extraction from text documents use TF-IDF to weight terms. In [2], the author applies a statistical relevance-scoring function to collecting web news documents to assign scores for each keyword based on maximal TFIDF. Depending on the user's requirement, the top N terms with the highest TF-IDF values can be occupied as the final list of keywords used in the ARM phase [1–3]. The major advantage of automated indexing based on TF-IDF is that it reduces the indexing cost [2].

4.2. Association Rule Mining (ARM) phase

Association Rule Mining (ARM) technique automatically finds information by extracting association rules which contain relationships such as "one implies the other" or "occur together" associated within a list of keywords in a collection of indexed documents. Following the user's requirement, the high-frequency keywords in the indexed documents are selected to generate association rules.

4.2.1. Definition and Notations in ARM

Given a set of keywords A = {w1, w2,..., wn} and a collection of indexed documents D={d1,d2,..., dm}, where each document d_i is a collection of keywords such that $d_i \subseteq A$. Let W_i be a set of keywords. A document d_i is said to contain W_i if and only if $W_i \subseteq d_i$. The general expression of an association rule is an implication of the form $W_i \Rightarrow W_j$ where W_i (antecedent) and W_j (consequent) are sets of keywords in a collection of indexed documents such that $W_i \subset A$, $W_j \subset A$ and $W_i \cap W_j = \varphi$.

The most significant two measures used in the Association Rule evaluation process are support(s) and confidence(c).

If s% of documents in the collection of indexed documents D contain $W_i \cup W_j$, then the Association rule $W_i \Rightarrow W_j$ is said to have support (W_i, W_j) that can be calculated using (5).

$$Support(W_i W_j) = \frac{Support\ count\ of\ W_i W_j}{Total\ number\ of\ documents\ (D)} \quad (5)$$

If C % of indexed documents in the collection of documents D that contain W_i also contain W_j, then the Association rule $W_i \Rightarrow W_j$ holds in D is said to have confidence (W_i, W_j) that can be calculated using (6).

$$Confidence\ (W_i \backslash W_j) = \frac{Support\ (W_i W_j)}{Support\ (W_i)} \quad (6)$$

The general definition of the above two basic measures depends on how a transaction is defined based on the domain. As an example, a transaction can be the records in the dataset or database, each document in the collection of documents [2].

The two threshold values, minimum support, and minimum confidence are the constraints of support and confidence values that the user establishes. An association rule whose support and confidence values are higher than the user-defined thresholds is said to be a strong association rule.

Association rules extraction is a two-step process [1-7]: First, find frequent itemsets, all the itemsets that have support above

minimum support value. Secondly, use the identified frequent keyword sets in step 1 to generate the association rules by discarding those rules below minimum confidence. The later step in the Association Rule generation process is more straightforward than step one, requiring more time and effort [2].

4.2.2. Apriori Algorithm in ARM

Apriori is one of the well-recognized ARM algorithms which supports finding frequent items in transactional data sources [1,2] using two main phases, where the first phase counts item or keyword occurrences to determine the large 1-itemsets and the second phase (k) is again two-fold. In the first fold, the large itemsets L_{k-1} found in the $(k-1)^{th}$ pass are used to generate the candidate itemsets C_k by the Apriori function. The database is scanned and the support of candidates in C_k is counted in the second fold.

The below steps demonstrate the general workflow in the Apriori algorithm associated with the terms in Table I.

Table 1. Notations Used in The Apriori Algorithm

Term	Description
k-keywordsets	A keyword set with k- keywordsets
L_k	Set of large k- keywordsets that satisfy min_{sup} value (in the kth pass)
C_k	Set of candidate k- keywordsets generated from potentially large k-keywordsets
min_{sup}	User-specified minimum support value
min_{conf}	User-specified minimum confidence value

Step 1: Scan the indexed documents or database and find all their accumulated counts. Select the set of frequent 1-itemsets that satisfy the threshold min_{sup} value which is known as L_k.

Step 2: Uses L_k to find the candidate 2- itemsets (C_{k+1}) set. This is a two-step process where step 1 is to use the L_k found in the kth pass to generate the candidate 2-itemsets C_{k+1}. The second step is to scan the indexed documents or database and calculate each candidate's support in C_{k+1}.

Step 3: Calculate the support count for all the itemsets in C_{k+1} and select the frequent 2-itemsets that satisfy the threshold minsup value known as L_{k+1}.

Step 4: Continues steps 2 and 3 until no new frequent (k+1)-itemsets are found.

Step 5: Find all the association rules that satisfy the threshold $min_{conf.}$ for each frequent itemsets.

Apriori follows an iterative level-wise search approach, where k-itemsets are used to explore (k+1)-itemsets [1,2,9].

As described in Table II, the input to the above algorithm is a transaction database where each entry is a transaction of the form <Transaction_id, Set of Items> which is also known as horizontal data format.

Table 2: Horizontal Data Format

Transaction _id	Set of Items
1	a,b,c,d
2	b,c,d
3	A c
4	c,d

The Apriori algorithm will encounter two problems when the frequencies in a set of items have a great variation.

- If the value of min_{sup} is set too high, then the association rules that involve rare items will not be detected. To find out the association rules that involve both frequent and rare items, the value of min_{sup} has to be set very low. Since the frequent itemsets will be associated with each other in all possible combinations, this may cause the generation of a large number of candidate sets [13,14].
- Apriori requires a repeated scan of the database for calculating the support count for candidate sets [13,14].

The vertical format was applied to solve the above issues arising in the DM field by transforming the horizontal data layout described in Table II into the vertical data layout described in Table III by scanning the data set for mining frequent itemsets [13,29,30]. The vertical data format in Table III shows the 1-itemset and it contains entries like <Item, Set of Transactions containing the item>.

Table 3: Vertical Data Format

Item	Transaction Ids	Frequency
a	1,3	2
b	1,2	2
c	1,2,3,4	4
d	1,2,4	3

Every item in Table II is considered as a candidate 1-itemset in the above transaction. This layout employed an intersection between the Transaction_id lists to calculate the support value of the 'candidate's itemset identified in subsequent steps. After counting their support values, all the candidate itemsets which are lesser than the minimum support are discarded.

The vertical layout is a refinement of the traditional Apriori algorithm because it substantially reduces the number of database scans [13,29,30]. This improvement is because the support count (Frequency) is stored in vertical data format just computing the cardinality of each Transaction Id set. Hence, the mining

efficiency is increased and the processing time is decreased compared to the default horizontal layout.

4.2.3. FP-Growth Algorithm in ARM

The FP-Growth is an efficient algorithm that has been proposed as an alternative to the Apriori algorithm for mining frequent itemsets without generating candidates [19–24]. It stores the database transactions in a tree structure which are denoted as FP-tree [19–24] where every item in a transaction has a linked list going through all other transactions that contain the item. The FP-tree structure uses a vertical and horizontal database layout to set aside the transactional database in the main memory. Every node in the FP-tree has a counter value that maintains the number of transactions shared by the node. All occurrences of an item in the FP-tree are linked with the next occurrence of the respective item in the FP-tree using links. Apart from that, a header table contains each item in the FP- tree and its support value and a link to refer to the first occurrence of the item in the FP tree. All the items in the FP tree are stored in the descending order of the support value. Hence, more frequently occurring items are organized closer to the root node of the FP-tree and such items are more likely to be shared than others.

The FP-Growth algorithm consists of 3 main stages:

- Generation of conditional pattern base
- Generation of conditional FP-Tree
- Searching for the frequent itemsets

Algorithm 1 demonstrates the general workflow in the FP-Growth algorithm that incorporates the above 3 main stages.

Algorithm 1: FP-Growth algorithm

```
Input: FP-Tree
Output: The complete set of frequent patterns
Method: FP-Growth (Tree, null)
Procedure: FP-Growth (Tree, α)
if the Tree consists of a single path P then;
    for each combination (denoted β) of the nodes in the path,
    P do
        generate patterns β ∪α with support = min support of
        nodes in β
else
    for each a_i in the Header of Tree do
        generate patterns β = a_i∪α with support = a_i.support
        construct β's conditional pattern base and FP-Tree,
        Treeβ
    if Treeβ ≠ null then
        CALL FP-growth (Treeβ, β)
end
```

The algorithm first compresses the database into a highly condensed FP-tree structure, avoiding costly database scans [23]. Next, it applies the divide-and-conquer approach to decompose the ARM tasks into smaller ones avoiding candidate generation [23, 24].

The FP-GrowthThe FP-Growth algorithm has overcome one major problem of the Apriori algorithm. It has overcome one major problem the Apriori algorithm does not generate a great number of candidate itemsets [13]. Furthermore, this algorithm transformed the entire text database into an FP-tree that holds all the database's information. Hence it requires only two scans of the database and does not scan it again and again as in Apriori [14–19]. The first scan is to find out the frequent words and their support count and the second scan is to sort out each transaction according to their descending support count. Moreover, searching frequent itemsets in FP-Tree does not use generated candidates as it is done on the Apriori algorithm. Hence, the FP-Growth algorithm becomes faster than the traditional Apriori algorithm [13-29].

4.3. Visualization phase

ARM often results in a massive number of association rules which requires the user to go through all the generated rules to discover the interesting rules. Manually analyzing such a massive number of rules is a time-consuming, tough, and tiring activity [20,21]. These issues can be addressed by interactive visualization of the extracted association rules in a textual or graphical format. Association rule visualization has an extensive history of transforming bulky amounts of data into a better accessible method for analysis. Interactive visualization techniques fall under three important categories named scatter plots, matrix visualization, and graph-based visualization [31]. The main elements used in these visualization categories are shown in Figure 1.

As shown in Figure 1, scatter plots utilize different interest measures such as confidence and support on the x and y-axis, and association rules with similar values for such interest measures are placed adjacent to each other. The graph-based visualization demonstrates how association rules are shared by individual items while matrix visualizations demonstrate association rules that have the same antecedent or consequent in the same column or row. The association rules in Figure 1 are shown in blue color.

The literature introduced different visualization techniques such as scatter plots, double-decker plots, mosaics, and parallel coordinates plots to visualize and analyze association rules [21]. Nevertheless, most of such techniques are still falling short with the presence of a large number of association rules [21,22]. However, the most traditional visualization techniques are not much appropriate for visualizing massive sets of association rules since they generally consist of many unique antecedents and consequents [21].

In [21], the author introduced a novel visualization method, a grouped matrix-based representation, which automatically creates nested groups of association rules by applying the k-means clustering algorithm to handle the massive number of generated rules. The hierarchical structure formed by these nested groups of association rules can be interactively explored down to each rule. This Matrix-based visualization [20,21] technique represents the antecedent and consequent of the association rule on the x and y-axes while any selected interest measure is displayed at the intersection of the x and y-axes of a given rule.

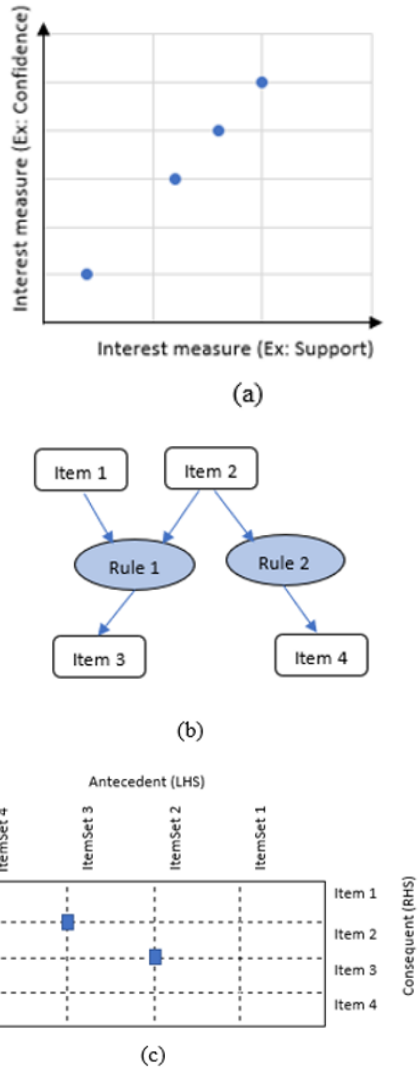


Figure 1. The main elements of association rule visualization using (a) a scatter plot, (b) graph-based visualization, and (c) matrix visualization.

This technique addresses the typical problem that is aroused with huge sets of association rules by grouping antecedents together and letting the end users explore the most interesting groups automatically using different colors and the position of elements in the plot. Grouped matrix-based visualization is easy to understand and unique [21] since most other visualization methods are not able to efficiently deal with huge sets of association rules.

The graph-based techniques [22] visualize the most important extracted association rules using vertices and directed edges where vertices represent each item or itemsets. In contrast, edges represent the relationships in terms of rules. Any selected interest measures are added to the edges of the graph as labels on the edges or as edges with different colors or edges with varying widths of the arrows. These graph-based techniques present a very comprehensible representation of association rules for relatively small sets of association rules selected based on their corresponding confidence score ($min_{conf.}$) value. Graph-based visualization techniques are well suited when the end users are concerned with an aggregated perspective on the most important rules [22]. But this technique offers a clear representation only for relatively small sets of rules, which can be easily chosen based on the selected interest measure.

5. ARM using e-news articles

This section describes the implementation details of the Apriori and FP Growth algorithm for a collection of e-news articles.

5.1. Environment and Dataset

The implementation and testing of Apriori and FP Growth algorithms were done in Python (version 3.7.0) using a workstation with Intel(R) Core (TM) i3-8130U CPU @ 2.20 GHz and 4 GB main memory.

This study evaluated the two algorithms: Apriori and FP Growth using the collection of e-news documents written in the English language released by the Document Understanding Conference (DUC) – 2002 which uses Text REtrieval Conference (TREC) data as the experimental dataset. This dataset compromises 128 topics-related e-news article sets and each article set has between 5 and 15 e-news articles, with an average of 10 articles. The news article must be at least 10 sentences long, but there is no maximum length. The list of topics-related e-news article sets used for the experiment is described in Table IV.

Table 4: The list of topics-related e-news article sets used for the experiment (DUC 2002)

Document set ID	No. of news articles	Total No. of sentences
d113h	5	82
d083a	6	95
d061j	7	124
d063j	8	136
d094c	9	137
d075b	10	183
d070f	11	208
d081a	12	210
d072f	13	229
d069f	14	237

5.2. Methodology

As shown in Figure 2, firstly the basic text pre-processing steps described in section 4.1 including filtration, stemming, and indexing was applied to the collection of e-news articles.

Next, the two algorithms: Apriori and FP Growth were applied to the indexed e-news article collections for generating association rules. We used the TF-IDF numerical statistic to generate index terms from the e-news articles and the count of such generated index terms which are above the threshold value of 0.6 are shown in Table V. All the calculations for identifying index terms were done following (1) to (4) described in the section 4.1.5. Depending on the user's requirement, the top N terms with

the highest TF-IDF values can be occupied as the final list of keywords used in the ARM phase [1–3].

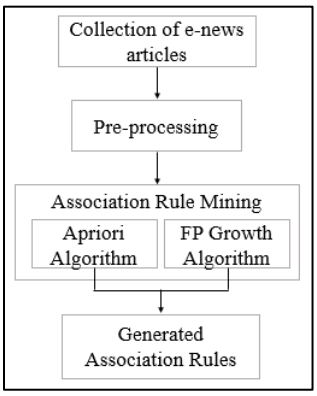


Figure 2. The process of Association Rule Generation

We generated the association rules with different combinations of min_{sup} (values: 60,70,80,90,100) and min_{conf} (values: 80,85,90,95,100) values. All the generated association rules that satisfy each combination of min_{sup} and min_{conf} values were taken for the experiment.

Table 5: The count of index terms (keywords) generated for each topics-related e-news article sets used for the experiment (DUC 2002)

Document set ID	No. Index Terms
d113h	528
d083a	593
d061j	789
d063j	847
d094c	698
d075b	1180
d070f	1271
d081a	1420
d072f	1542
d069f	1782

Table 6: Association Rules generated for the document set ID: d113h using the Apriori algorithm

min_{sup} \ min_{conf}	80	85	90	95	100
60	2909	2221	1912	1459	1127
70	2793	2442	1789	1009	808
80	2543	1992	1405	828	576
90	2046	1551	1079	518	251
100	1506	1119	731	385	73

Our experiment generated the association rules for document sets in Table V using different value combinations of min_{sup} and min_{conf}. Support and Confidence value for each generated association rule were calculated using (5) and (6) in section 4.2.1. All the association rules that satisfy each combination of min_{sup} and min_{conf} values were taken for evaluation.

Table 7: Association Rules generated for the document set ID: d113h using the FP Growth algorithm

min_{sup} \ min_{conf}	80	85	90	95	100
60	2329	2110	1794	1391	978
70	2172	1825	1490	1121	873
80	1856	1513	1201	705	584
90	1621	1297	903	561	240
100	1349	1067	871	378	81

The total number of Association Rules generated using the document set ID: d113h which satisfy the specified min_{sup} and min_{conf} values are listed in Table VI and Table VII.

6. Results & Discussion

The major concern of the experiment described in this study includes a comparison of the execution time of both algorithms for each document set listed in Table V to identify the most effective algorithm for e-news articles.

We monitored the complete execution time taken by each algorithm for generating the set of association rules with different combinations of min_{sup} and min_{conf} values. The execution time taken by each algorithm for generating association rules for the document set ID: d113h with 100% support and 100% confidence is listed in Table VIII.

Table 8: Execution time of Apriori and FP-Growth algorithms for the document set ID: d113h

	Apriori Algorithm		FP Growth Algorithm	
Document set ID	Execution time	No. of rules	Execution time	No. of rules
d113h	40 minutes	73	3 minutes	81
d083a	50 minutes	70	4 minutes	76
d061j	65 minutes	81	5 minutes	80
d063j	70 minutes	85	6 minutes	91
d094c	63 minutes	72	5 minutes	68
d075b	75 minutes	88	6 minutes	76
d070f	87 minutes	89	7 minutes	93
d081a	116 minutes	107	9 minutes	98
d072f	125 minutes	114	11 minutes	91
d069f	146 minutes	123	13 minutes	126

Our implementations of the Apriori and FP Growth algorithms consider each sentence in the e-news article as one database transaction. Hence, each document collection generates a large number of TF-IDF weighted index terms (Table V) or keywords to be considered in association rule generation.

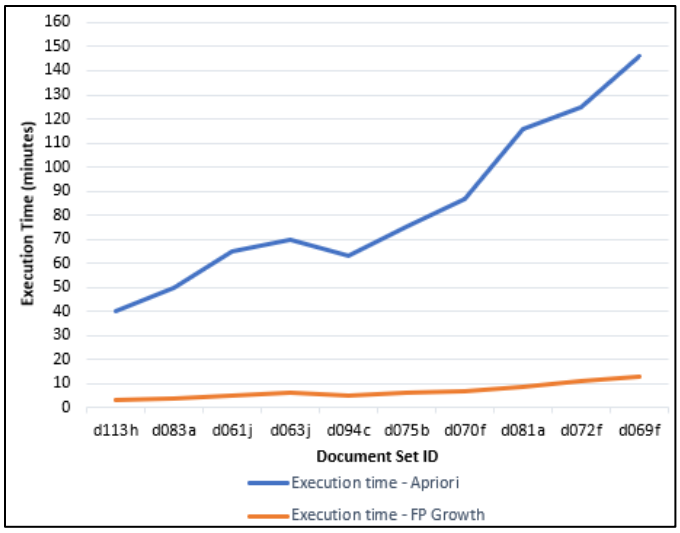


Figure 3: The comparison of algorithm execution time for generating association rules with 100% support and 100% confidence

When the number of generated association rules getting increased, the execution time of both algorithms also getting increased. The details presented in Table VIII and the comparison of execution time in Figure 3 represent that out of these two algorithms, the FP Growth algorithm always performs much better than the Apriori algorithm disregarding the number of generated association rules and several index terms. Mainly this happens due to the major weaknesses in the Apriori algorithm such as several transactional database scans, large memory utilization, and higher execution time.

6.1. The complications in Apriori Algorithm in ARM

Following the experimental results, we evident the below compilations in the Apriori algorithm for association rule generation using textual data.

The Apriori algorithm requires a repeated scan of the database for calculating the support values for each generated candidate set [1-14] which is one of the major problems in the Apriori algorithm. The literature describes the GARW algorithm [1-3] as a solution to this issue since it 'doesn't scan the original documents multiple times as in the Apriori algorithm. Instead, it scans the generated XML file, which includes only the keywords identified during the large frequent keyword sets generation in each document It also includes only the keywords identified during the large frequent keyword sets generation in each document and satisfies the user-defined threshold values. Therefore, the execution time of the GARW algorithm is less than the Apriori algorithm [1,2].

Another major bottleneck in the Apriori algorithm is candidate set generation by handling extremely large candidate sets. Moreover, frequent pattern matching using such massive candidates by searching through each transaction in the database is very expensive and time-consuming. When the frequencies in a set of items have a significant variation and if the value of min_{sup} is set too high, then the association rules that involve infrequent items will not be detected. To identify the association rules that involve both frequent and infrequent items, the value of the min_{sup} has to be set very low. Since each frequent item is getting associated with other frequent items making all possible combinations, this may cause the generation of a massive number of candidate sets [13,14].

As a solution to this issue, the study in [8] presented a fast and memory-efficient scalable algorithm that uses the divide-and-conquer approach. This algorithm uses the bitwise vertical representation of the original database and formalizes the problems that arise in closed itemsets mining by partitioning the search space. This algorithm presents a condensed but lossless representation of all the extracted frequent itemsets in the transactional database and then identifies all the possible closed frequent Itemsets for discovering closed frequent itemsets. Furthermore, this algorithm uses several optimization techniques to reduce space and time utilization in generating itemset closures and their support values. But, one major drawback encountered in this algorithm is the generation of the same closed itemset multiple times [8]. Apart from this algorithm, in [8] the author proposes an effective and efficient pruning technique that does not require storing the set of extracted closed patterns in the main memory. Moreover, this technique allows each visited partition of the search space to be mined parallelly or separately despite the order.

The pruning technique used in the Apriori algorithm does not assure an effective pruning of candidates due to the rapid growth in the frequent itemsets. In this case, the proposed algorithm in [8] ensures that frequent closed itemsets extract all frequent closed itemsets in the transactional database without any duplications while saving memory space and reducing the execution time in itemset closures generation [8].

Association rule generation using the Apriori algorithm considers all the keywords without considering the importance of each keyword. Owing to this, a large collection of association rules will be generated, becoming a tedious and time-consuming task. More importantly, the knowledge discovery using these association rules will not be so worthwhile as it considers all the keywords in the transactional database, including lower-weight keywords [3]. Applying the Apriori algorithm to the vertical data format described in Table III instead of the horizontal data format addresses most of the limitations discussed above [2,3].

6.2. The complications in FP-Growth Algorithm in ARM

Following the experimental results, we evident the below compilations in the FP Growth algorithm for association rule generation using textual data.

The major complication in the FP-Growth algorithm is the cost and time consumption in constructing the FP-Tree. The FP-tree construction in this algorithm has a high requirement for memory space. Hence, the algorithm may not fit in the shared memory when the transactional database gets larger. Apart from that, the observations explained in [25] state that the resulting FP-tree cannot be considered a unique representation of the same logical database that is used in FP-tree generation. But once the construction of the FP-tree is over, it can be easily used to read off the itemsets [19].

The other complication is that the support value can only be calculated after the entire FP-Tree is constructed. Furthermore, when the threshold value of support is getting higher, extensive time is wasted since the pruning can be applied only to one item at once [19,24]. Also, the FP-growth algorithm requires at most two complete database scans [13-29], and scanning the database twice reduces the efficiency of the FP-Growth algorithm.

Even though the Painting-Growth and N Painting-Growth algorithms [28] only scan the database once and are easy to implement, there are still several complications in them. The Painting-Growth algorithm builds the association picture, leading to a large memory overhead. The N Painting-Growth algorithm performs well with macroscale datasets and the implementation of the algorithm is more complex than the Painting-Growth algorithm. Third, these two algorithms repeatedly scan the generated frequent item association sets and extract multi-item frequent sets. This results in reducing the time efficiency of these algorithms [28].

The exponential growth in association rule extraction produces a large number of rules. Association rule extraction approaches are also broadly expanding, and novel algorithms such as Apriori-hybrid are increasing their performance in many real-world applications [6]. Generally, the existing data mining techniques discover innumerable patterns or knowledge from a training data set. Furthermore, noise in the discovered patterns makes the effective knowledge discovery process a big challenge. The concept of closed patterns [8] deals with the noise up to a significant level, but the discovered knowledge still retained many meaningless patterns. Apart from that, the generation and identification of only the interesting patterns based on pre-defined constraints improve the quality of the ARM process [5].

7. Future Directions

Knowledge discovery is an increasing field of research. Still, there is a tremendous need to develop TM approaches using AMR which can guide the user on knowledge discovery in text documents such as web articles identifying important information, and how to generate knowledge. Even though knowledge refinement has been carried out in the literature, the refinements in the ARM are still comparatively less [27].

7.1. Refinements to the Apriori Algorithm

Support and confidence are the two factors considered by the Apriori algorithm to measure the statistical strength of the generated association rules. However, these factors are not useful in deciding the generated rules' convenience or detecting conflicts between the generated rules [27]. Hence, a refinement of the existing factors by introducing a novel factor that considers the convenience of generated rules and identification of conflicts between rules is important for obtaining consistent and interesting patterns.

A few of the major future directions in the Apriori algorithm are,

- Reducing the multiple passes over the chosen data to speed up the algorithm execution.
- Identifying only the most important keywords for generating association rules to reduce generating a large collection of association rules which will not be so meaningful.
- Generating efficient candidate sets by handling exceedingly large candidate sets and matching frequent patterns with a large number of candidates.

Furthermore, the current pruning technique used in the Apriori algorithm does not assure an effective pruning of candidate sets due to the rapid growth in the frequent itemsets.

7.2. Refinements to the FP-Growth Algorithm

The most needed refinement in the FP-Growth algorithm is reducing the cost of development and time consumption in constructing the FP-Tree. Another primary direction in the FP-Growth algorithm is to enhance the algorithm to fit in the shared memory when the transactional database gets larger.

N Painting-Growth and Painting-Growth algorithms developed in [28] provide a direct link for the subsequent ARM research. Enhancing the performance of the Painting-Growth and N Painting-Growth algorithms [28] while increasing the efficiency in infrequent item removal and multi-item frequent sets mining will be some of the leading future works.

8. Conclusion

Effective information extraction using text documents is a massive challenge due to the duplications and ambiguities of data values in the text. The works of literature related to Knowledge Discovery from text documents using Association Rule Extraction have become a more frequent and imperative research approach to finding out the most significant information, patterns, and features in the text documents. This survey provides a comprehensive review of Association Rule extraction using text documents covering a comprehensive review of multiple aspects.

Furthermore, the experimental results discussed in this study answer the question 'Which ARM algorithm: Apriori or FP Growth has fast execution in association rule generation using e-news articles?'. Out of the Apriori and FP Growth algorithms, the FP Growth algorithm always performs much better than the Apriori algorithm disregarding the number of textual contents included in E-news articles. Mainly this happens due to the major weaknesses in the Apriori algorithm such as several transactional database scans, large memory utilization, and higher execution time.

The main outcomes of the study we have conducted are,

- A thorough review of extracting knowledge in text sources using Association Rules.
- A review of the most popular application domains of ARM along with its usage and applications.
- A review of more frequently used association rule extraction algorithms, mainly the Apriori and FP-Growth algorithms in ARM, their usage, recent breakthroughs, advantages, and major bottlenecks in TM.
- Detailed information on the ARM process including a comprehensive review of the Text Pre-processing phase, Association Rule Mining phase, and Visualization phase.
- Detailed information on the perspective directions to facilitate future research.

The major complications in the Apriori algorithm were addressed by introducing a novel compact data structure, FPtree. The FP-growth algorithm that uses the FP tree allows efficient discovery of frequent itemsets and is faster than other ARM algorithms. Even though the value of the support factor influenced the performance of the Apriori algorithm, the performance of the FP-growth algorithm is not influenced by the value of the support factor. Thus, the algorithms with candidate set generation behave well with small databases with the size of a maximum number of transactions of around 50,000 and a support factor of at least 30% [25].

Later, FP-tree-based pattern recognition algorithms such as FEM [26], and DynFP-growth algorithm [25] were developed based on the completeness and compactness of the FP-growth structure. DynFP-growth [25] and FP-growth algorithms that don't require candidate generation perform much better. The experimental results of the FEM [26] and FP-growth [25] algorithms have shown a significant improvement in the performance of mining frequent patterns compared to the FP-Growth algorithm. N Painting-Growth algorithm and Painting-Growth algorithm presented in [28] reduce the overhead of scanning a database multiple times. The use of two-item permutation sets in these algorithms gives several advantages: running faster, occupying less memory space, having low complexity, and being easy to maintain. Based on the survey presented in our paper, it's evident that pattern-based growth algorithms like FP-Growth and their refinements based on the FP-tree are more efficient than algorithms based on candidate generation such as Apriori.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

I am grateful to the University of Colombo School of Computing (UCSC), Sri Lanka for all the facilities provided for this research work.

References

[1] H. Mahgoub, "Mining association rules from unstructured documents'', International Journal of Applied Mathematics and Computer Sciences, **1** (4), 201-206, 2008, doi:10.5281/zenodo.1330457.

[2] H. Mahgoub, D. Rösner, N. Ismail, F. Turkey, "A text mining technique using association rules extraction", International Journal of Computer, Electrical, Automation, Control, and Information Engineering, **2**(6), 2044-2051, 2008.

[3] M. Kulkarni, S. Kulkarni, "Knowledge discovery in text mining using association rule extraction", International Journal of Computer Applications, **143(**12), 30–35, 2016, doi: 10.5120/ijca2016910144.

[4] A. A. Lopes, R. Pinho, F. V. Paulovich, R. Minghim, "Visual text mining using association rules", Comput. Graph., **31(**3), 316–326, 2007, doi: 10.1016/j.cag.2007.01.023.

[5] Y. Li, N. Zhong, "Rough association rule mining in text documents for acquiring web user information needs", in 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06), Hong Kong, China, 226–232, 2006, doi: 10.1109/WI.2006.151.

[6] J. Manimaran, T. Velmurugan, "A survey of association rule mining in text applications", in 2013 IEEE International Conference on Computational Intelligence and Computing Research, Enathi, Tamilnadu, India, 1–5, 2013, doi: 10.1109/ICCIC.2013.6724258.

[7] M. Delgado, M. J. MartÍn-Bautista, D. Sánchez, J. M. Serrano, M. A. Vila, "Association rule extraction for text mining", In: Carbonell, J.G., Siekmann, J., Andreasen, T., Christiansen, H., Motro, A., Legind Larsen, H. (eds) Flexible Query Answering Systems - FQAS 2002, Lecture Notes in Computer Science(), Springer, Berlin, Heidelberg, **2522**, 154-162, 2002, doi: https://doi.org/10.1007/3-540-36109-X_12.

[8] C. Lucchese, S. Orlando, R. Perego, "Fast and memory-efficient mining of frequent closed itemsets", IEEE Trans. Knowl. Data Eng., **18**(1), 21–36, 2006, doi: 10.1109/TKDE.2006.10.

[9] C. M. Rahman, F. A. Sohel, P. Naushad, S. M. Kamruzzaman, "Text classification using the concept of association rule of data mining", in 2010 Proceedings of the International Conference on Information Technology, Kathmandu, Nepal, 234-241, 2010, doi: https://doi.org/10.48550/arXiv.1009.4582.

[10] R. Agrawal, R. Srikant, "Fast algorithms for mining association rules" , in Proceedings of the 20th International Conference of very Large Data Bases, VLDB, Santigo, Chile, 487-499, 1994.

[11] T. Ouypornkochagorn, "Mining rare association rules on Banpheo Hospital (Public Organization) via Apriori MSG-P algorithm", ECTI Transactions on computer and information Technology, **6**(2), 156–165, 1970, doi: 10.37936/ecti-cit.201262.54337.

[12] V. Chandola, V. Kumar, "Summarization - compressing data into an informative representation", in 2005 5th IEEE International Conference on Data Mining (ICDM'05), 1-8, 2005, doi: 10.1109/ICDM.2005.137.

[13] Ali Ikhwan, Milfa Yetri, Yohanni Syahra, Jufri Halim, Andysah, Putera Utama Siahaan, Solly Aryza, Yasmin Mohd Yacob, "A novelty of data mining for promoting education based on FP-Growth algorithm", in 2018 International Journal of Civil Engineering and Technology (IJCIET), **9**(7), 1660-1669, 2018.

[14] Wei Zhang, Hongzhi Liao, Na Zhao, "Research on the FP Growth algorithm about association rule mining", in 2008 International Seminar on Business and Information Management, Wuhan, 315–318, 2008, doi: 10.1109/ISBIM.2008.177.

[15] R. K. Soni, N. Gupta, A. Sinhal, "An FP-Growth approach to mining

association rules", International Journal of Computure Science and Mobile Computation, **2**(2), 1-5, 2013.

[16] Satya Wacana Christian University, I. S. Rajagukguk, S. Y. J. Prasetyo, I. Sembiring, "The analysis and implementation of algorithm of Frequent Pattern – Growth to support the promotion strategy in Victory University Sorong", International Journal of Comput. Sci. Eng., **4**(10), 24–31, 2017, doi: 10.14445/23488387/IJCSE-V4I10P106.

[17] J. Panjwani, "Application of FP Tree Growth algorithm in text mining", M.S. thesis, Department of Computer Science and Engineering, Jadavpur University, Kolkata, India, 2010.

[18] K. Gadia, K. Bhowmick, "Parallel text mining in multicore systems using FP-tree algorithm", in 2015 International Conference on Advanced Computing Technologies and Applications (ICACTA), Procedia Computer Science , **45**, 111–117, 2015, doi: 10.1016/j.procs.2015.03.100.

[19] A. Kaur1, G. Jagdev, "Analyzing working of FP-Growth algorithm for frequent pattern mining", International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), **4**(4), 22-30, 2017, doi: http://dx.doi.org/10.20431/2349-4859.0404003.

[20] Zulham Zulham, Ibnu Rusydi, Ananda H Elyas, "Pattern analysis of health equipment procurement system using the FP-Growth algorithmin", in 2nd International Conference on Inovations in Social Sciences Education and Engineering (ICoISSEE), **1**(1), 2021.

[21] M. Hasler, R. Karpienko, "Visualizing association rules in hierarchical groups", Journal of Business Economics**, 87**(3), 317–335, 2017, doi: 10.1007/s11573-016-0822-8.

[22] M. Hasler, S. Chelluboina, "Visualizing association rules in hierarchical groups", in 42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms (Interface 2011), Intelligent Data Analysis group, Southern Methodist University, 2011.

[23] O. Netzer, R. Feldman, J. Goldenberg, M. Fresko, "Mine your own business: market-structure surveillance through text mining", Marketing Science, **31**(3), 521–543, 2012, doi: 10.1287/mksc.1120.0713.

[24] J. Arora, "An efficient ARM technique for information retrieval in data mining", International Journal of Engineering Research & Technology (IJERT), **2**(10), 1079-1084, 2013, doi: 10.17577/IJERTV2IS100229.

[25] A. Bala, M. Z. Shuaibu, "Performance analysis of Apriori and FP-Growth algorithms (Association Rule Mining)", International Journal of Computer Technology & Applications, **7**(2), 279-293, 2016.

[26] C. Gy, "A comparative study of association rules mining algorithms", in 2004 1st Romanian- Hungarian Joint Symposium on Applied Computational Intelligence (SACI), 1-10, 2004, doi:10.13140/2.1.1450.3365.

[27] G. Alaghband, L. Vu, "A fast algorithm combining FP-tree and TID-list for frequent pattern mining", in 2011 Proceedings of Information and Knowledge Engineering, 472-477, 2011.

[28] M. N. M. García, S. Segrera, V. F. López, "Association rules: problems, solutions, and new applications", in 2005 Proceedings of the III National Workshop on Data Mining and Learning, 317-323, 2005.

[29] Y. Zeng, S. Yin, J. Liu, M. Zhang, "Research of improved FP-Growth algorithm in association rules mining", Scientific Programming, **2015**, 1–6, 2015, doi: 10.1155/2015/910281.

[30] I. Aqra, N. Abdul Ghani, C. Maple, J. Machado, N. Sohrabi Safa, "Incremental algorithm for association rule mining under dynamic threshold", Applied Sciences, **9**(24), 5398, 2019, doi: 10.3390/app9245398.

[31] M. Hasler, "arulesViz: interactive visualization of association rules with R", The R Journal, **9**(2), 163-175, 2017, doi: 10.32614/RJ-2017-047.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 195-203 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Estimating Subjective Appetite based on Cerebral Blood Flow

Lai Kecheng*, He Qikun, Hu Ning, Fujinami Tsutomu

Creative Society Design Research Area, Japan Advanced Institute of Science and Technology, Nomi City, Japan

ARTICLE INFO

Article history:
Received: 15 September, 2022
Accepted: 21 November, 2022
Online: 20 December, 2022

Keywords:
Visual stimuli
Subjective evaluation
Biological response
Cerebral blood flow
fNIRS

ABSTRACT

This study aims to develop and validate a biological food preference task that simultaneously evaluates biological responses to visual stimuli of various food states and subjective evaluations of foods and to examine how these biological responses are related to food preference behavior. We recruited seventeen healthy male and female subjects to observe changes in cerebral blood flow related to salivation and the prefrontal cortex region while performing a food preference task related to visual stimuli of various food states. We also examined the relationship between these changes and the subjects' subjective evaluations. The results showed that subjective evaluations of the various states of visual stimuli differed from subjective evaluations of the different food states. Furthermore, comparing the hemodynamic response function of cerebral blood flow to each visual stimulus, we observed a trend of activation of brain activity in the prefrontal and parotid regions during task execution. In addition, correlations were calculated between the subject's subjective evaluation and cerebral blood flow in the prefrontal and parotid regions and between cerebral blood flow in the prefrontal and parotid regions, and significant differences were observed. Our findings demonstrate the potential of combining the evaluation of food in different states with cerebral blood flow indices in biological responses to visual cues of food.

1. Introduction

This paper is an extension of a work originally presented in ICMHI 2021 [1], which used visual stimuli of foods to examine the relationship between salivary responses and appetite. In this study, we developed a food preference task using visual stimuli of various food states rather than visual stimuli of food. We also measured the biological responses with cerebral blood flow in both parotid and prefrontal regions and examined its relationship to the subjects' evaluation of appetite.

The visual appearance of foods may influence the choice of foods. The association between nutrient and energy content of food, as well as pleasant or unpleasant sensations toward food, are paired with different sensory features. Among the various sensory cues, the visual stimulation of food is an important regulator of appetite. When subjects were presented with visual stimuli, a significant difference was found in blood flow in the parotid region between the low-rating and the high-rating groups concerning subjects' appetite [1]. People can predict its characteristics through photos and videos of food, and this action also influences people's eating behavior. The chemical senses, sight, smell, and taste, play a key role in the sensory effects on appetite, food choice, and intake [2].

Saliva is recognized to have a variety of functions, and it affects oral defense as well as oral and systemic health. Saliva secretion depends on a complex set of factors, including food-related cues, health status, gender, etc.[3,4]. Salivary glands are under the dual control of sympathetic and parasympathetic nerves, and different glands have different salivary characteristics. Sympathetic nervous system activity causes the parotid gland to secrete saliva rich in serum and α-amylase while the parasympathetic nervous system activity causes the submandibular and sublingual glands to produce viscoelastic, mucin-rich saliva [5–7]. Each component of saliva is regulated to perform a particular function. Sensory food cues, such as appearance, fragrance, and taste, may induce a rapid secretion of saliva in the oral cavity, which is called the cephalic phase salivary response [8–11]. Cephalic phase salivary response includes physiological responses to food-related cues such as the thought, fragrance, appearance, and taste of food [12]. Many studies have shown that sensory exposure to various foods increases saliva production.

*Corresponding Author: Lai Kecheng, laikecheng199222@gmail.com

https://dx.doi.org/10.25046/aj070621

Food choice and intake processes consist of various cognitive, sensory, and metabolic components, such as sensory pleasure, metabolic hunger, and knowledge about food[13]. In addition, these non-homeostatic aspects are an important area of research, as cognitive, sensory, and emotional processes often take precedence over food choice decisions in the current food-abundant obesogenic environment [14]. Traditional measures of appetite often use subjective assessment methods that rely on self-report, such as questionnaires or interviews. These subjective assessment methods require cognitive information processing and reasoning by the subject, influenced by factors such as social desirability [13]. However, implicit motivation is difficult to measure, as humans make various daily food decisions. In addition, people's choices and actions regarding food and food cues have characteristics such as motivational processes that we are unaware of, cannot articulate, or do not want to do [13,15]. In recent years, technological advances in biometric systems that measure psychophysiological parameters have made it possible to examine subjective rates of food and the implicit processes involved in food intake [16]. Biometrics, a non-invasive behavioral and physiological measure that may reflect motivation and emotional responses to food, has features that can identify individual characteristics based on biological and physiological properties and is commonly used in food science and consumer science [17].

However, it is still unclear why people respond to foods this way and what psychophysiological processes are possible before and during behavioral responses to food cues. Furthermore, as far as previous studies are concerned, since the number of studies on biometrics is limited, the results are inconsistent, which makes it difficult to compare study results. On the other hand, we believe that biometric measures need to be combined with measures of appetite, as it is necessary to investigate individual motivations and responses behind food choices and intake to promote healthier eating habits.

Therefore, this study aims to develop and test a biological food preference task that simultaneously rates biological responses to visual stimuli of various food states, and subjective evaluations of food. The study then examines how these biological responses are related to food preference behavior. Specifically, differences in biological responses (cerebral blood flow in the prefrontal cortex, salivation) to visual food stimuli under different states of food are rated in a biometric food preference task. Specifically, a database and validation task will be developed using visual stimuli of different states of food. Next, the differences between biological responses to visual stimuli (cerebral blood flow in the prefrontal cortex and parotid region) and the evaluation of foods in response to visual stimuli of foods will be evaluated in a biometric food preference task. Finally, we examine the relationship between biological responses and food preferences.

2. Methodology

2.1. Participants

Seventeen healthy students (5 females, 12 males, aged 25-30) were recruited at Japan Advanced Institute of Science and Technology. Subjects were randomly selected from graduate students who agreed to take part in the study. Before the experiment, all subjects were given a screening session to determine eligibility. A questionnaire survey was conducted on the subjects' age, gender, and medication status. As a result, all subjects had no evidence of systemic or oral diseases such as periodontal disease, and all were nonsmokers. In addition, participants were selected to be subject if their body mass index (BMI) is within the range of 18.5-25kg/m2 because being overweight or obese tends to impact the secretion of saliva as they secrete more saliva due to food cues[18]. Subjects were excluded if they had dental pathologies or issues chewing and swallowing due to the muscle movement, which could possibly affect their saliva secretion[19,20]. The subjects who volunteered were told they would be watching some photos and videos about ice cream and completing some rating tasks for the experiment. The experimental investigation period was from March 15, 2022 to March 23, 2022. The experiment was conducted on those who had consented to the study after the purpose of the study was explained. On the day of the experiment, participants were instructed not to eat for one hour before the experiment. Drinking water was allowed, but drinks containing sugar were not. Each participant was paid 3,000 yen as an honorarium for his participation in the experiment.

2.2. Data Collection

2.2.1. Experimental Device

In this experiment, two types of experimental devices were used. One was the HOT-2000 device, which can measure cerebral blood flow in the prefrontal cortex, associated with cognitive function, and the other was the WOT-S20 device, which can measure cerebral blood flow in the parotid gland, associated with saliva secretion.

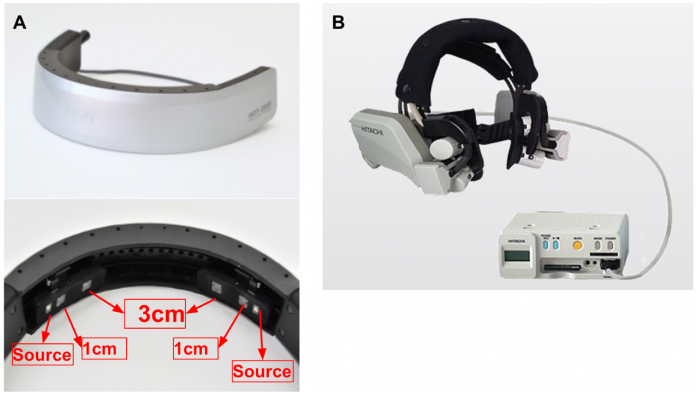


Figure 1: Experimental Devices. A: HOT-2000 device; B: WOT-S20 device.

We measured cerebral blood flow using the HOT-2000. It is a portable brain activity measurement device that can be used in everyday scenarios to simultaneously measure cerebral blood flow change, pulse and head acceleration in real-time [21]. The device used in this study has two Source-Detector (SD) optode pairs. Each optode pair has two optical detectors located at a distance of 1 cm and 3 cm from the light source. The device emits a wavelength of infrared light (about 800 nm) that is easily absorbed by hemoglobin in the blood and the detected light was sampled at a frequency of 10 Hz. The measurement principle is that the infrared light irradiated from the scalp diffuses, and if the area of the brain that is in the path of the light is activated, the amount of light that returns to the detector decreases due to increased blood flow and increased light absorption. The intensity change of the detected light was converted to total-hemoglobin (HbT) on the optical path

of concentration using the modified Beer-Lambert law[22]. The 1cm-SD optode pair is the short separation channel, which provides auxiliary signals from shallow tissues such as the scalp and skull, and we use it to measure what is happening in the scalp and separate brain signals from non-neuronal physiological signals. In contrast, the 3cm-SD optode pair is the long separation channel that reaches the cerebral cortex and reflects blood flow changes associated with neural activity, usually measuring the brain.

We used WOT-S20, a near-infrared optical measuring device for the lateral face (developed by NeU Corporation), to measure cerebral blood flow changes in response to saliva secretion. The WOT-S20 device consists of three components: a headset equipped with a near-infrared light receiving/emitting sensor (one channel each for left and right), a portable control box for wireless data transmission and backup functions, and a measurement controller for real-time display and storage of measurement settings and results [23]. Among them, the source emits infrared light at wavelengths of 705[nm] and 830[nm] to detect hemoglobin changes (Oxy-hemoglobin (HbO), Deoxy-hemoglobin (HbR), Total-hemoglobin (HbT)) in the blood and to capture data at a frequency of 5 Hz. The basic principle of the measurement is that an increase in saliva secretion in the parotid region promotes an increase in blood volume, which decreases the amount of light transmission.

2.2.2. Experimental Materials

For the experimental material, we chose ice cream, a food commonly found in Japanese culture. Four of the most popular ice cream flavors were selected from various commercially available ice cream products and used as experimental materials. The four ice cream flavors selected were matcha, chocolate, strawberry, and vanilla. The ice cream melting process was documented on video by placing each ice cream on a wooden plate and filming the process from hard to wholly melted. Then, we used programming to calculate the area of each state of ice cream during the process from hard to melt. Finally, the ice cream was divided into four states based on the calculated area. In the first stage, only the middle of the ice cream has the characteristic of beginning to melt. In the second stage, the ice cream melts rapidly. In the third stage, the ice cream is still melting, but the melting rate of the ice cream is easing, and finally, in the fourth stage, the general area has not changed much. The fourth stage was a state in which only the middle of the ice cream was melting, although the general area did not change much.

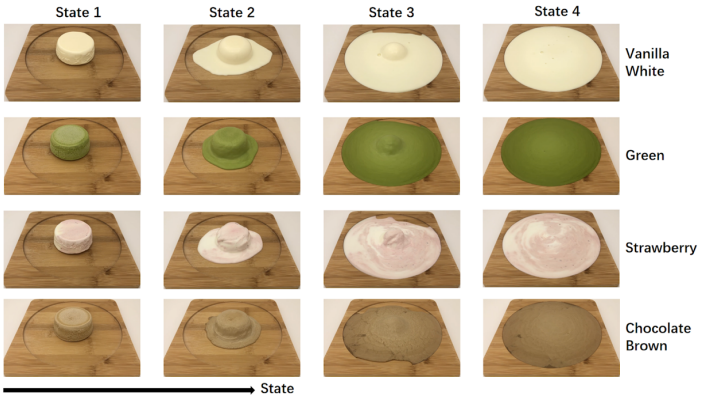


Figure 2: Experimental Material

Then, one representative photo was extracted from each video, and the 16 photos and 16 videos constituted the experimental material database for this experiment.

2.2.3. Experimental Environment

The experimental environment is shown in Figure 3. Figure C on the left shows the subject using the HOT-2000 device and doing the experimental task. Figure D on the right shows the subject wearing the WOT-S20 device and participating in the experiment.

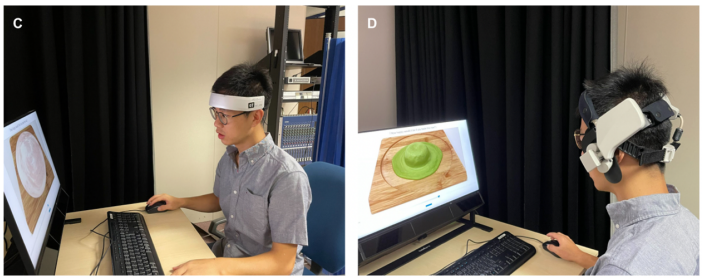


Figure 3: Experimental Environment. C: Experimenting with the HOT-2000 device, D: Experimenting with the WOT-S20 device.

2.2.4. The Flow of the Experiment

According to previous research, the Leeds Food Preference Questionnaire [24]was developed to examine the "liking" and "wanting" components of subjective evaluations of food, which has been reported as a valid method for measuring explicit and implicit aspects of appetite. Moreover, the Leeds Food Preference Questionnaire is a computer task with features that collect ratings, choices, and reaction times to visual food stimuli from different food categories [25].

The task of this study was designed to collect biological responses, including cerebral blood flow in the prefrontal and parotid regions, to food images and video stimuli based on the Leeds Food Preference Questionnaire procedure. A web application was built using JavaScript and HTML for stimulus presentation, and image and video stimuli were presented on the browser. As shown in Figure 2, the visual stimuli of ice cream displayed in the images and video vary in four different states.

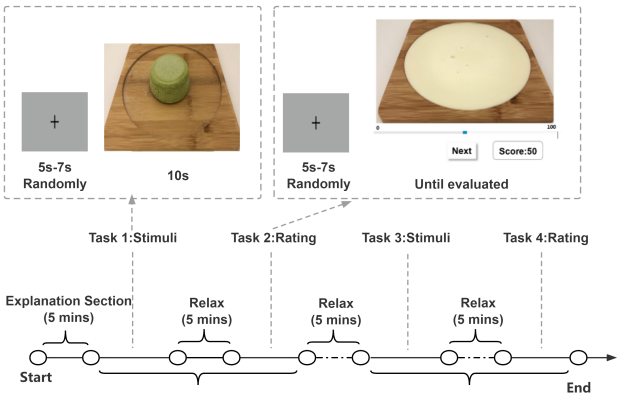


Figure 4: Experimental Procedure

The experimental procedure is illustrated in Figure 4. Participants were instructed to eat one hour before the experiment, and upon arrival at the laboratory, only water was allowed while

the experiment was in progress. After explaining the criteria for participation in the experiment, the purpose, and content of the experiment, and confirming the subjects' eligibility, a preliminary questionnaire was conducted regarding the subjects' age, BMI, frequency of ice cream consumption, and preferences.

As shown in Figure 4, the task is consisted of four parts (Task 1 to Task 4). Task 1 involves passively viewing 16 ice cream pictures on two separate occasions, and Task 2 requires the participants to rate the stimuli presented in Task 1. The specific procedure was that in Task 1, pictures of ice cream were displayed. Then, subjects were allowed to look at them for ten seconds, followed by Task 2, in which subjects were asked to rate them on a scale from 0 to 100 (0 means the lowest degree and 100 means the highest degree of liking and wanting). Then, a new image was presented in Task 1 for ten seconds, and the subject was asked to rate the image presented in Task 1 in Task 2. Thus, sixteen ice cream images were displayed randomly, and subjects were asked to rate them. The subjects were asked to rate each ice cream image using two questions when rating the images. First, they are rated on one food preference component ("How happy would it be if you taste this now?") and then on the other ("How much would you want some of this now?"), in random order. Tasks 3 and 4 followed the same procedure as Tasks 1 and 2. Task 3 consisted of passively viewing 16 ice-cream videos on two separate occasions, and Task 4 evaluated the stimuli of the videos presented in Task 3.

During passive viewing of the images or videos, subjects were presented with the ice cream image or videos for 10 seconds, sufficient to capture the salivary response. After passively viewing each image or video, subjects were asked to rate their feel of "liking" ("How happy would you be if you ate this food now?") or "wanting" ("How much would you like this food now?") explicitly using a 100-point visual analog scale (VAS). In total, 32 passive viewings and ratings were conducted, 16 of which were explicit " liking" ratings and 16 of which were explicit "wanting" ratings. Before each passive viewing, participants were presented with a fixation target[26] with a randomly varying exposure time (5s-7s) to ensure that their gaze was directed to the center of the screen and that participants could not accurately predict the onset of the next stimulus. Before rating, subjects were presented with a fixation target to ensure they were looking at the center of the screen.

2.2.5. fNIRS Data Collection

In this study, we focused on the prefrontal and parotid brain regions. During the experiment, data were collected by measuring each subject's cerebral blood flow data in each task. When measuring cerebral blood flow in the prefrontal cortex, subjects' resting cerebral blood flow was measured for 5 minutes before the experiment. When measuring the cerebral blood flow for resting, subjects were instructed to sit on a chair, close their eyes, and relaxed as much as possible. The cerebral blood flow data of resting for five minutes were used as training data for fNIRS data processing.

2.3. Data Analysis

2.3.1. fNIRS Data Processing (saliva)

The raw data of the cerebral blood flow on saliva includes HbO, HbR, and HbT, where HbT is the sum of HbO and HbR. For the processing of fNIRS data in the parotid region, we first convert the data output from the machine to Shared Near Infrared Spectroscopy Format (SNIRF), developed by the Society for functional Near Infrared Spectroscopy [27]. Then, a general linear model (GLM)[28] was employed for processing the fNIRS data in the parotid region using the open-source HOMER3 toolbox[29] in MATLAB (Mathworks Inc.). The procedure was as follows: Firstly, we used the prune channel function to prune bad or low SNR channels from the measurement list. Next, the raw light intensity signal was converted to optical density. Bandpass filtering was applied to the fNIRS data in the 0-0.5 Hz range. The optical density was then converted to total hemoglobin (HbT) concentration with a default partial path length factor. Finally, a general linear model (GLM) was used to remove whole-body artifacts. The hemodynamic response function (HRF) was estimated using the mean of the long separation channels. A continuous Gaussian function (stdev=1.0, step=1.0) was calculated from rest to speech.

2.3.2. fNIRS Data Processing (PFC)

For the processing of fNIRS data in the prefrontal region, we converted the data to the Shared Near Infrared Spectroscopy Format (SNIRF). Then, the general linear model with temporally embedded canonical correlation analysis (GLM with tCCA)[30] was conducted using the open-source HOMER3 toolbox in MATLAB (Mathworks Inc.). Firstly, the prune channel's function was used to prune bad or low SNR channels from the measurement list. Then we converted the raw optical intensity signal to an optical density. Subsequently, both fNIRS and auxiliary data were conducted the bandpass filtering within 0-0.5 Hz. Then, we converted the optical density to total-hemoglobin (HbT) concentration with default partial path-length factor. Now, to reduce nuisance signals in fNIRS and create optimal regressors, the temporally embedded Canonical Correlation Analysis function was conducted. Finally, a general linear model (GLM) was used to remove the systemic artifacts. The hemodynamic response function (HRF) was estimated using the average of the long separation channels. It was calculated with the consecutive sequence of Gaussian functions (stdev=1.0, step=1.0) from rest to the speech period.

2.3.3. Statistics Analysis

Subjective rating scores for each ice cream were collected during the subject's experiment, and the mean value of each rating score was calculated. After processing the cerebral blood flow data, the mean value and the change of cerebral blood flow signal for each channel were calculated. After confirming the normality of the data, correlations were calculated between subjects' subjective evaluation data and objective fNIRS data.

3. Results

3.1. Results of Ice cream Evaluation

The evaluation results for ice cream are shown in Table1. In addition, photos and videos of ice cream were categorized by state, and the mean value of each rating result was calculated.

Comparing ice cream by state, when subjects were asked how happy they would be if they tasted the ice cream at that moment in response to a photo of ice cream, the mean rating of ice cream in

each state was 56.59 for State 1 and 15.04 for State 2, 9.52 for State 3, and 9.36 for State 4. When subjects were asked how much they would like to eat the ice cream in the picture, the mean rating of ice cream in each state was 63.59 for state 1 and 22.53 for state 2, 8.77 for state 3, and 5.43 for state 4, on the other hand, when subjects were asked how happy they would be if they ate the ice cream in response to a video of ice cream, the mean rating of ice cream in each state was 52.59 for state 1 and 28.22 for state 2, 9.77 for state 3, and 5.89 for state 4, when subjects were asked how much would you like this food now in response to a video of ice cream, the mean rating of ice cream in each state was 56.41 for state 1 and 26.84 for state 2, 8.23 for state 3, and 1.89 for state 4.

Table 1: Rating results for ice cream divided by state

State 1

	State 1	**State 2**	**State 3**	**State 4**
				
Liking-Photo	56.59	15.04	9.52	9.36
Wanting-Photo	63.59	22.53	8.77	5.43
Liking-Video	52.59	28.22	9.77	5.89
Wanting-Video	56.41	26.84	8.23	1.89

Note: The number in the table is the mean value of the evaluation

3.2. Results of Hemodynamic response function (HRF)

After processing the fNIRS data, the hemodynamic responses of cerebral blood flow in the prefrontal and parotid regions were investigated during the visual stimulation and evaluation of the ice cream. There is a positive correlation between the hemoglobin signal of cerebral blood flow in the parotid region and salivary secretion, meaning that increased cerebral blood flow activity tends to increase saliva secretion. The results of activation in prefrontal and parotid regions by visual stimulation of ice cream in each state are shown in Figure 5 to Figure 7. However, little activation trend was observed in cerebral blood flow in the prefrontal and parotid regions of the subjects to the stimulation of ice cream in State 3, which is not shown in the results.

3.2.1. Results of Hemodynamic response function (HRF) in the parotid region

Figure 5 and Figure 6 shows the results of the brain activation trend observed in the parotid region during the visual stimulation and rating of ice cream. The horizontal axis shows the time course, and the vertical axis shows the average change in cerebral blood flow. The purple dotted line shows the group-averaged mean cerebral blood flow change, and the error bars show the standard error of the group-level mean. In addition, cerebral blood flow activity from -3s to 5s was observed. Time 0 on the horizontal axis represents the stimulation or evaluation start time. Therefore, the mean value of 3s before the start time of stimulation or evaluation was used to estimate the change in cerebral blood flow during 5s from the start time of stimulation or evaluation.

Figure 5 shows the trend of the hemodynamic response function to the stimulation of photos for each ice cream state. According to the results of the hemodynamic response function, when subjects were asked how happy they would be if they tasted the food in response to a photo of ice cream, we observed a tendency for a stimulation-induced of increase in cerebral blood flow in the left parotid region when subjects were evaluating photos of ice cream in state 1 and state 4, and when they were receiving stimulation of the photo in state 4. Moreover, when subjects were asked how much they would like to eat this food in response to a photo of ice cream, we observed a tendency for cerebral blood flow due to stimulation to increase in the right parotid region when they were receiving stimulation of the ice cream photo in state 1. Furthermore, when they were receiving stimulation of the photo in state 4, an increase of the cerebral blood flow was observed in the left parotid region during the stimulation.

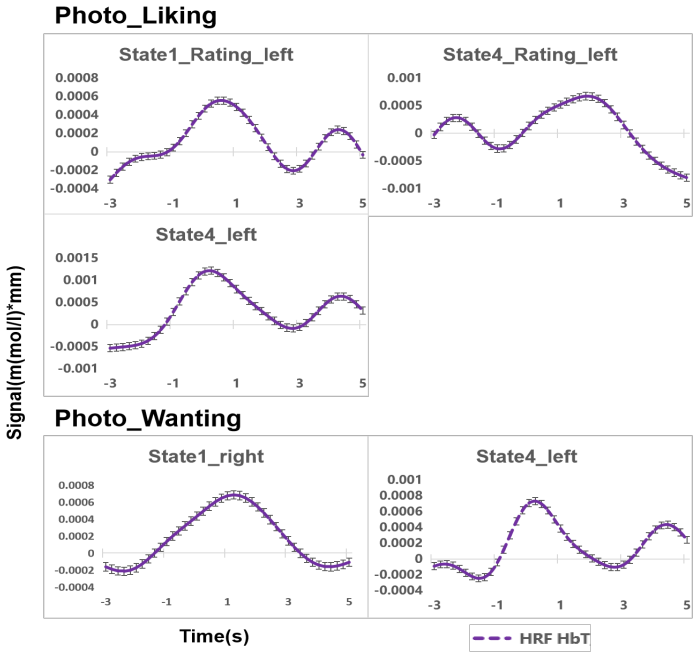


Figure 5: The trend of hemodynamic response function time courses of parotid region to photo stimuli in different states

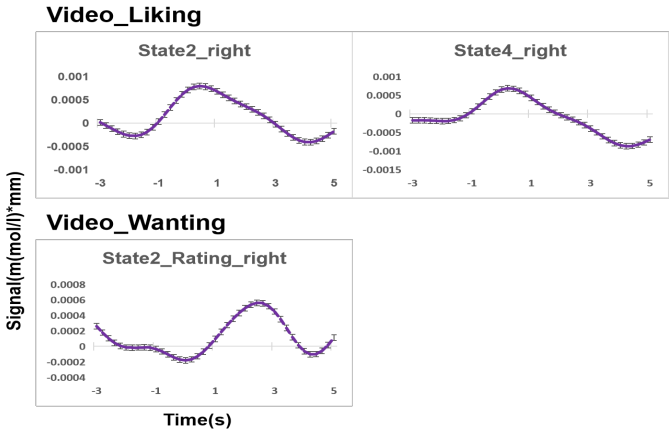


Figure 6: The trend of hemodynamic response function time courses of parotid region to video stimuli in different states.

Figure 6 shows the trend of the hemodynamic response function to the stimulation of videos for each ice cream state. When subjects were asked how happy they would be if they tasted this in response to a video of ice cream, we observed a tendency for a stimulation-induced of increase in cerebral blood flow in the right parotid

region when subjects were evaluating videos of ice cream in state 2 and state 4. Furthermore, when subjects were asked how much they would like to eat this food in response to a video of ice cream, an increase of cerebral blood flow was observed in the right parotid region when subjects were rating videos of ice cream in state 2.

3.2.2 Results of Hemodynamic response function (HRF) in the prefrontal region

Figure 7 shows the results of the observed brain activation trend in the prefrontal cortex during the visual stimulation of ice cream and rating. The horizontal axis shows the time course, and the vertical axis shows the average change in cerebral blood flow. The red solid line shows the group-averaged mean cerebral blood flow change, and the error bars show the standard error of the group-level mean. In addition, cerebral blood flow activity from -2s to 5s was observed. Time 0 on the horizontal axis represents the stimulation or evaluation start time. Therefore, the mean value of 2s before the start time of stimulation or evaluation was used to estimate the change in cerebral blood flow during 5s from the start time of stimulation or evaluation.

According to the results of the hemodynamic response function, when subjects were asked how much they would like to eat this food in response to a photo of ice cream, we observed a tendency for cerebral blood flow due to stimulation to increase in the left prefrontal region when they were receiving stimulation of the ice cream photo in state 2 and state 4. Moreover, when subjects were asked how happy they would be if they tasted the food in response to a video of ice cream, a tendency for cerebral blood flow due to stimulation to increase in the right prefrontal region when they were receiving stimulation of the ice cream photo in state 2. Furthermore, when subjects were asked how much they would like to eat this food in response to a video of ice cream, a tendency for cerebral blood flow to increase was observed in the left prefrontal region when subjects were rating videos of ice cream in state 1.

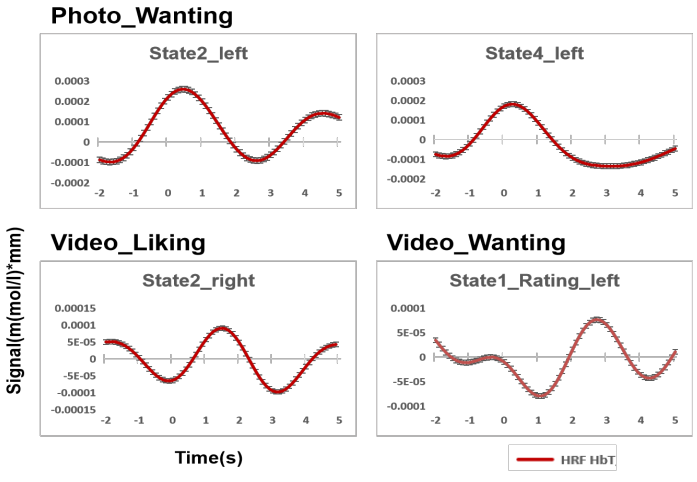


Figure 7: The trend of hemodynamic response function time courses of prefrontal region to photo and video stimuli in different states.

3.3. Results of Correlation

After processing the fNIRS data, the mean value of concentrations of HbR, HbO, and HbT were calculated. Then, the mean values were used to analyze the correlations, and the results are shown in Table2 to Table4. Since there is a positive correlation between the hemoglobin signal in cerebral blood flow in the parotid region and saliva secretion, a high level of hemoglobin signal in cerebral blood flow in the parotid region means a high level of saliva secretion.

Table 2: Correlation coefficients between subjects' ratings and the signal mean of cerebral

		HbR		HbO		HbT	
		left	right	left	right	left	right
Liking -Photo	stimuli	0.510	0.382	0.821**	0.553	0.726**	-0.520
	evaluate	0.636	-0.733**	-0.601	-0.868**	0.177	-0.942**
Wanting -Photo	stimuli	-0.878**	0.271	0.378	0.991**	-0.183	0.571
	evaluate	-0.831**	-0.952**	-0.350	-0.171	0.820**	-0.311
Liking -Video	stimuli	0.957**	0.573	0.453	0.070	0.228	-0.598
	evaluate	0.430	0.537	0.778**	0.269	0.419	-0.464
Wanting -Video	stimuli	0.076	-0.241	-0.178	-0.349	0.484	-0.174
	evaluate	0.668	0.420	-0.76**	0.597	-0.479	0.313

note: **: p<0.01
The numbers in the table refer to correlation coefficients.

Table 2 shows the correlation coefficients between subjects' ratings and the change in the signal mean of cerebral blood flow related to saliva were determined separately for each ice cream state. When subjects were asked how happy it would be if they tasted this in response to a photo of ice cream, a strong positive correlation was found between subjects' ratings and HbO (r = 0.821) and HbT (r = 0.726) on the left side during stimulation. Therefore, it can be said that the mean values of HbO and HbT on the left side tend to increase as the subject's evaluation of the ice cream increases. On the other hand, there was a strong negative correlation between the subject's ratings and the HbR (r = -0.733), HbO (r = -0.868), and HbT (r = -0.942) on the subject's right side while rating the ice cream photos. Thus, it can be concluded that the mean values of HbR, HbO, and HbT on the right side tend to become smaller as the subject's evaluation of the ice cream increases.

When subjects were asked how much they would like this food in response to a photo of ice cream, there was a strong negative correlation between the subject's ratings and HbR (r = -0.878) on the left side and a strong positive correlation with HbO (r = 0.991) on the right side during stimulation. Thus, it can be concluded that the mean value of HbR on the left side tends to become smaller and the mean value of HbO on the right side tends to become larger as the subject's evaluation of the ice cream increases. On the other hand, when subjects rated the stimulus pictures, there was a strong negative correlation between HbR on the left (r = -0.831) and right (r = -0.952) sides and a strong positive correlation between HbT (r = 0.820) on the left side. Therefore, it can be concluded that the mean values of HbR on the left and right sides become smaller as the subject's evaluation of the ice cream increases, but the mean value of HbT on the left side tends to increase.

When subjects were asked how happy they would be if they tasted this in response to a video of ice cream, a strong positive correlation was found between subjects' ratings and HbR (r = 0.957) on the left side during stimulation. Therefore, it can be concluded that the mean values of HbR on the left side tend to increase as the subject's evaluation of the ice cream increases. On

the other hand, there was a strong positive correlation between the subject's ratings and the HbO (r = 0.778) on the subject's left side while rating the ice cream videos. Thus, it can be concluded that the mean values of HbO on the left side tend to become larger as the subject's evaluation of the ice cream increases.

When subjects were asked how much they would like if they could eat it in response to a video of ice cream, no correlation was found between subjects' ratings and subjects' cerebral blood flow during stimulation. However, there was a strong negative correlation between the subject's ratings and the HbO (r = -0.760) on the subject's left side while rating the ice cream videos. Thus, it can be concluded that the mean values of HbO on the left side tend to become smaller as the subject's evaluation of the ice cream increases.

Table 3: Correlation between subjects' ratings and the signal mean of cerebral blood flow in the prefrontal region by state

	Photo_Liking _HbT		Photo_Wanting _HbT		Video_Liking _HbT		Video_Wanting _HbT	
	left	right	left	right	left	right	left	right
stimuli	-0.236	-0.171	-0.137	0.036	0.977**	-0.227	0.118	0.221
evaluate	0.895**	-0.252	-0.888**	-0.193	0.515	0.729**	0.134	0.601

note: **: p<0.01
The numbers in the table refer to correlation coefficients.

Table3 shows the correlation coefficients between subjects' ratings and the change in the signal mean of cerebral blood flow related to cognitive function were determined separately for each ice cream state. When subjects were asked how happy they would be if they tasted this at that moment in response to a photo or video of ice cream, a strong positive correlation was found between subjects' ratings and HbT (r = 0.895) on the left side of the prefrontal cortex while rating the ice cream photos. Moreover, a strong positive correlation was found between subjects' ratings and HbT (r = 0.977) on the left side of the prefrontal cortex during stimulation, and a strong positive correlation was found between subjects' ratings and HbT (r = 0.729) on the right side of the prefrontal cortex while rating the ice cream videos. Thus, it can be said that the higher the subject's evaluation, the larger the mean values of HbT on the left side of the prefrontal cortex tended to be while rating the ice cream photos; the higher the subject's evaluation, the larger the mean values of HbT on the right side of the prefrontal cortex tended to be while rating the ice cream videos, and the larger the mean values of HbT on the left side of the prefrontal cortex tended to be during stimulation of the ice cream videos.

On the other hand, when subjects were asked how much they would like to eat this food in response to a video of ice cream, a strong negative correlation was found between subjects' ratings and HbT (r = -0.888) on the left side of the prefrontal cortex while rating the ice cream photos. Therefore, it can be concluded that the higher the subject's evaluation, the smaller the mean values of HbT on the left side of the prefrontal cortex tended to be while rating the ice cream photos.

Table 4 shows the correlation coefficients between the change in the signal mean of cerebral blood flow related to saliva and the change in the signal mean of cerebral blood flow related to cognitive function were determined separately for each ice cream state. When subjects were asked how happy they would be if they tasted the food in response to a video of ice cream, a strong positive correlation (r = 0.843) was found between HbT on the right side of the prefrontal cortex and HbT on the right side of the parotid region during stimulation. Thus, it can be concluded that the mean values of HbT on the right side of the parotid region become larger as the the mean values of HbT on the right side of the prefrontal cortex increases.

Table 4: The correlation coefficients of the signal mean change of cerebral blood flow between the parotid and prefrontal regions (By state)

	Photo_Liking _HbT		Photo_Wanting _HbT		Video_Liking _HbT		Video_Wanting _HbT	
	left	right	left	right	left	right	left	right
stimuli	0.110	-0.474	0.541	-0.227	0.409	0.843**	0.681	-0.986**
evaluate	0.450	-0.245	-0.742**	-0.857**	-0.569	-0.077	0.323	-0.123

note: **: p<0.01 **The numbers in the table refer to correlation coefficients.**

On the other hand, when subjects were asked how much they would like this food at that moment in response to a photo of ice cream, a strong negative correlation was found between HbT on the left (r = -0.742) and right (r = -0.857) sides of the parotid region and HbT on the left and right sides of the prefrontal cortex while rating the ice cream photos. Furthermore, when subjects were asked how much they would like this food in response to a video of ice cream, there was a strong negative correlation (r = -0.986) between the HbT on the subject's right side of the parotid region and the HbT on the subject's right side of the prefrontal cortex while rating the ice cream videos. Therefore, it can be concluded that the mean values of HbT on the left and right sides of the parotid region become smaller as the mean values of HbT on the left and right sides of the prefrontal cortex increases, and the mean values of HbT on the right side of the parotid region become smaller as the mean values of HbT on the right side of the prefrontal cortex increases.

3.4. Summary of the results

In this study, to explore the individual motivations behind food choices and how biological responses are related to evaluations of food, we employed an approach for combining biological indicators from prefrontal and parotid regions and examined them in response to visual stimuli of food. We developed and tested a biological food preference task that simultaneously rates biological responses to visual stimuli of various food states and subjective evaluations of food. We also examined how these biological responses are related to food preference behavior. First, the results revealed the differences in subjects' subjective evaluations of the visual stimuli of different states of ice cream. Then, according to the hemodynamic response function of cerebral blood flow during visual stimulation and evaluation, activation was observed in the prefrontal and parotid regions of the brain during the task. Furthermore, we found significant correlations between subjects' subjective evaluation and the mean change in cerebral blood flow in the prefrontal and parotid regions and between the mean change in cerebral blood flow in the prefrontal region and the mean change in cerebral blood flow in the parotid region.

4. Discussion

According to the result of the subjects' subjective evaluations, when the subjects' evaluations were segmented according to the

state of the ice cream, there was a large difference in the subjects' subjective evaluation of each state of the ice cream. In particular, State 1 ice cream had the highest evaluation, and State 4 ice cream had the lowest evaluation. Consequently, when the state of the ice cream changed from hard to completely melted, the subjects' evaluations tended to change from high to low. The visual sense plays an essential role in food selection [2], and some studies have reported that the visual sense of food strongly influences appetite [31–33]. In this study, we controlled the visual changes in food state as cues for visual stimuli of food. During the experimental task, subjects visually process food cues, and along with learned knowledge about the food, there is likely to be a process of evaluating the food as expectations about the food are generated from the visual cues of the food. The fact that there was a significant difference in the evaluation of ice cream by the state is due to the fact that ice cream in a hard state is commonly expected, and as the state of ice cream changes, the subjects tend to feel more uncomfortable with the melting ice cream. Therefore, the subject's evaluation of the ice cream was also lower. Furthermore, many studies have reported that visual changes with aging significantly impact appetite [33–35]. Particularly, appetite is strongly influenced by visual factors such as the state of food [34,35]. In our experiment, the influence of state is significant, which indicated the same results as previous studies.

It has been reported that subjects process visual information about food and that, along with learned knowledge about food, expectations about food are generated from visual cues of food, and various predictive responses can be caused in the body to visual stimuli of food [36]. By detecting changes in oxygen saturation, fNIRS can identify the activation of different brain regions in response to a task or stimulus and has been widely used in many research fields, including cognitive behavior and brain injury[37,38]. The activation of the cerebral cortex causes the concentration of HbT and HbO to increase while the concentration of HbR decrease [39]. In this experiment, biological responses with activation of cerebral blood flow in the prefrontal and parotid regions were observed as predictive responses in the body. Therefore, visual stimulation in different states of food may have led to an increase in cerebral blood flow concentration in the prefrontal and parotid regions. Furthermore, these responses include cognitive and physiological processes [36]. In other words, measuring the different aspects of cognitive and physiological processes may provide insight into facilitating human food selection activities. As the state has an important influence on food selection, in this study, we used a food preference selection task in different states of ice cream. In addition, we asked subjects to rate their liking and wanting to measure brain blood flow responses in prefrontal and parotid regions during the task. Information from the subjects' subjective ratings and cerebral blood flow allowed us to examine how food cues from different states and colors in food categories affect cognitive and physiological processes and how these factors are related.

It was reported that vision is a major factor influencing human food selection[40], visual stimulation of food is an essential regulator of appetite, and visual stimulation of food can activate the reward center of the brain [41]. In this study, we measured the response of cerebral blood flow during passive viewing and evaluation of visual stimuli of food. The results showed that the food evaluation variables differed between each category of different food states. For example, when subjects were asked how happy they would be if they ate the food in response to a photo of ice cream in different states, a strong negative correlation was observed between the mean change in HbO in the right parotid region and the subject's evaluation. On the other hand, when subjects were asked how much they would like to eat this food in response to a photo of ice cream in different states, a strong positive correlation between the subject's evaluation and the mean change in cerebral blood flow in the left parotid region while rating the pictures detected, but a strong negative correlation with the mean change in cerebral blood flow in the left prefrontal region is observed. Although our results revealed various correlations between subjective evaluation and cerebral blood flow from the approach of food in different stats, the results are inconsistent, and there is currently a limited number of comparable studies on the results of each of the correlations revealed. Some studies demonstrated the differences in biological responses to the taste of different juices[42] and the olfactory cues of smell [43]. In the future, it is conceivable that it may be possible to evoke more cerebral blood flow responses by approaching elements such as the smell and taste of food compared to visual stimuli to food cues.

It was reported that increased cutaneous electrical activity in response to food aversive visual stimuli[42]. Since the materials used in this experiment ranged from normal ice cream shape and state to completely melted ice cream, it is likely that as the ice cream melted, subjects' feelings of liking or disliking each state of ice cream would also change. Similarly, subjects' biological responses to each feeling would change. Based on the results of this experiment, it was found that the subjects' subjective rating scores for each state of ice cream also decreased as the ice cream melted. On the other hand, the characteristics of cerebral blood flow in the prefrontal and parotid regions for each state of ice cream showed that changes in cerebral blood flow did not change with the ice cream state but were somewhat irregular. Furthermore, the characteristics of cerebral blood flow caused by visual stimulation of ice cream revealed differences in cerebral blood flow changes between the left and right parotid brain regions. However, to the best of our knowledge, although the components of saliva secreted by the left and right parotid regions were different, the significance of the differences in cerebral blood flow changes in the left and right parotid regions is still unknown and needs further investigation.

From the hemodynamic response function results and the correlation between the subject's subjective evaluation and cerebral blood flow, it can be inferred that there is more activity in the left hemisphere of the subject's brain. The visual stimuli used in this study led to rational judgments by the subjects, which may have resulted in more activity in the left hemisphere of the brain in the results of this experiment. However, if it is a whole-brain response, then not only visual stimuli but also olfactory and gustatory sensations are essential in the whole-brain response. The appetite examined in this experiment using only visual stimuli is different from the appetite of daily life and is worth improving in the future.

5. Future Works

In the future, we will increase the number of subjects and expand the subjects' age range while investigating whether this

experiment's results can be reproduced. In particular, we will investigate whether visual stimulation can improve appetite and xerostomia in the elderly. We will also examine the response of the brain when visual stimuli are combined with the senses of smell and taste.

References

[1] K.C. Lai, Y.Q. Liu, Q.K. He, M. Yi, T. Fujinami, "Saliva Secretion as Indicator of Appetite," in ICMHI 2021: 2021 5th International Conference on Medical and Health Informatics, Association for Computing Machinery New York NY United States, Kyoto Japan: 249–253, 2021.

[2] S. Boesveldt, K. de Graaf, "The differential role of smell and taste for eating behavior," Perception, **46**(3–4), 307–319, 2017. Doi: 10.1177/03010066166855

[3] M. Levine, "Saliva," Topics in Dental Biochemistry. Springer, Berlin, Heidelberg, 203–230, 2011.

[4] A. M. L. Pedersen, C. E. Sørensen, G. B. Proctor, G. H. Carpenter, "Salivary functions in mastication, taste and textural perception, swallowing and initial digestion," Oral Diseases, **24**(8), 1399–1416, 2018. Doi: 10.1111/odi.12867

[5] G. H. Carpenter, "The secretion, components, and properties of saliva," Annual Review of Food Science and Technology, **4**, 267–276, 2013.

[6] J. R. Stokes, G. A. Davies, "Viscoelasticity of human whole saliva collected after acid and mechanical stimulation," Biorheology , **44**(3), 141–160, 2007.

[7] C. Dawes, A. M.L. Pedersen, A. Villa, J. Ekström, G. B. Proctor, A. Vissink, D. Aframian, R. McGowan, A. Aliko, N. Narayana, Y. W. Sia, R. K. Joshi, S. B. Jensen, A. R. Kerr, A. Wolff, "The functions of human saliva: A review sponsored by the World Workshop on Oral Medicine VI," Archives of Oral Biology, **60**(6), 863–874, 2015, doi: 10.1016/J.ARCHORALBIO.2015.03.004.

[8] R. D. Mattes, "Physiologic Responses to Sensory Stimulation by Food: Nutritional Implications," Journal of the American Dietetic Association, **97**(4), 406–413, 1997, doi:10.1016/S0002-8223(97)00101-6.

[9] P. A.M. Smeets, A. Erkner, C. de Graaf, "Cephalic phase responses and appetite," Nutrition Reviews, **68**(11), 643–655, 2010, doi:10.1111/j.1753-4887.2010.00334.x.

[10] M.A.Zafra, F. Molina, A. Puerto, "The neural/cephalic phase reflexes in the physiology of nutrition," Neuroscience & Biobehavioral Reviews, **30**(7), 1032–1044, 2006, doi:10.1016/J.NEUBIOREV.2006.03.005.

[11] C. SPENCE, "MOUTH-WATERING: THE INFLUENCE OF ENVIRONMENTAL AND COGNITIVE FACTORS ON SALIVATION AND GUSTATORY/FLAVOR PERCEPTION," Journal of Texture Studies, **42**(2), 157–171, 2011, doi: 10.1111/j.1745-4603.2011.00299.x.

[12] M.P. Lasschuijt, M. Mars, C. de Graaf, P.A.M. Smeets, "Endocrine Cephalic Phase Responses to Food Cues: A Systematic Review," Advances in Nutrition, **11**(5), 1364–1383, 2020, doi:10.1093/advances/nmaa059.

[13] R. A. de Wijk He, W. He, M.G. Mensink, R.H. Verhoeven, C. de Graaf, "ANS responses and facial expressions differentiate between the taste of commercial breakfast drinks," PloS One, **9**(4), 2014.

[14] P.C. Lee, J.B. Dixon, "Food for Thought: Reward Mechanisms and Hedonic Overeating in Obesity," Current Obesity Reports, **6**(4), 353–361, 2017, doi:10.1007/s13679-017-0280-9.

[15] H. Münzberg, E. Qualls-Creekmore, S. Yu, C.D. Morrison, H.R. Berthoud, "Hedonics Act in Unison with the Homeostatic System to Unconsciously Control Body Weight," Frontiers in Nutrition, 3, 2016.

[16] C. Gibbons, M. Hopkins, K. Beaulieu, P. Oustric, J.E. Blundell, "Issues in Measuring and Interpreting Human Appetite (Satiety/Satiation) and Its Contribution to Obesity," Current Obesity Reports, **8**(2), 77–87, 2019, doi:10.1007/s13679-019-00340-6.

[17] L. Bell, J. Vogt, C. Willemse, T. Routledge, L.T. Butler, M. Sakaki, "Beyond Self-Report: A Review of Physiological and Neuroscientific Methods to Investigate Consumer Behavior," Frontiers in Psychology, **9**, 2018.

[18] D. Ferriday, J.M. Brunstrom, "'I just can't help myself': effects of food-cue exposure in overweight and lean individuals," International Journal of Obesity, **35**(1), 142–149, 2011, doi:10.1038/ijo.2010.117.

[19] G.H. Carpenter, "The secretion, components, and properties of saliva," Annual Review of Food Science and Technology , **4**, 267–276, 2013.

[20] Y. ILANGAKOON, G.H. CARPENTER, "IS THE MOUTHWATERING SENSATION A TRUE SALIVARY REFLEX?," Journal of Texture Studies, **42**(3), 212–216, 2011, doi:https://doi.org/10.1111/j.1745-4603.2011.00290.x.

[21] Neu co., BRAIN ACTIVITY MEASUREMENT , Dec. 2022.

[22] D.T. Delpy, M. Cope, P. van der Zee, S. Arridge, S. Wray, J. Wyatt, "Estimation of optical pathlength through tissue from direct time of flight measurement," Physics in Medicine & Biology, **33**(12), 1433, 1988, doi:10.1088/0031-9155/33/12/008.

[23] Neu co., WOT-S20, Dec. 2022.

[24] P. Oustric, D. Thivel, M. Dalton, K. Beaulieu, C. Gibbons, M. Hopkins, J. Blundell, G. Finlayson, "Measuring food preference and reward: Application and cross-cultural adaptation of the Leeds Food Preference Questionnaire in human experimental research," Food Quality and Preference, **80**, 103824, 2020, doi:10.1016/J.FOODQUAL.2019.103824.

[25] M. Dalton, G. Finlayson, "Psychobiological examination of liking and wanting for fat and sweet taste in trait binge eating females," Physiology & Behavior, **136**, 128–134, 2014, doi:10.1016/J.PHYSBEH.2014.03.019.

[26] L. Thaler, A.C. Schütz, M.A. Goodale, K.R. Gegenfurtner, "What is the best fixation target? The effect of target shape on stability of fixational eye movements," Vision Research, **76**, 31–42, 2013, doi:10.1016/J.VISRES.2012.10.012.

[27] The Society for functional Near Infrared Spectroscopy, SNIRF, Dec. 2022.

[28] A. von Lühmann, A. Ortega-Martinez, D.A. Boas, M.A. Yücel, "Using the General Linear Model to Improve Performance in fNIRS Single Trial Analysis and Classification: A Perspective.," Frontiers in Human Neuroscience, **14**, 30, 2020, doi:10.3389/fnhum.2020.00030.

[29] T.J. Huppert, S.G. Diamond, M.A. Franceschini, D.A. Boas, "HomER: a review of time-series analysis methods for near-infrared spectroscopy of the brain," Applied Optics, **48**(10), D280–D298, 2009, doi:10.1364/AO.48.00D280.

[30] A. von Lühmann, X. Li, K.R. Müller, D.A. Boas, M.A. Yücel, "Improved physiological noise regression in fNIRS: A multimodal extension of the General Linear Model using temporally embedded Canonical Correlation Analysis," NeuroImage, **208**, 116472, 2020, doi: 10.1016/J.NEUROIMAGE.2019.116472.

[31] K. Tomita, "Effects of the color of Tablecloths on the Sence of Taste," Journal of the Color Science Association of Japan, **28**, 38–39, 2004.(in Japanese)

[32] M. Mitsusada, T. Ohkusa, "Effects of brightness of lighting in room on appetite in elder people," 2019.(in Japanese)

[33] S. Kobayashi, "Appetite for food illuminated with vivid color lights," Architectural Institute of Japan's Journal of Environmental Engineering, **74**(637), 271–276, 2009.(in Japanese)

[34] T. Feiniwa, T. Ueda, A. Hamada, "Research on food distinction and preference through visual sense of the elderly," Journal of the Japanese Society for Culinary Science, **41**(1), 35–41, 2008.(in Japanese)

[35] M. Mitsusada, "Examination of the effects of food morphology on the mental and physical health of elderly people: from autonomic nervous system responses and subjective evaluation," Journal of the Japanese Society for Eating and Swallowing Rehabilitation, **17**(3), 226–232, 2013.(in Japanese)

[36] K.N. Boutelle, M.A. Manzano, D.M. Eichen, "Appetitive traits as targets for weight loss: The role of food cue responsiveness and satiety responsiveness," Physiology & Behavior, **224**, 113018, 2020, doi:10.1016/J.PHYSBEH.2020.113018.

[37] S. Benavides-Varela, D. Gómez, J. Mehler, "Studying neonates' language and memory capacities with functional near-infrared spectroscopy," Frontiers in Psychology, 2, 2011.

[38] H. Obrig, "NIRS in clinical neurology — a 'promising' tool?," NeuroImage, **85**, 535–546, 2014, doi:10.1016/J.NEUROIMAGE.2013.03.045.

[39] T. Ma, H. Lyu, J. Liu, Y. Xia, C. Qian, J. Evans, S. He, "Distinguishing bipolar depression from major depressive disorder using fnirs and deep neural network," Progress In Electromagnetics Research, **169**, 73–86, 2020.

[40] L.N. van der Laan, D.T.D. de Ridder, M.A. Viergever, P.A.M. Smeets, "The first taste is always with the eyes: A meta-analysis on the neural correlates of processing visual food cues," NeuroImage, **55**(1), 296–303, 2011, doi:10.1016/J.NEUROIMAGE.2010.11.055.

[41] L. Nummenmaa, J.K. Hietanen, M.G. Calvo, J. Hyönä, "Food Catches the Eye but Not for Everyone: A BMI–Contingent Attentional Bias in Rapid Detection of Nutriments," PLOS ONE, **6**(5), e19215-, 2011.

[42] L. Danner, S. Haindl, M. Joechl, K. Duerrschmid, "Facial expressions and autonomous nervous system responses elicited by tasting different juices," Food Research International, **64**, 81–90, 2014, doi:10.1016/J.FOODRES.2014.06.003.

[43] W. He, S. Boesveldt, C. de Graaf, R. de Wijk, "Dynamics of autonomic nervous system responses and facial expressions to odors," Frontiers in Psychology, 5, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 204-211 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Ensemble Extreme Learning Algorithms for Alzheimer's Disease Detection

Vanamala H R[*,1], Samriddha Shukla[1], Vijaya krishna A[2]

[1]*Department of Electronics and Communication Engineering, PES University,Bangalore, India*

[2]*Bosch Global Software Technologies Pvt.Ltd., Bangalore, India*

ARTICLE INFO

Article history:
Received: 30 August, 2022
Accepted: 17 November, 2022
Online: 20 December, 2022

Keywords:
Ensemble ELM
Ridge ELM
Bagging and Boosting ELM
Negative Correlation ELM

ABSTRACT

Alzheimer's disease has proven to be the major cause of dementia in adults, making its early detection an important research goal. We have used Ensemble ELMs (Extreme Learning Models) on the OASIS (Open Access Series of Imaging Studies) data set for Alzheimer's detection. We have explored various single layered light-weight ELM networks. This is an extension of the conference paper submitted on implementation of various ELMs to study the difference in the timing of execution for classification of Alzheimer's Disease (AD) Data. We have implemented various ensemble ELMs like Ridge, Bagging, Boosting and Negative correlation ELMs and a comparison of their performance on the same data set is provided.

1. Introduction

Advances in healthcare have made significant contributions to the longer survival and healthy lifestyle of human beings. But there are diseases that still pose a daunting challenge to the research community, Alzheimer's Disease (AD) being an important example. AD is a major neuro degenerative disease. The expense of Caring AD patients is quiet high. With the increase in quality of healthcare, the aging population is bound to increase and hence, we will see a greater number of people facing this disease caused by neuro degenerative changes. A treatment that follows an early detection leads to lower severity in the coming times and lowers the risk of damage. The need for having a Computer Aided Diagnosis system (CAD) for early and accurate AD detection and classification is crucial. In the past years, Multi Layer Perceptrons (MLP) have been extensively used for the computer vision analysis on medical images. But the less complex Single Hidden Layer Feed Forward Neural Networks (SLFNs) have been relatively less explored. In this research work, we focus on the SLFNs, taking a step forward in realizing the potential of the ensemble ELM algorithms with the help of iterative error optimization, bagging, boosting and correlation coefficients. This work mainly focuses on evaluation of different Ensemble ELM models and comparison of performance on OASIS imaging data set. Bagging and Boosting ELM gave us good results followed by negative correlation and Incremental ELMs. We have implemented these methods on Brain MR images unlike the different 2D data sets used in the reference implementations. As a result of the study, we intend to bring forth techniques which shall facilitate clinicians in classifying Alzheimer's experiencing patients from normal individuals in an early stage using MRI imagery.

[*]Corresponding Author: Vanamala H R, vanamalahr@pes.edu

2. Related work

This paper is an extension of the originally submitted work in the conference "2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON). [1]

We used Modified Extreme Learning Machine algorithms to study the training computational requirements in comparison to conventional machine learning methods.[2,3] The variations of ELMs used were, Regularized ELM (RELM) and MLPELM [4,5]. Convolution Neural network takes more time for training and the desired accuracy may be difficult to achieve where as ELM is a lightweight SLFN with randomly initialized biases and weights, which when propagated onto the next layer create a Multi-Layer Perceptron without BPA (Back propagation algorithm) with ELM as the backbone of the network, and this network delivers at a faster training speed with similar accuracy rates [6].

In our implementations we had used the ELM algorithm which randomly generates the weights and biases. It tends to have

a good generalizability and all the hidden layers are treated as a whole system. This way, once the feature of the previous hidden layer is extracted, the weights or parameters of the current hidden layer will be fixed and need not be fine-tuned.Therefore, this helps in getting a better accuracy and trains the network faster compared to using MLP with back propagation algorithms. It was observed that Multi Layer Perceptron had the highest accuracy while the training speed was slower, meanwhile, the RELM had lesser accuracy compared to Multi Layer Perceptron but the training speed was faster. There is a trade off between training speed and accuracy [7]-[9].

Lot of research on new approaches with ELMs are done. Dual-tree complex wavelet transforms (DTCWT), Principal Component Analysis(PCA), Linear Discriminant Analysis(LDA), and ELM implementations were used to identify AD conditions. On ADNI dataset, they could get an accuracy of 90%, specificity of 90% and sensitivity of 90% [9]. Another approach where Key Features Screening method based on ELMs (KFS-ELM) was implemented From 920 key functional connections screened from 4005 the accuracy is obtained for detecting AD was 95.33% [10].

Many state-of-the art techniques like Support Vector Machine (SVM), Random Neural Network (RNN), Radial Basis Function Neural Network (RBFNN), Hopfield Neural Network (HNN), Boltzmann Machine (BM), Restricted Boltzmann Machine (RBM), Deep Belief Network (DBN), and other DL methods are compared with ELM implementations like Circular ELM (CELM) and Bootstrap aggregated (B-ELM) and other variants. The observations are that ELMs are faster than these techniques . Many ELM variations are being used and they are comparatively more robust. Implementation is simple and performance in terms of accuracy is also considerably better [11]-[13].

Considering the advantages of ELMs we have incorporated the less utilized ensemble ELMs with necessary modifications on OASIS Brain MR Data set and study the classification performance.

3. Data set

The data set is acquired from the official OASIS repository. The data set used is OASIS-1. The data is made up of 416 subjects with their respective cross-sectional scans from 434 scan sessions. The number of T1-weighted MRI scans varies from 3 to 4 in a single-imaging session. The data acquired is restricted to right-handed individuals of both genders. The number of participants with an age above 60 with diagnosis of AD at various stages is 100.

Every imaging session was stored in its respective directory labeled with its subject ID. The subject ID format is OAS1_xxxx, where 'xxxx' denotes the 4digit ID of a patient(eg: OAS1_0027). All the sessions have been assigned with an ID formatted as OAS1_xxxx_MRy, here y denotes the session number of the participant (eg: OAS1_0034_MR1). A particular session is accompanied by a zip-compressed file with its respective session ID. We have considered totally 1412 images of which 179 Mild Demented, 145 Moderately Demented, 640 Non Demented and 448 Very Mild Demented, from the data set after Preprocessing.

4. Implementation

4.1. Proposed Method

Our objective here is to incorporate the less explored ensemble ELMs with relevant modifications on Brain MR Data and study the classification performance on the OASIS Dataset.

The same Preprocessing steps used in our work earlier [1] are used for processing the images. The steps involved are initially Preprocessing, segmenting and feature extracting leading the vectors into our desired ELM classifier. The sample images and the processed images are shown below in the results section in Figure 2 and Figure 3. Biological implications of AD is that it results in two distinctive abnormalities in the brain, firstly Neuro-fibrillary tangles and secondly senile plaques. A Neuro-fibrillary tangle gets originated in the cytoplasm of neurons in the entorhinal cortex, whereas the plaques found in the neocortex of the brain. Structural changes in the early stages are more dominant in the medial temporal lobe, especially in the entorhinal cortex. With progression of the disease both entorhinal cortex and hippocampus are the indications of the variations observed.

4.1.1. Image Preprocessing

Firstly the unrefined images are denoised and given a level of uniformity. The primary goals are echo reduction, contrast imbalance correction and image resolution rectification. A brain image is primarily comprised of Cerebrospinal Fluid (CSF), White Matter (WM) and Grey Matter (GM). The overview of the situation is that the precision with which the Regions of Interest (RoI) are ascertained among the noise created by the aforementioned features even after threshold removal of skull. We use filtering to make our classification efforts easier as it gets rid of low frequency echoes. A fixed re sizing of 256x256 and grayscale conversions are made. The finalized Preprocessing method chosen upon is median filtering which is a nonlinear method of noise elimination. It functions by iterating through each pixel and substituting each pixel with the median value of the pixels in a certain proximity range all while preserving the edges. It works on the principle of averaging and which effectively eliminates noise and blurring of sharp edges [14].

4.1.2. Image Segmentation

The objective of this process is to segregate the image into non intersecting regions based on similarity indexes. The segmentation technique adopted here is threshold method which relies on normalizing a pixel with respect to its intensity level. It is optimal for object and background separation which comes in handy for skull removal here [15,16].

4.1.3. Feature Extraction

It involves the recreating values by deriving them via consolidation of a large collection of data in order to bring generalization while maintaining its relevance to the task intended in future. Our focus here is texture analysis based on understanding the nature of spatial variation. with the Gray-Level Co-occurrence Matrix (GLCM). We have used the features extracted for earlier algorithms [17]. GLCM is a square matrix representing the frequency count of appearance of a reference pixel with a particular intensity value i with respect to an adjacent

neighbor pixel having an intensity value of j. Therefore, a cell (i,j) represents their cumulative occurrence count. Their separation distance, d is also taken into account for the GLCM function computation along with regional frequency count. Further a matrix consisting of statistical values to increase the descriptive understanding is created [18].

Our earlier implementation was of RELM and MLPELM [19,20]. Further we have worked on the performance of various Ensemble ELMs on Alzheimer's data. Various ELMs with sigmoid , hardlim, gaussian and leaky Relu functions which helps in speeding up of the training for classification and Regression. Incremental Learning and increase in hidden nodes assist in minimizing the error. Multiclass Adaboost ELM (MAELM), Bagging Stepwise ELM and Negative Correlation ELM are suitable for imbalanced data. They are simple, cost effective, supporting many kernels or feature mapping functions for multi class data.Knowing ELMs require lesser time than conventional networks for training from our study performed earlier[1], our objective here is to incorporate the above modifications to ELMs and study the classification performance on the same OASIS Dataset.We also have compared the performance of other state of art techniques like CNNs and SVM. We have studied different Ensemble ELMs like ridge bagging, boosting, (Adaboost ELM)AELM and Negative Correlation ELMs(Incremental ELM) IELM.

4.2. Algorithms

We have implemented different ensemble ELMs and the this section will give the mathematical treatment and the methodology of implementation of these algorithms.

4.2.1. Vanilla ELM (VELM)

ELM for the SLFNs was invented in 2006.The Algorithm 1 represents the basic ELM flow. The Vanilla Extreme Learning Machine network architecture is shown in Figure 1.

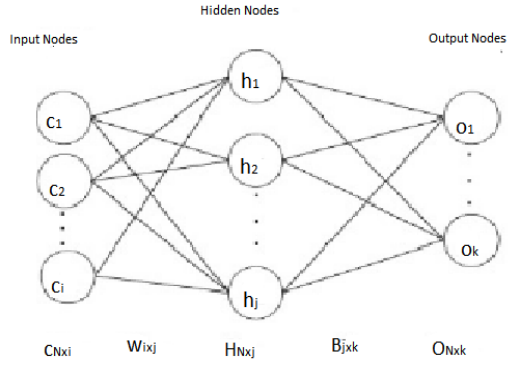


Figure 1: ELM architecture.

Assume N distinct samples (c_r, o_r) where $c_r \in R_i$ and $o_r \in R_k$, r =1,…,N. The output hidden layer matrix H is presented as

$$H = \begin{bmatrix} G(w_1, b_1, c_1) & \cdots & G(w_i, b_j, c_1) \\ \vdots & \ddots & \vdots \\ G(w_1, b_1, c_N) & \cdots & G(w_i, b_j, c_N) \end{bmatrix} \quad (1)$$

G(.) gives the activation function, $W_{N\times j}=w_1,\ldots,w_j$ represents the weight matrix between the input layer and the hidden layer and b_r represents the rth hidden node in a particular hidden layer bias. Both of the parameters in an particular layer are randomly initialized [19,20].

Weights B in the output layer are to be adjusted using the Moore–Penrose Generalized Inverse is used to compute the weights B in the between the the hidden layer and output layer

$$B=H\dagger O. \quad (2)$$

here H† denotes the Moore–Penrose inverse of matrix H and

$$O = o_1,\ldots, o_N^T \quad (3)$$

Algorithm 1

Consider

Training set $T=\{(c_r, o_r), c_r \in R_{i,r} \in R_k, r=1,.., N\}$

With G• and j representing activation function and the number of hidden nodes respectively

i: Randomly initialize desire parameters (w_r,b_r), r =1,…..j

ii: Compute output hidden layer matrix H

iii: Successively compute matrix B

iv: Resultants are B, W and b

4.2.2. Ridge ELM

Ridrin order added to the robustness of Regularized ELM algorithm by accompanying the matrix B (output hidden layer matrix) with l_2 norm through addition to minimize the matrix B constituents by a certain degree in accordance with Bartwtt's theory which is 'Smaller the weights in the output layer better the generalization performance in case of feed forward neural networks.'[21,22,23] The mathematical model is

$$M_{in} = \frac{1}{2}\sum_{i=1}^{k} \gamma_i^T \gamma_i + \lambda \frac{1}{2}\|B\|^2$$

$$where: \quad y_r = 0_r - \gamma \ , \ \gamma = 1 \ldots k \quad (4)$$

$$B_{jxk} = [\beta_1 \text{----} \beta_k] \quad (5)$$

$$Y_r = H_{\beta r} \text{ and} \quad (6)$$

$$O_{Nxj} = [o_1 \text{----} o_j] \quad (7)$$

here λ has a varying range as $\lambda \in (0,\infty)$. In the scenario of the data set being comparatively smaller, we use

$$B = H^T\left(\frac{I}{H} + HH^T\right)^{-1}T \quad (8)$$

where as for the latter scenario where N ≫ j, we opt for

$$B = (\frac{1}{\lambda} + H^T H)^{-1} H^T T \quad (9)$$

here N is the number of training samples and j is the dimension of feature space.

Algorithm 2

Consider

Training set

T=c_r, or, $c_r \in R_i$, or $\in R_k$, r=1,…,N

i: Randomly initialize desire parameters (w_r,b_r), r =1,…..j

ii: Compute output hidden layer matrix H

iii: Successively compute matrix B using Eq.8 or Eq.9

iv: resultants are B, W and b

4.2.3. Bagging ELM Algorithm

Bagging primarily consists of the processes of bootstrapping and aggregation. Bootstrap sampling is adopted to introduce the concept of experts generation which means the further division of training and validation set catered differently by each expert along with the putting back of samples. The models for all experts are trained with the basic algorithm for learning. Bagging uses voting for classification in order to aggregate [24,25,26]. Algorithm 3 explains this bagging technique.[27]

Algorithm 3
Given training set T=c_r, or, cr$\in R_i$, or $\in R_k$, r=1,…N

Learning parameter Φ and E: number of experts

To do: Aggregation of all E experts

Loop in range r = 1:*E*

(i) By bootstrap sample in *D*; get the training and validation set

(ii) train the respective expert with Φ in training set and validation set

End: Resultants are E models of experts

4.2.4. Boosting Ridge ELM

It is an ELM integrated with Boosting Ridge Regression. It consists of two initialization steps where hidden nodes are randomly initialized and the Boosting Ridge ELM chooses bias and input weights randomly. Then, a response matrix is computed for the hidden nodes. Boosting ridge then computes the output weights using boosting ridge regression. Then using feedback based on any output weights being zero, the respective hidden nodes are deleted successive to which bias and input weights matrix update is made .[28,29]The algorithm is as follows:

Algorithm 4:

Given training set T=c_r, or, cr$\in R_i$

i. Randomly initialize input weight w_i and bias bi

ii. Compute hidden layer output matrix

iii. Find output weight $\beta_{boosting\text{-}ridge}$ with the help of boosting ridge regression

$$\beta = H^T (\frac{I}{C} + WHH^T)^{-1} W^T \quad (10)$$

iv. Delete the redundant hidden node and do successive updation according to $\beta_{boosting\text{-}ridge}$

v. Adjust the parameter values and number of iterations to get best fit

4.2.5. Negative Correlation ELM

NCELM is made up of i learners, each of which is an ELM, c = 1. . . N, where N is the number of classifiers. The averaged output of a sample x ∈ R

$$p_j(x) = \frac{1}{C}\sum_{c=1}^{C} p_j{}^{(c)}(x) = \frac{1}{C}\sum_{c=1}^{C} h^{(c)^T}(x)\beta_j{}^{(c)} \quad (11)$$

The end goal is to generalize using minimization of each learner and the using penalization (λ ∈ (0, +∞)), we take care of controlling the diversity which is dynamic to the problem. The minimization can be shown as :

$$Min\ (\beta_j^i \in R^{DXJ})[(\ ||\beta_j^{(i)}||^2 + S\ ||H^{(i)}\beta_j^{(i)} + Y_j||^2 + \lambda < H^{(i)}\beta_j^{(i)}, P_j >^2\ )] \quad (12)$$

Below p_j denotes the ensemble output,

$$P_j = \sum_{c'=1}^{C} H^{(i')}.\beta_j{}^{(i')} \quad (13)$$

Because $\beta_j{}^{(c)}$ appears in p_j , the proposed solution for Eq. 12 is to transform the problem in an iterated sequence, with solution of the first iteration $\beta_j{}^{(c)}$,(1) for c = 1, . . . , C. The output weight matrices in the rth iteration β (c) j,(r) ,c = 1, . . . , C, for each individual are obtained from the following optimization problem.[30]

The output weight matrices are obtained iteratively with the rth iteration corresponding to $\beta_j{}^{(c)}$,(r) ,c = 1, . . . , C, shown as,

$$Min\ (\beta_{j(r)}^{(c)} \in R^{DXJ})\ (\ ||\beta_{j,(r)}^{(c)}||^2 + S\ ||H^{(c)}\beta_{j,(r)}^{(c)} - Y_j||^2 + \lambda < H^{(c)}\beta_{j,(r)}^{(c)}, P_{j(r-1)} >^2\ ) \quad (14)$$

updation of Pj,(r−1) takes place in this manner,

$$P_{j(r-1)} = \frac{1}{C}\sum_{c=1}^{C} H^{(c)}.\beta_{j(r-1)}^{(c)} \quad (15)$$

Solution can be obtained for Eqn.14 by deriving it and now by equating to 0.

$$\beta_{j,(r)}^{(c)} = (\frac{I}{S} + H^{(c')}H^{(c)} + \frac{\lambda}{S}H^{(c')}P_{j,(r-1)}P'_{j,(r-1)}H^{(c)})^{-1}H^{(c')}Y \quad (16)$$

Iteratively we calculate $(r+1)^{th}$ term and the convergence is assured by the Banach fixed-point theorem.[30]

4.2.6. Incremental ELM

Characteristic of IELM is that the error decreases for every iteration and is dwindled to zero with increase in number of hidden nodes. The trade off is between training time and accuracy as it is no more a one-shot process. The residual error is :

$$\mathbf{E`} = \mathbf{E} - \beta\mathbf{H} \quad (17)$$

To minimize the residual error if we are not satisfied with the minimization magnitude we restart the training of the particular node with new randomly initialized values. [31]

Algorithm 5

Given training set {(C, O)}, C is a $l \times N$ matrix which is the input of N data sets. O is a $k \times N$ matrix which is the output of N data sets

Step 1: Set the max value J for number of nodes, initial value of residuals ε and set current number of nodes j =0

Step 2: loop until j < J and E > ε :

i. Increment j by 1
ii. Randomly initialize weight ωj and bias bj of the hidden layer neuron hj
iii. Use the activation function g(c`) to calculate the output for the node o_l
c` = ω_jC + b_j
iv. Find the ouput vector H of hidden layer neurons:
H = g(c`)
v. Calculate β
vi. Re-calculate the residuals error

In the analysis 2D MRI scans were converted to 1D Numpy Arrays and then fed as input to the respective ELMs implemented.The above algorithms were implemented in Python and trained on the OASIS data set considered in our previous study as mentioned earlier. Four class classification is performed and the comparison of the performance is discussed in the results section.

5. Results

In our analyses using the 4 different variations of ELM and Vanilla ELM we predict the stages of the Alzheimer's disease which were non-demented, very mildly demented, mildly demented and moderately demented as indicated in Figure2. The various preprocessing techniques have varying effects which are visible in the Figure 3.

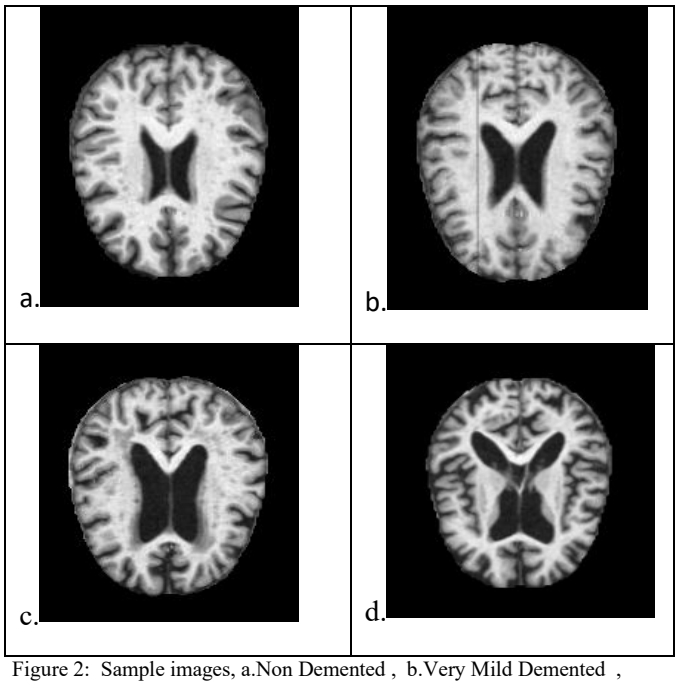


Figure 2: Sample images, a.Non Demented , b.Very Mild Demented , c.Mild Demented , d. Moderate Demented

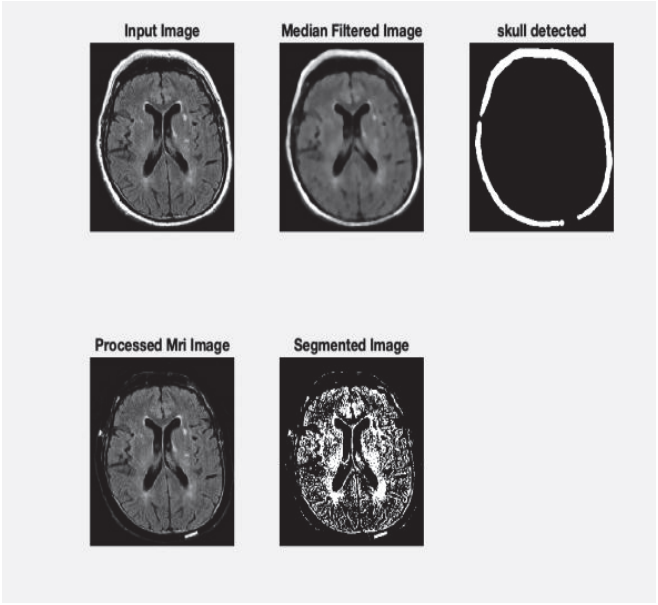


Figure 3: Sample Preprocessed images

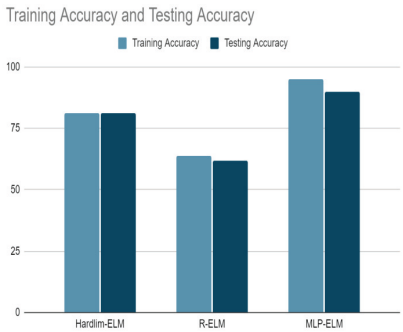


Figure 4: Accuracies of different ELMs implemented

The results obtained for the four class classification of OASIS Data for Hardlim, RELM and MLPELM implemented in the earlier work is shown in Figure4. Timing analysis was done for which MPELM gave us good performance.

The research was focused primarily on implementation of algorithms proposed by earlier researchers and then using the appropriate quantitative initialization parameters associated with them for our Data set.

Table 1 indicates the testing accuracies obtained for different state of art techniques implemented to compare the results with that of the ensemble ELMs.SVM and CNN-LSTM(Long Term Short Term Memory) gave us good results as indicated. With the advantages of ELMs , simple network and faster we find the ELMs more apt. Table 2 shows the comparison of testing accuracy for different activation functions used. Bagging and Boosting gave good results with Sigmoid and Leaky ReLU.

Table 1: Accuracies with our other implementations

Sl. No.	ELM	Testing Accuracy(%) Gaussian	Testing Accuracy (%) Sigmoid	Testing Accuracy(%) Leaky-ReLU
1	Vanilla-ELM	77.05	74.86	71.66
2	Boosting-ELM	88.42	92.32	93.47
3	Incremental-ELM	82.55	82.13	81.09
4	Bagging ELM	98.11	97.63	96%
5	Negative Correlation-ELM	83.49	87.56	84.38

Table 2: Accuracy with other activation functions

l.No.	Model	Testing Accuracy(%)
1	CNN	84.01
2	CNN-LSTM	98.23
3	Decision Tree	81.23
4	SVM	97.41

Table 3: Accuracies of different implemented ELMs

Sl. No.	ELM	Training Accuracy (%) at N=50	Training Accuracy (%) at N=5000	Testing Accuracy (%) at N=5000
1	Vanilla-ELM	53.43	74.86.	72.34
2	Boosting-ELM	62.72	92.32	90.69
3	Incremental-ELM	61.63	82.13	81.23
4	Bagging ELM	63.48	97.63	97.41
5	Negative Correlation-ELM	43.00	87.56	86.03

The bagging and boosting algorithms were tried over with different activation functions which are hardlim, triangular, sigmoid, (with beta=1, used leaky Relu to overcome vanishing gradient problem) and gaussian . We also tuned the negative correlation coefficient in accordance with the bias and the values varied in the order of 10^{-3} to 10^{-4} and the best accuracy was achieved near $6x10^{-3}$. We observed that bagging and boosting algorithms took a training time of 18-20 minutes for 5000 nodes in spite of being stepwise whereas the negative correlation ELM took 13 minutes time for training and tuning 500 nodes. The training time of negative correlation ELMs is also dependent on the designated negative correlation coefficient. The incremental ELM algorithm's residual error value acts as a threshold value for stopping which was optimized over multiple epochs for the particular data set. Training and Testing accuracies are shown for Bagging, Boosting , Incremental and NE-ELM in the Figure 5, Figure 6 and Figure 7.

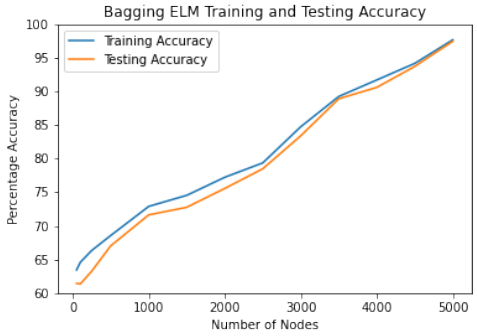


Figure 5: Accuracy of Bagging ELM

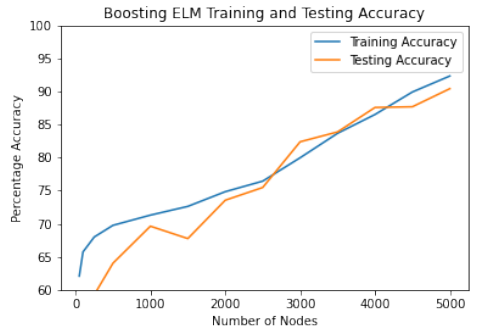


Figure 6: Accuracy of Boosting ELM

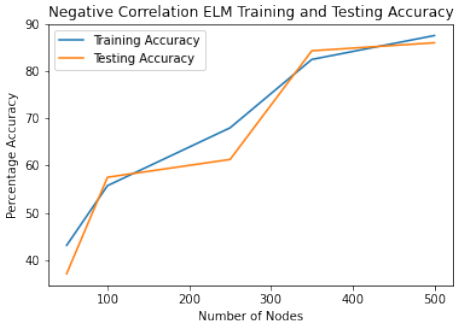


Figure 7: Accuracy of Negative-Correlation ELM

Table 3 indicates Training and testing accuracies obtained for the algorithms implemented. Bagging gave better performance, followed by boosting and NE-ELM for 5000 nodes.

6. Conclusion

We have implemented various ensemble ELMs like Ridge, Bagging, Boosting and Negative correlation ELMs and studied their performance on the OASIS data set. Bagging and Boosting ELM gave us good results followed by negative correlation and Incremental ELMs. These methods are implemented on Brain MR images unlike other 2D data sets used in the reference implementations.

Our approach results in better accuracy and convergence rate. The onset of impact of Alzheimer's disease is delayed if an early diagnosis is made possible which overall helps in prevention of the incurable disease. The regularly used machine learning algorithms such as Import Vector Machine (IVM), Support Vector Machine (SVM) and Radial Basis Functions require iterative training by adjustment of parameters to reduce the error. They are time-intensive and require manual assistance for obtaining the desired accuracy. In our proposed method, we have used the ELM algorithms which rely entirely on random initialization which creates a generalized inductive bias, thus obviating the need for fine-tuning for each layer or node addition.

7. Future Scope

The research on Ensemble ELM networks can be taken forward with the infusion of deep learning kernels such as Convolution ELMs and RNN-ELMs which promise the successful results of dense multi-layered networks without the time indulgence put into fine tuning. If algorithms like RNN-ELM which make use of temporal space weights are used along with Dual-Tree Complex Wavelet Transform (DTCWT) we stand to gain great benefits from the directional sensitivity subsiding shift in variance. It can be implemented and tested on PET (Positron Emission Tomography) Data. Can further work on improving on timing constraints [1].The time required for execution can be optimized.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments

Data set is from OASIS - OASIS-1: Cross-Sectional: Principal Investigators: D. Marcus, R, Buckner, J, Csernansky J. Morris

References

[1] A. Hari, V. Rajanna, V. A, R. Jain, "An Extreme Learning Machine based Approach to Detect the Alzheimer's Disease," 1–6, 2021, doi:10.1109/GUCON50781.2021.9573589.

[2] R.K. Lama, J. Gwak, J.-S. Park, S.-W. Lee, "Diagnosis of Alzheimer's Disease Based on Structural MRI Images Using a Regularized Extreme Learning Machine and PCA Features," Journal of Healthcare Engineering, **2017**, 5485080, 2017, doi:10.1155/2017/5485080.

[3] K.A. Johnson, N.C. Fox, R.A. Sperling, W.E. Klunk, "Brain imaging in Alzheimer disease," Cold Spring Harb. Perspect. Med., **2**(4), a006213, 2012.

[4] R.N. and G.A. Palimkar Prajyot and Shaw, "Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach," in: Bianchini Monica and Piuri, V. and D. S. and S. R. N., ed., in Advanced Computing and Intelligent Technologies, Springer Singapore, Singapore: 219–244, 2022.

[5] X. Liu, L. Wang, G.-B. Huang, J. Zhang, J. Yin, "Multiple kernel extreme learning machine," Neurocomputing, **149**, 253–264, 2015, doi:https://doi.org/10.1016/j.neucom.2013.09.072.

[6] S. Mandal, V.E. Balas, R.N. Shaw, A. Ghosh, "Prediction Analysis of Idiopathic Pulmonary Fibrosis Progression from OSIC Dataset," in 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), 861–865, 2020, doi:10.1109/GUCON48875.2020.9231239.

[7] G. Huang, G.-B. Huang, S. Song, K. You, "Trends in Extreme Learning Machines: A Review," Neural Networks, **61**, 2014, doi:10.1016/j.neunet.2014.10.001.

[8] S. Suhas, C.R. Venugopal, "MRI image preprocessing and noise removal technique using linear and nonlinear filters," in 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT), 1–4, 2017, doi:10.1109/ICEECCOT.2017.8284595.

[9] D. Jha, S. Alam, G.-R. Kwon, "Alzheimer's Disease Detection Using Extreme Learning Machine, Complex Dual Tree Wavelet Principal Coefficients and Linear Discriminant Analysis," Journal of Medical Imaging and Health Informatics, **8**, 881-890(10), 2018.

[10] J. Lu, W. Zeng, L. Zhang, Y. Shi, "A Novel Key Features Screening Method Based on Extreme Learning Machine for Alzheimer's Disease Study," Front Aging Neurosci, **14**, 888575, 2022.

[11] J. Wang, S. Lu, S.-H. Wang, Y.-D. Zhang, "A review on extreme learning machine," Multimedia Tools and Applications, **81**(29), 41611–41660, 2022, doi:10.1007/s11042-021-11007-7.

[12] S. Grueso, R. Viejo-Sobera, "Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review," Alzheimer's Research & Therapy, **13**(1), 162, 2021, doi:10.1186/s13195-021-00900-w.

[13] M. Odusami, R. Maskeliunas, R. Damaseviv cius, T. Krilavicius, "Analysis of Features of Alzheimer's Disease: Detection of Early Stage from Functional Brain Changes in Magnetic Resonance Images Using a Finetuned ResNet18 Network," Diagnostics (Basel), **11**(6), 2021.

[14] S. Leandrou, S. Petroudi, P.A. Kyriacou, C.C. Reyes-Aldasoro, C.S. Pattichis, "Quantitative MRI Brain Studies in Mild Cognitive Impairment and Alzheimer's Disease: A Methodological Review," IEEE Rev Biomed Eng, **11**, 97–111, 2018.

[15] G.V.S. Kumar, V.K. .R, "REVIEW ON IMAGE SEGMENTATION TECHNIQUES," International Journal of Scientific Research Engineering & Technology (IJSRET), **3**, 992–997, 2014.

[16] M. Al-amri, D. Kalyankar, D. S.D, "Image segmentation by using edge detection," International Journal on Computer Science and Engineering, **2**, 2010.

[17] D. Tian, "A review on image feature extraction and representation techniques," International Journal of Multimedia and Ubiquitous Engineering, **8**, 385–395, 2013.

[18] T. and S.R.N. and G.A. Sinha Trisha and Chowdhury, "Analysis and Prediction of COVID-19 Confirmed Cases Using Deep Learning Models: A Comparative Study," in: Bianchini Monica and Piuri, V. and D. S. and S. R. N., ed., in Advanced Computing and Intelligent Technologies, Springer Singapore, Singapore: 207–218, 2022.

[19] W. Deng, Q. Zheng, L. Chen, "Regularized Extreme Learning Machine," in 2009 IEEE Symposium on Computational Intelligence and Data Mining, 389–395, 2009, doi:10.1109/CIDM.2009.4938676.

[20] J. Tang, C. Deng, G.-B. Huang, "Extreme Learning Machine for Multilayer Perceptron," IEEE Transactions on Neural Networks and Learning Systems, **27**(4), 809–821, 2016, doi:10.1109/TNNLS.2015.2424995.

[21] R. Yangjun, S. Xiaoguang, S. Huyuan, S. Lijuan, W. Xin, "Boosting ridge extreme learning machine," in 2012 IEEE Symposium on Robotics and Applications (ISRA), 881–884, 2012, doi:10.1109/ISRA.2012.6219332.

[22] B. Jin, Z. Jing, H. Zhao, "Incremental and Decremental Extreme Learning Machine Based on Generalized Inverse," IEEE Access, **5**, 20852–20865, 2017, doi:10.1109/ACCESS.2017.2758645.

[23] G. Feng, G.-B. Huang, Q. Lin, R. Gay, "Error Minimized Extreme Learning Machine With Growth of Hidden Nodes and Incremental Learning," IEEE Transactions on Neural Networks, **20**(8), 1352–1357, 2009, doi:10.1109/TNN.2009.2024147.

[24] Z.X. Chen, H.Y. Zhu, Y.G. Wang, "A modified extreme learning machine with sigmoidal activation functions," Neural Computing and Applications, **22**(3), 541–550, 2013, doi:10.1007/s00521-012-0860-2.

[25] C. Perales-González, "Global convergence of Negative Correlation Extreme Learning Machine," Neural Processing Letters, **53**(3), 2067–2080, 2021, doi:10.1007/s11063-021-10492-z.

[26] P. Khan, Md.F. Kader, S.M.R. Islam, A.B. Rahman, Md.S. Kamal, M.U. Toha, K.-S. Kwak, "Machine Learning and Deep Learning Approaches for Brain Disease Diagnosis: Principles and Recent Advances," IEEE Access, **9**, 37622–37655, 2021, doi:10.1109/ACCESS.2021.3062484.

[27] M. Yang, J. Zhang, H. Lu, J. Jin, "Regularized ELM bagging model for Tropical Cyclone Tracks prediction in South China Sea," Cognitive Systems Research, **65**, 50–59, 2021, doi: 10.1016/j.cogsys.2020.09.005.

[28] A.O.M. Abuassba, D. Zhang, X. Luo, A. Shaheryar, H. Ali, "Improving Classification Performance through an Advanced Ensemble Based Heterogeneous Extreme Learning Machines," Computational Intelligence and Neuroscience, **2017**, 3405463, 2017, doi: 10.1155/2017/3405463.

[29] Y. Jiang, Y. Shen, Y. Liu, W. Liu, "Multiclass AdaBoost ELM and Its Application in LBP Based Face Recognition," Mathematical Problems in Engineering, **2015**, 1–9, 2015, doi:10.1155/2015/918105.

[30] S. Wang, H. Chen, X. Yao, "Negative correlation learning for classification ensembles," in The 2010 International Joint Conference on Neural Networks (IJCNN), 1–8, 2010, doi:10.1109/IJCNN.2010.5596702.

[31] S. Song, M. Wang, Y. Lin, "An improved algorithm for incremental extreme learning machine," Systems Science & Control Engineering, **8**(1), 308–317, 2020, doi:10.1080/21642583.2020.1759156.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 212-221 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Emerging Trends in Green Best Practices and the Impact on Government Policy

Emily Holt*, Casey Corrado

The MITRE Corporation, Bedford MA, 01730, USA

ARTICLE INFO

Article history:
Received: 22 August, 2022
Accepted: 27 November, 2022
Online: 20 December, 2022

Keywords:
Sustainability
Green Best Practices
Green Policy

ABSTRACT

While it is commonly accepted that climate change needs to be addressed to protect both human and environmental health, it is not widely understood what steps need to be taken to accomplish this daunting task. Additionally, there is currently no formal definition of what constitutes a 'green' company or 'green' best practice, despite the rising usage of the term. We found that companies that are considered 'green' have well-documented, quantifiable improvements in their sustainability plans and initiatives. These plans are published yearly in publicly available progress reports. Multi-year goals, with progress mapped from year to year, follow trends in the following areas: reduction in carbon emissions, energy obtained through renewable energy sources, amount of waste diverted from landfills, third-party certifications for buildings, water conservation, increasing 'green' requirements from suppliers, and sustainable fleet management. To address the gap between industry and government practices, and to capitalize on recent interest and investment in 'green', we recommend that all U.S. government agencies formalize and publicly release sustainability policies with quantifiable goals, identify practices to be implemented, and define metrics to measure progress. To effectively develop and implement these plans, we recommend: (1) each agency evaluate their current organization to develop a baseline, (2) define milestones and targets using the baseline as a starting point such that industry standards can be reached, and (3) release a finalized, publicly available sustainability plan.

1. Introduction

The Fourth National Climate Assessment defines climate change as the "changes in average weather conditions that persist over multiple decades or longer" and "encompasses both increases and decreases in temperature, as well as shifts in precipitation, changing risk of certain types of severe weather events, and changes to other features of the climate system" [1]. Over the last decade, climate change has become a central research area with interest and investment across both government and industry. While it is commonly accepted that climate change needs to be addressed to protect both human and environmental health, it is not widely understood what steps need to be taken to accomplish this daunting task. Inconsistencies in sustainability plans and policies between U.S. government and industry showcase the confusion and misunderstanding within the 'green' space. To properly address this issue, the government needs information on what proven and effective 'green' practices exist. With this, various approaches can be unified under a single definition to enact widespread change.

This work expands upon work previously done by the authors, going into a more detailed survey of the exploding popularity of the term 'green' [2]. In this survey, we both demonstrate and evaluate the lack of recognized, concrete definitions or metrics for this term and provide recommendations for action through policy. We will lay out what 'green' practices the U.S. government has already adopted, and, in contrast, what state-of-the-art sustainability practices are being implemented by industry. The authors will analyze the differences between the two groups, using the results as the baseline for developing the set of policy-based recommendations. The guiding research questions we will explore include: who are the industry leaders in the 'green' field; what are the 'green' best practices being utilized by these leaders; and, finally, what quantifiable metrics or features make a company 'green'?

2. Background

2.1. Key Definitions

After assessing the terminology across industry and government, a few key definitions must be articulated to describe

*Corresponding Author: Emily Holt, egholt@mitre.org

'green' best practices. It should be noted that there is no formalized, universally accepted definition for 'green'. Here, 'green' will refer to anything that is accepted to be sustainable or environmentally friendly, including products, practices, technologies, etc. Table 1 below gives an overview of the key terms referenced throughout this paper.

Table 1: Key Definitions in Green Policy

Term	Definition
Green	Anything that is commonly accepted to be sustainable or environmentally friendly
Sustainability	The ability for a community to maintain environmental, health, and economic stability over an extended period. It should be noted that 'green' and sustainable are sometimes used interchangeably in media
Renewable Energy	Energy sources that are derived from sources that are naturally replenishing. These include solar, wind, geothermal, hydropower, and biomass (ex. wood, landfill/biogas, ethanol, biodiesel)
Non-Renewable Energy	Energy sources that once they are depleted, they cannot be replaced. These include petroleum, coal, nuclear, diesel, etc.
Clean Energy	Non-pollutant producing energy sources, including solar, wind, hydropower, geothermal, bioenergy, nuclear, and hydrogen/fuel cells. It should be noted that clean energy sources are not always renewable (ex. nuclear)

2.2. Best Practices

'Green' best practices are a set of actions that can be implemented to improve environmental impact, such as reducing greenhouse gas emissions, minimizing waste, saving water, etc. We observed six main categories of these practices, including renewable energy, sustainable buildings, fleet management, waste reduction, water conservation, and sustainable purchasing. Table 2 below provides a few examples of each category.

Table 2: Overview of Green Best Practices

Best Practice	Example
Renewable Energy	Increasing use of solar, wind, hydropower, and geothermal energy, purchasing of renewable energy credits (RECs)
Sustainable Buildings	Requirements for new constructions to incorporate sustainable design plans, 'green' building certification
Fleet Management	Develop a vehicle fleet with more fuel efficient vehicles, or vehicles that use biofuels/renewable energy
Waste Reduction	Reducing demolition and construction waste, composting, paper reduction, increased recycling
Water Conservation	Installing dual plumbing for incorporating non-potable water sources, reducing landscaping costs, monitoring consumption to better understand use
Sustainable Purchasing	Committing to purchasing recycled, biobased, and other sustainable products

3. Benefits of 'Green' Practices for Government Agencies

In addition to environmental impact, there are many other benefits of implementing 'green' best practices. Some of the key benefits include: cost savings, improved safety and security, and public health. The section below will provide an overview of each of these benefits and how they can impact government agencies.

3.1. Cost Savings

One of the biggest benefits of implementing 'green' practices is the cost-savings, particularly on long term investments. Renewable energy is a great example of substantial cost-savings over time. Many renewable sources, such as solar, wind, or hydropower, do not require fuel costs. Traditional fossil fuel sources, like coal and oil, require not only high material costs, but often involve costly transportation fees as well. The cost of solar photovoltaics (PV) systems has fallen dramatically over the past decade [3] and currently wind is "either competitive with, or less expensive than, coal-generated electricity – and it is a form of Clean Energy. Ongoing cost reduction will soon make wind energy the least expensive source of electricity, perhaps within a decade" [4].

In addition to decreasing energy costs, there are other economic impacts from the implementation of sustainable practices. For example, studies have demonstrated that renewable energy policies within the Rust Belt have the potential to reduce air pollution to such a degree that the savings from improved human health would exceed the costs of the policies by 2030, with more stringent implementation having an even more pronounced cost-savings effect [5]. The decreased energy costs as well as health-related savings makes implementation of 'green' best practices a favorable financial decision for the U.S. government.

3.2. Safety

From a military perspective, safety is a major benefit of 'green' practices. There are direct relationships between increased safety to soldiers through the implementation of these best practices. One example is the use of more energy efficient vehicles. In active war zones, refueling missions are high risk. In Afghanistan, for example, oil refueling operations document one casualty for every 24 field-resupply convoys [6]. More fuel-efficient vehicles not only decrease fossil fuel use and carbon emissions, but can reduce the number of refueling missions required, saving soldiers' lives.

Extending beyond the military, implementing 'green' practices can increase safety for the entire nation. The National Climate Assessment (NCA) report states that "Climate change creates new risks and exacerbates existing vulnerabilities in communities across the United States, presenting growing challenges to human health and safety, quality of life, and the rate of economic growth" and notes that future impacts of climate change are "expected to further disrupt many areas of life, exacerbating existing challenges to prosperity posed by aging and deteriorating infrastructure, stressed ecosystems, and economic inequality" [1]. The report also emphasizes that vulnerable communities, lower-income, historically marginalized, etc., may experience greater impacts [1]. Wide-scale implementation of 'green' practices protects both military and civilian lives.

3.3. Security

'Green' practices are a unique opportunity to fortify national security. Researchers from the National Renewable Energy Lab (NREL) note that "renewable energy can support energy security by adding diversity to an overall electricity generation portfolio…

A more spatially diverse generation and storage energy portfolio can better withstand shocks to the system. With more resources across different geographic areas, such diversity could power infrastructure during disasters, cyberattacks, or other extreme events" [7]. With wider use of these energy sources, there is a decreased dependency on fossil fuels. Disruptions in the supply chain could lead to disastrous effects for both the nation and our military. Transitioning to renewable-energy sources not only addresses supply chain risk but can also increase the physical security of the energy generation process by decreasing reliance on centralized production and subsequent distribution. Many renewable sources can be implemented on-site, such as at military instillations, strengthening the resiliency of the that location's energy supply. The electrical grid is particularly susceptible to attacks, both cyber as well as physical, which could disrupt operation at critical facilities [6]. Localized power on-site using renewable sources decreases these risks and thus greatly improves security.

In addition, climate change is a critical driver of military investment in 'green' practices. Rising sea levels, for example, are a particular concern for the Navy. There is an expected increase in the demand for the Navy's military and humanitarian services in response to the effects of climate change. Both domestic and global bases and ports will be at high risk as sea levels rise and weather patterns become more severe [6]. Slowing the rate of climate change not only gives the Navy and other military branches more time to prepare for these challenges but can also prevent some of these concerns from even becoming a reality.

3.4. Public Health

Implementing these practices will also impact public health through the reduction in greenhouse gas emissions. The National Climate Assessment report notes that the "health and well-being of Americans are already affected by climate change, with the adverse health consequences projected to worsen with additional climate change" due to the effects of "exposures to heat waves, floods, droughts, and other extreme events; vector-, food- and waterborne infectious diseases; changes in the quality and safety of air, food, and water; and stresses to mental health and well-being" [1].

Implementing 'green' practices and enacting sustainability-focused policies has the potential to reduce the risks and impacts from climate-sensitive health outcomes, and researchers project that "additional benefits to health arise from explicitly accounting for climate change risks in infrastructure planning and urban design" [1]. The report claims that "reducing greenhouse gas emissions would benefit the health of Americans in the near and long term" and that "thousands of American lives could be saved and hundreds of billions of dollars in health-related economic benefits gained each year under a pathway of lower greenhouse gas emissions" [1].

4. Methodology

As described above, the benefits of 'green' practices are undeniable, particularly from a government perspective. The goal of this work was to provide an overview of the different definitions of 'green', highlight the emerging trends, and illustrate gaps between government and industry practices. In this process, we illustrate the growing need to set goals, assign priorities, establish regulations, and make investments as well as guidance for where to focus these efforts. This is intended to provide foundational information for implementing green policy, as well as identified resources for where to look for more information when setting or writing future sustainability policies.

The team conducted a survey of 'green' practices and policies across both government and industry. To assess the current state of the government sustainability efforts, the authors reviewed government agency sustainability plans and published 'green' goals, 'green' focused legislation and executive orders (EOs), as well as the timeline of when and how these documents were put into effect. For industry, key players were identified based on a combination of publicly released policies, data, and status on third-party 'green' rankings. The publicly released 'green' initiatives and policies of these companies were then compiled as a list of best practices as well as analyzed for industry-wide trends. The results of these evaluations were then used to identify gaps within current government 'green' efforts. Based on these findings, recommendations were made for what steps the government should take moving forward, as well as what research still needs to be done to properly implement these actions.

4.1. Sustainability Evaluation

While there are plenty of organizations with sustainability practices already in place, it can be challenging to compare different companies to one another. To help with this, several third-party organizations have developed sustainability certifications and rankings to help consumers better understand leaders in industry. These accolades have considerable impact on a company, affecting everything from brand image to product success. Having a high sustainability ranking or 'green' certified products and facilities can attract talent and generate new business. Pew Research studies have shown that adults under 50 believe protecting environment and increasing reliance on renewable energy sources should be a high priority for America's energy policies [8] and adults under 30 place a high priority on protecting the environment [9]. Additionally, both Glassdoor and Forrester Research observe that both job seekers and consumers are gravitating towards companies whose mission and culture align with their values [10] [11]. In addition, as today's consumers become more concerned about the environmental impact of their purchasing decisions, these certifications and rankings can influence purchasing decisions.

4.2. Sustainability Certifications

Sustainability certifications are third-party evaluations that assess the sustainability of a building, business, product, etc. These programs often provide both an assessment of the current sustainability practices or designs in place as well as what steps need to be taken to make further improvements. Additionally, many certifications have varying levels (ex. gold, silver, etc.) that indicate what that particular service has achieved. Table 3 gives an overview of available programs in each of the main certification categories.

Table 3: Sustainability Certification Overview

Category	Certification	Summary
Buildings	LEED	Leadership in Energy and Environmental Design (LEED), certification for buildings based on sustainable design features [12]
	WELL	Building Standard to create buildings and originations with thoughtful and intentional spaces that enhance human health and well-being [13]
	EDGE	Building certification focused on making buildings more resource-efficient [14]
	ParkSmart	Certification program for developing sustainable parking garages [15]
	Energy Star	Certification for products and buildings following standards set by the EPA; Energy Star buildings must save energy, money, and help protect the environment through limited generation of greenhouse gas emissions [16]
	SITES	Sustainability-focused framework for landscape design (reduce water demand, reduce energy consumption, improve air quality, etc.) [17]
	Living Building Challenge Certification	Certification program with a goal to create buildings that generate more energy than they use, capture and treat all water on site, and are built using healthy materials. Several different certifications available: Core Green Building Certification, Zero Energy Certification, and Zero Carbon Certification [18]
	Green Building Initiative	Certification program (Green Globe Certification) designed to allow building owners to select which sustainability features best fit their building and its occupants, creating a custom-tailor approach that helps from the most sustainable outcomes [19]
	BREEAM	Building Research Establishment Environmental Assessment Method (BREEAM); method of assessing, rating, and certifying the sustainability of buildings [20]
Businesses	Green Business Bureau Seal	Online green business certification program using the Green Business Bureau (GBB)'s Eco Assessment tool [21]
	Green America Certified Business	Certification that evaluates companies according to principles of social justice and environmental sustainability [22]
Products/ Materials	Energy Star	Certification for products as well as buildings following standards set by the EPA; products must reach a defined set of criteria, applicable for computers, servers, appliances, heating and cooling systems, electronics, lighting, etc. [16]
	WaterSense	Voluntary partnership program sponsored by the EPA, products and services are certified to use at least 20 percent less water, save energy, and perform as well as or better than regular models [23]
	EPEAT	Certification for electronics; leading ecolabel covering products and services from the electronics sector [24]
	Safer Choice	Certification through the EPA Pollution Prevention which identifies products that are both high performing and contain ingredients that are safer for human health and the environment [25]
	USDA Organic	U.S. certification for foods to be labeled organic according to federal guidelines on soil quality, animal raising practices, pest control, and use of additives [26]
	Fair Trade Certified	Certification evaluating a company's use of equitable trade practices across their supply chain, ensuring fair treatment, prices, and environmental impact [27]
	Green Seal	Green Seal Verification confirms that a raw material, concentrate, formula, or parent roll is proven to meet human and environmental health standards [28]
Other	PEER	Certification for the power and energy sector [29]
	GRESB	Global Real Estate Sustainability Benchmark Certification, assesses the sustainability and ethical impact of real estate and infrastructure [30]
	WasteWise	Program through the EPA to encourage companies to reduce waste, practice environmental stewardship, and incorporate sustainable materials management into their waste-handling processes [31]

4.3. Sustainability Rankings

In addition to various certifications, ranking systems have also come about for the top sustainable companies. The leading classifier is the Global 100, an annual ranking of corporate sustainability performance performed by Corporate Knights (CK). This list is released each January at the World Economic Forum in Davos and is considered the gold standard for sustainability ranking evaluations. For companies to be eligible for the Global 100, they must meet several eligibility requirements. Only publicly listed companies with a gross revenue of a minimum of one billion dollars can be considered. From this initial list, all industries and locations are eligible before screening [32].

CK uses only publicly disclosed data, such as financial filings and sustainability reports, to evaluate 21 different key performance indicators (KPIs) which cover resource management, employee management, clean revenue, and supplier performance. Out of this list, all companies are evaluated on the eight universal KPIs: percentage tax paid, pension fund status, supplier sustainability, women in executive management, women on boards, sustainability pay link score, sanctions deductions, and clean revenue. Other KPIs are considered based on a company's corresponding CK Industry Group. It should be noted that there are automatic disqualifiers such as companies that produce weapons, tobacco, or lobby to block climate change policy [32].

5. Sustainability Policies and Legislation within the United States Government

5.1. Federal Government

Before the early 2020s, the U.S. federal government made minimal steps toward implementing sustainability policy. There was over a decade lull where, for example, no major law regarding sustainability was passed between the mid-2000s through the early 2020s. The Energy Policy Act of 2005 [33] introduced tax incentives and loan guarantees for various energy sources, while the Energy Independence and Security Act of 2007 [34] aimed to move the U.S. toward greater energy independence and security through the increase of clean renewable fuels.

During this lull, only a handful of Executive Orders (EO) and government guidelines were released for implementing sustainable practices. In 2015, a comprehensive document, Planning for Federal Sustainability in the Next Decade [35], was put into effect, outlining the federal government's 10-year plan for implementing more sustainable practices. Following this, several other guidelines were released, including Executive Order 13834, Efficient Federal Operations [36] in 2018 and the Guiding Principles for Sustainable Federal Buildings and Associated Instructions [37] in 2020. EO 13834 "affirms that it is the policy of the United States that agencies meet energy and environmental performance statutory requirements in a manner that increases efficiency, optimizes performance, eliminates unnecessary use of resources, and protects the environment" and that "agencies are tasked to prioritize actions that reduce waste, cut costs, enhance the resilience of federal infrastructure and operations, and enable more effective accomplishment of its mission. Additionally, EO 13834 has been revoked by Executive Order 13990, Protecting Public Health and the Environment and Restoring Science To Tackle the Climate Crisis, which "directs all executive departments and agencies to immediately review and, as appropriate and consistent with applicable law, take action to address the promulgation of federal regulations and other actions during the last 4 years that conflict with these important national objectives, and to immediately commence work to confront the climate crisis" [38].

A major shift in 'green' policy has been seen in the early 2020s, with multiple examples of legislation focused on climate change and sustainability going into effect in just a few short years. This change will be discussed later in this section, but it should be noted that a need still exists to further develop legislation, policy, and regulations to make substantial impact on a 'greener' nation.

5.2. Military Agencies

Many of the benefits of 'green' best practices are valuable to the U.S. military. As discussed above, these practices can help to improve safety for soldiers in the field and increase security of military bases, all while decreasing operational costs. For these reasons, the Department of Defense (DOD) has released yearly Sustainability Reports and Implementation Plans (SRIP), and branches of the military have already identified sustainability as an organization-wide goal. For example, the Navy, Army, Air Force, and National Guard, all have sustainability policies already in place, though these policies largely only reference goals and plans outlined in EOs and the DOD SRIP.

The DOD Sustainability plan released in 2020 identifies four major priorities: energy resilience, water efficiency, sustainable acquisition, and electronics stewardship [39]. When detailing the progress and goals for energy resilience, the report emphasizes goals to reduce energy consumption, while noting that there are goals in place for the DOD "to produce or procure greater than or equal to 25% of the total quantity of facility energy DOD consumes within its facilities during FY 2025 and each fiscal year thereafter from renewable energy sources," but that in 2019, only "6.0% of facility electricity consumption was procured from renewable energy sources" [39]. Although the Department does not have specific water efficiency goals, DOD is successfully using innovative approaches to conserve water, save costs, and assure access to an adequate water supply for mission success [39].

DOD policies and guidance encourage the construction of high performance and sustainable buildings. Policy requires "new construction and major renovations to adhere to the Guiding Principles for Federal Sustainable Buildings" and instructs DOD buildings to obtain "at least the LEED silver level (when cost effective)" [39]. DOD policies also outline quantifiable goals for waste reduction, such as "2% reduction in non-hazardous solid waste generated in FY21 from FY20 40% diverted and 60% sent to treatment and disposal facilities in FY21" [39]. While the report contains quantifiable goals for reducing waste sent to landfills as well as reducing energy consumption and, the report mentions the goal for "continuous improvement" for other 'green' practice adoption goals [39].

In addition to implementing sustainability plans, many military branches have achieved impressive feats in implementing best practices. For example, the Weed Army Community Hospital in Irwin, California has been awarded the LEED Platinum standard, the highest level given. The facility is both carbon neutral and has net zero energy output, meaning that all electricity is generated

onsite from renewable energy sources [40]. While the military is beginning to take steps to make their facilities, fleets, and operations more sustainable, more resources, education, and support is required to make substantial lasting change across the entire organization.

5.3. Environmental Agencies

The leading U.S. government organization focused on the environment is the Environmental Protection Agency (EPA), whose mission is to protect both human health and the environment. In 2018, the EPA published Working Together: FY 2018-2022 U.S. EPA Strategic Plan [41], which gave three goals for accomplishing the above mission. These are outlined in Table 4.

Table 4: EPA FY2018-2022 Strategic Plan Goals Overview [41]

Number	Overview	Description
Goal 1	A cleaner, healthier environment	Deliver a cleaner, safer, and healthier environment for all Americans and future generations by carrying out the Agency's core mission
Goal 2	More effective partnerships	Provide certainty to states, localities, tribal nations, and the regulated community in carrying out shared responsibilities and communicating results to all Americans
Goal 3	Greater certainty, compliance, and effectiveness	Increase certainty, compliance, and effectiveness by applying the rule of law to achieve more efficient and effective agency operations, service delivery, and regulatory relief

In addition to the EPA, the Department of Energy (DOE) does considerable work related to sustainability in the energy sector. The mission of the DOE is to "ensure America's security and prosperity by addressing its energy, environmental, and nuclear challenges through transformative science and technology solutions" [42]. The DOE helps to fund research and development of clean energy technologies, collaborating across industry, academia, and government to produce new and innovative approaches to the energy crisis. Recently, the DOE announced two renewable energy investments, one with the goal to "cut the cost of solar energy by 60% within the next ten years, in addition to nearly $128 million in funding to lower costs, improve performance, and speed the deployment of solar energy technologies" and one with the goal to deploy 30 gigawatts (GW) of offshore wind by 2030 [43]. In addition, the DOE has 17 labs where they operate research topics like climate change. Within the DOE, the Office of Energy Efficiency and Renewable Energy (EERE) focuses specifically on renewable and clean energy.

5.4. Increased Government Sustainability

As stated above, there was minimal investment in 'green' before the early 2020s. No major laws and only few sustainability-focused EOs were implemented over a 15-year period. Beyond just legislation, several climate initiatives were canceled in the late 2010s, such as the Navy's Climate Task Force, and the U.S. officially withdrew from the Paris Climate Agreement in 2020 [44].

Beginning in 2021, however, drastic changes were made across the government that showcased their commitment to creating a more sustainable nation. Not only did the U.S. re-enter the Paris Climate Agreement [45], but many climate change focused EOs were put into place: Executive Order 13990: Protecting Public Health and the Environment and Restoring Science to Tackle the Climate Crisis [38], Executive Order 14008: Tackling the Climate Crisis at Home and Abroad [46], and Executive Order 14013: Rebuilding and Enhancing Programs to Resettle Refugees and Planning for the Impact of Climate Change on Migration [47]. These orders not only established various environmental policies but made a strong statement about the government's stance on the environment moving forward. For example, EO 13990 revoked the permit for the controversial Keystone XL Pipeline [38].

These initiatives grew beyond just EOs and several pieces of legislation focusing on climate change and sustainability have been put into effect. In 2021, the Infrastructure Investment and Jobs Act [48] addressed climate change concerns, specifically on its impact to the transportation system. Furthermore, the Inflation Reduction Act of 2022 [49] marked billions of dollars for climate change and energy research and development, making it the largest U.S. government investment in climate change to date. This trend was seen throughout government activities during 2022. For example, the National Strategy for Advanced Manufacturing [50] stressed the need to invest in sustainable manufacturing to further address the climate crisis.

As a nation, we need to continue this trend and continue to develop further legislation, policy, and regulations to make substantial impact on a 'greener' nation. With this increased investment in sustainability, now is the time to harmonize government 'green' policy to make lasting, impactful changes. By developing sustainability plans and practices like those used by industry, the U.S. can become a leader in the international 'green' space. By actively working on climate change and global warming issues at the national level, the government can influence other nations and make world-wide impact.

6. Industry Sustainability Trends

The observed key players in industry are considered 'green' leaders due to the publicity surrounding their environmental initiatives and publicly available sustainability policies. These yearly sustainability reports contain the company's short and long-term goals, current metrics and measurements, as well as documented progress. This showcases both their sustainability efforts and effort to cultivate the culture of sustainability, in addition to asserting the company's influence in the 'green' space.

6.1. Renewable Energy

All industry leaders investigated in this study place strong emphasis on renewable energy adoption and carbon emission reduction. Google, for example, has achieved carbon neutral-status for the last 12 years and has purchased enough renewable energy to match 100% of their global electricity consumption since 2017 [51]. Similarly, starting in 2018, Apple announced its global facilities (including retail stores, offices, data centers and co-located facilities in 43 countries) are powered with 100%

renewable energy. These renewable energy projects include Apple Park, Apple's headquarters in Cupertino, which is powered by 100% renewable energy. Power comes from multiple sources, including a 17-megawatt onsite rooftop solar installation and four megawatts of biogas fuel cells, all controlled by a microgrid with battery storage [52]. Currently, Starbucks purchases enough renewable energy to power 100% of its company-operated stores in the U.S., Canada, and the U.K. as well as making strategic investments in solar and wind farms [53].

6.2. Circular Economy

The observed industry leaders' sustainability policies also place a strong emphasis on reducing waste through a shift towards a "circular economy", one that benefits businesses, society, and the environment [54]. Google, for example, states that their goal is to "design out waste and pollution", with the aim being to eliminate release of greenhouse gases, the use of toxic and hazardous substances, the pollution of air, land, and water, and landfilling and incineration of waste at the design phase. Their designs aim to maximize product use and reuse: designing for durability, repair, reuse, remanufacturing, and ultimately recycling. Through these more durable and recyclable products, Google seeks to "create demand for recycled materials in order to accelerate the transition to a circular economy" [55]. Similarly, Apple has expanded their refurbished devices program over recent years in their shift towards a "circular economy" mentality. More than 11 million devices were sent by Apple to be refurbished for new users in 2019, a 42% increase from the previous year [53]. While Starbucks does not focus on explicitly implementing the 'circular economy' concept within their business practices, they do set goals to "reduce waste sent to landfills from stores and manufacturing by 50%, driven by a broader shift toward a circular economy" [56].

6.3. Sustainable Buildings

These companies are not only redesigning their products and packaging, but also the buildings that house their operations. By the end of 2019, over 1.4 million square meters (13 million square feet) of Google office facilities had achieved LEED certification [51]. Moving beyond LEED certification, Google is also pursuing the Living Building Challenge (LBC) Materials Petal certification for its facilities, which would certify that every building product on-site has been vetted against the LBC's Red List of worst-in-class chemicals that pose human and environmental health concerns. Similarly, "more than 50 Apple sites have received LEED or BREEAM (Building Research Establishment Environmental Assessment Method) certifications" [53]. Apple facilities feature many 'green' practices such as natural ventilation systems to highly efficient hydronic radiant heating and cooling, smart LED lighting, and high-efficiency water-cooled chillers, low-flow fixtures, and an irrigation system that uses 90% recycled water help to optimize water consumption [53]. Starbucks has similarly built more than 1,600 LEED-certified stores, but intends to go "beyond LEED, expanding the scope and breadth of [their] greener stores commitment with an open-source Greener Stores framework for design, construction and operation" [57].

6.4. Fostering a Sustainable Culture

Industry leaders are changing not only the company's buildings, but also the culture within their organization. For example, Google recognized that "reducing single-use beverages relied on behavior science insights in order to raise the desirability of sustainable and healthy options", and that a broader "culture change also matters" [55]. Similarly, Starbucks provides 'green' education materials and has "14,800 Greener Apron partners" with their goal to "empower 10,000 partners to be sustainability champions by the end of 2020" [57].

6.5. External Sustainable Investments

Companies are also making strategic investments outside of their organization. For example, Google committed to "invest roughly $150M into renewable energy projects in key manufacturing regions" to further reduce carbon emissions. These investments are not limited to just renewables - the company also "requires the highest ethical standards throughout [their] supply chain" and is "working to promote meaningful improvements in the communities in which [they] operate" [58]. Similarly, Apple has made strategic investments in renewable energy technologies. Apple not only runs its facilities on 'green' power (including solar roofs) but has "also convinced 23 companies in its supply chain to sign a pledge to get to 100% renewable energy for the portion of their business relating to Apple products" [59].

Additionally, industry leaders are creating and contributing to funds that will invest in the restoration and protection of forests and natural ecosystems globally. Apple has made several steps in this effort, including creating "a fund in partnership with Conservation International" with the goal to 'help protect and restore the world's forests, wetlands, and grasslands to remove excess carbon from our atmosphere" [53] in addition to working with the Alliance for Water Stewardship, "investing in their work to raise awareness throughout the Asia-Pacific region on this issue, to build tools and training for the AWS certification process, and to increase support for suppliers working to be certified" while encouraging suppliers to pursue AWS certifications [53].

Starbucks has stated a commitment to ethically sourced products [60] marked by their 'Coffee and Farmer Equity (C.A.F.E.) Practices, one of the coffee industry's first set of sustainability standards, verified by third-party experts', which was developed in collaboration with Conservation International (CI), and 'includes over 400,000 coffee farmers in 28 countries that are committed to improving working conditions and spanning more than a million hectares of land committed to sustainable growing practices', recognizing that the 'longevity of the coffee industry is directly linked to the social, economic and environmental conditions of coffee communities' [61]. Starbucks has also invested in reforestation efforts with '40 million trees distributed since 2015' and the goal to provide '100 million coffee trees to farmers by 2025' [57], and links social responsibility to 'green' practices, and notes in their sustainability report that the company sells "99% ethically sourced tea" and have contributed "$46 million invested in farmer loans" [57].

7. Emerging Trends

Across industry, companies are investing in renewable energy to reduce their carbon emissions and transition away from fossil fuels. At a more granular level, companies are implementing 'green' practices in their buildings using new technologies, such as water conservation techniques, natural ventilation systems, and

smart LED lighting. Companies are pursuing LEED certification for their buildings, as well as new, more rigorous certifications. We also observe a focused shift towards a "circular economy" that reuses materials where possible. Each company investigated had a strong focus on employee engagement and fostering a culture that emphasizes sustainable practices. Many of these companies are also exerting their influence on suppliers and making strategic investments (both domestic and abroad) to shape 'green' practice adoption, such as reforestation efforts and renewable energy microgrids.

It should be noted that 'green' work goes beyond industry efforts. Research is being conducted across academia to develop new technologies and practices, as well as identify areas where further 'green' work needs to be done. One example of this is data centers, which have been identified as a major contributor to carbon emissions. These centers contribute exorbitantly to power consumption and their continued growth will be unsustainable such that the energy needed to power these capabilities might be infeasible, contribute greatly to greenhouse gasses, or both. Research is currently being done to assess the state of the art in sustainable computing, as well as identifying and characterizing challenges and potential solutions [62] [63].

8. Conclusions and Research Findings

This work provides an overview of the many different definitions of 'green', in addition to highlighting emerging trends and illustrating the growing gaps between government and industry practices. These findings demonstrate the need to set goals, assign priorities, promulgate regulations, and make investments, as well as guidance for where to focus these efforts. This work is intended to provide foundational information for implementing green policy, as well as identify resources for where to look for more information when creating sustainability policies.

One of the first steps in enacting 'green' practices across the government is to formalize a single definition for the term 'green'. The EOs enacted in 2021 emphasize a need for a "whole-of-government approach to combatting the climate crisis" (which would require coordinated and deliberate approach) as well as a need to "leverage the federal government's footprint and buying power to lead by example" (which would require a consistent example and well understood priorities to follow) [46]. With a formalized definition and articulated 'green' priorities, the combined "whole of government" approach with the federal government's buying power provides unique opportunities to evaluate the national benefits and costs of pursuing more nation-wide transformational 'green' initiatives. Examples of these actions could include incentivizing a large-scale transition to all-electric vehicles and a large-scale shift to a decentralized power grid by using renewable energy based DERs and microgrids.

The phenomenon of "greenwashing," where corporations mislead consumers about their environmental performance or the environmental benefits of a product or service [64], has become increasingly widespread [65] and has eroded consumers and citizens' trust [66]. These incontinences around the term 'green' have created a misunderstanding of the true benefits of implementing more sustainable practices in the government and across the U.S. We recommend the federal government enact a top-down 'green' definition as well as 'green' priorities that will standardize how environmental suitability and planning is approached.

To capture the findings from our survey of 'green' best practices, we provide answers to our original research questions in the following sections.

8.1. What quantifiable metrics or features make a company 'green'?

This work observed that companies considered 'green' have well-documented, quantifiable improvements in their sustainability initiatives. These are published in yearly, publicly available progress reports that often include multi-year goals as well as progress over time. The goals could include reduction in carbon emissions, amount of energy obtained through renewable sources as opposed to traditional fossil fuel sources, amount of waste diverted from landfills compared to previous years and obtaining third-party certifications for sustainable buildings. We also observed quantifiable goals and developments in the areas of fleet management, water conservation techniques, and increasing 'green' requirements for suppliers.

8.2. What are 'green' best practices?

The U.S. government identifies 'green' best practices as practices that improve energy resilience, water efficiency, sustainable acquisition and lifecycle management, and electronics stewardship [39]. As mentioned, the DOD has goals to source greater than 25% of facility energy from renewable energy sources, as well as goals to incorporate innovative approaches to conserve water. DOD policies also prioritize high performance and sustainable buildings that target at least the LEED silver level. DOD policies also outline quantifiable 'green' goals for waste reduction, including quantifiable reduction in waste generated, as well as quantifiable goals for increasing the percentage waste diverted from landfills.

Industry leaders identify 'green' best practices as those that target carbon emissions, focusing on investments in both renewable energy and developing a 'circular economy' business approach. Companies are also implementing more granular 'green' practices such as water conservation techniques (including dual-plumbing systems to leverage non-potable water source where possible), natural air circulation ventilation systems, and smart LED lighting. Additionally, industry leaders emphasize employee engagement and culture shifts to promote a company culture that implements 'green' practices. Many of these companies have also invested in ways to shape 'green' practice adoption outside of their direct organization, such as reforestation efforts and renewable energy microgrids.

As discussed in Section 4.2, third-party certifications help identify 'green' best practices in the areas of sustainable buildings, business values, product manufacturing, and water use. Beyond this, third-party rankings use key performance indicators such as resource management, employee management, clean revenue, and supplier performance to grade companies. Using the well-regarded Corporate Knights method, companies are evaluated on the eight universal KPIs: percentage tax paid, pension fund status, supplier sustainability, women in executive management, women on boards, sustainability pay link score, sanctions deductions, and clean revenue.

8.3. Who are the industry leaders in the 'green' field?

Through our survey, Google, Apple, and Starbucks were identified as viable case studies for 'green' leaders in industry. It should be noted that Disney, IBM, and Dell also appear on sustainability rankings lists but were not used as the primary focus in this investigation [32]. The three chosen companies are not only acknowledged as leading 'green' institutions, but also offer publicly available sustainability reports that could both be reviewed for the purpose of this investigation but can be used as examples for other organizations developing their own plans.

9. Recommendations

To be competitive with industry and international partners, we recommend that all U.S. government agencies formalize and publicly release sustainability policies with (1) quantifiable goals aligned with the major areas industry leaders are targeting, (2) identified 'green' practices the agency plans to implement to meet these goals, and (3) metrics the agency plans to use to measure sustainable progress. As a first step to develop these goals, we recommend each agency evaluate their current organization as a baseline for improvements. The major sustainability areas discussed in this paper (renewable energy use, sustainable building, fleet management, waste reduction, water conservation, and sustainable purchasing) should be included in this investigation. Existing evaluation methods, such as the Corporate Knights ranking procedure, should be used as a guideline during the process. Once the baseline is determined, milestones and targets should be planned such that the government can meet industry standards for sustainable practices (carbon neutral, net-zero to landfill, buildings maintaining LEED certification, etc.). A new EO that standardizes agency green practices across the U.S. government would help catalyze individual agency efforts. Further U.S. government funded research on how 'green' best practices slow climate change and its adverse societal consequences, improve public health, and enhance environmental protection is also recommended.

10. Future Work

While this report outlines the changing landscape of 'green' practices, we recognize that there are a multitude of barriers inhibiting wider-spread adoption of these practices and an evaluation needs to be performed to determine what these barriers are and how to best address them. These barriers include, but are not limited to, the following: perceived cost, conflicting codes and guidelines, and lack of government resources. One of the most widely documented barriers to 'green' practice adoption is the upfront expense for implementation. Additionally, we observed that LEED certification conflicts with certain COVID guidelines, specifically for ventilation.

The next step for this research is to formally investigate these barriers to adoption to propose novel solutions to remove those barriers and enable implementation of 'green' practices and climate mitigation strategies. The authors completed an initial investigation to these barriers specially for solar photovoltaics [67] and are continuing the research.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This project was funded by MITRE's independent research and development program.

References

[1] Climate Science Special Report: Fourth National Climate Assessment, **I**, Washington, DC, 2017, doi:10.7930/J0J964J6.
[2] E. Holt, C. Corrado, "Emerging Trends in Green Best Practices," in 2021 IEEE International Symposium on Technologies for Homeland Security (HST), IEEE: 1–5, 2021, doi:10.1109/HST53381.2021.9619852.
[3] V. Ramasamy, D. Fledman, U.S. Solar Photovoltaic System and Energy Storage Cost Benchmarks: Q1 2021, 2021.
[4] P. Hawken, Drawdown: The Most Comprehensive Plan Ever Proposed to Reverse Global Warming, Penguin Books, New York, 2017.
[5] E.G. Dimanchev, S. Paltsev, M. Yuan, D. Rothenberg, C.W. Tessum, J.D. Marshall, N.E. Selin, "Health co-benefits of sub-national renewable energy policy in the US," Environmental Research Letters, **14**(8), 085012, 2019, doi:10.1088/1748-9326/ab31d9.
[6] F. Reinhardt, M. Toffel, "Managing Climate Change: Lessons from the U.S. Navy," Harvard Business Review, 102–111, 2017.
[7] S.L. Cox, L. Beshilas, E.L. Hotchkiss, Renewable Energy to Support Energy Security, Golden, CO (United States), 2019, doi:10.2172/1569691.
[8] C. Funk, B. Kennedy, Public Divides Over Environmental Regulation and Energy Policy, 2017.
[9] Pew Research Center, After Seismic Political Shift, Modest Changes in Public's Policy Agenda, 2017.
[10] A. Stansell, Which Workplace Factors Drive Employee Satisfaction Around the World, 2019.
[11] J. Nail, The Power of The Values-Based Consumer - And Of Authentic Brand Values, 2020.
[12] U.S. Green Building Council, LEED Rating System, 2022.
[13] International WELL Building Institute, WELL, 2022.
[14] GBCI, EDGE, 2022.
[15] GBCI, ParkSmart, 2022.
[16] Environmental Protection Agency, Energy Star, United States Department of Energy, 2022.
[17] GBCI, SITES, 2022.
[18] International Living Future Institute, Living Building Challenge Certification, 2022.
[19] Green Building Initiative, Green Building Initiative, 2022.
[20] BRE, BREEAM, 2022.
[21] Clearyst GBB LLC, Green Business Bureau Seal, 2022.
[22] Green America, Green America Certified Business, 2022.
[23] United States Environmental Protection Agency, WaterSense, United States Department of Energy, 2022.
[24] Global Electronics Council, EPEAT, 2022.
[25] United States Environmental Protection Agency, Safer Choice, United States Department of Energy, 2022.
[26] U.S. Department of Agriculture, USDA Organic, 2022.
[27] Fair Trade USA, Fair Trade Certified , 2022.
[28] Green Seal, Green Seal, 2022.
[29] GBCI, PEER, 2022.
[30] GRESB, GRESB, 2022.
[31] Waste & Compliance Management Inc, WasteWise, 2022.
[32] Corporate Knights, The 2020 Global 100: Overview of Corporate Knights Rating Methodology, 2021.
[33] H.R.6 - 109th Congress, The Energy Policy Act of 2005, 2005.
[34] H.R.6 - 110th Congress, Energy Independence and Security Act of 2007, 2007.
[35] D. Trump, Planning for Federal Sustainability In The Next Decade, 2015.
[36] D. Trump, Efficient Federal Operations, 2018.
[37] Environmental Protection Agency, Guiding Principles for Sustainable Federal Buildings and Associated Instructions, 2022.
[38] J. Biden, Executive Order on Protecting Public Health and the Environment and Restoring Science to Tackle the Climate Crisis, 2021.
[39] United States Department of Defense, 2020 Sustainability Report and Implementation Plan, 2020.
[40] J. Morgan, "Army hospital goes above call of duty in sustainable design," Health Facilities Management, 2017.
[41] United States Environmental Protection Agency, Working Together: FY 2018-2022 U.S. EPA Strategic Plan, 2018.

[42] United States Department of Energy, United States Department of Energy, 2022.
[43] United States Department of Energy, Energy Secretary Granholm Announces Ambitious New 30GW Offshore Wind Deployment Target by 2030, United States Department of Energy, 2021.
[44] J. Simkins, "Navy quietly ends climate change task force, reversing Obama initiative," Military Times, 2019.
[45] J. Biden, Paris Climate Agreement, 2021.
[46] J. Biden, Tackling the Climate Crisis at Home and Abroad, 2021.
[47] J. Biden, Rebuilding and Enhancing Programs to Resettle Refugees and Planning for the Impact of Climate Change on Migration , 2021.
[48] H.R.3684 - 117th Congress, Infrastructure Investment and Jobs Acts, 2021.
[49] H.R.5376 - 117th Congress, Inflation Reduction Act of 2022, 2022.
[50] Subcommittee on Advanced Manufacturing, National Strategy for Advanced Manufacturing , 2022.
[51] R. Porat, Google, Our 2020 Environmental Report, 2021.
[52] Apple, Apple now globally powered by 100 percent renewable energy, 2018.
[53] Apple, Environmental Progress Report: Fiscal Year 2019, 2020.
[54] Ellen MacArthur Foundation, The Circular Economy in Detail, Ellen Macarthur Foundation, 2017.
[55] Google, A Circular Google, Google, 2019.
[56] J. Warnick, 5 things to know about Starbucks new environmental sustainability commitment, Starbucks, 2020.
[57] Starbucks, Global Social Impact Report, 2019.
[58] Google, Our Approach to Sustainability , Google, 2019.
[59] M. Sullivan, "Apple Now Runs On 100% Green Energy, And Here's How It Got There," Fast Company, 2018.
[60] Starbucks, 2020 Report: Coffee, 2021.
[61] Starbucks, Making Coffee the World's First Sustainably Sourced Agricultural Product, Starbucks Stories & News, 2015.
[62] S. DeLong, A. Tolk, "Sustainable Computing and Simulation: A Literature Survey," in 2021 Winter Simulation Conference (WSC), IEEE: 1–12, 2021, doi:10.1109/WSC52266.2021.9715447.
[63] C.R. Corrado, S.M. DeLong, E.G. Holt, E.Y. Hua, A. Tolk, "Combining Green Metrics and Digital Twins for Sustainability Planning and Governance of Smart Buildings and Cities," Sustainability, **14**(20), 12988, 2022, doi:10.3390/su142012988.
[64] M.A. Delmas, V.C. Burbano, "The Drivers of Greenwashing," California Management Review, **54**(1), 64–87, 2011, doi:10.1525/cmr.2011.54.1.64.
[65] L.M. Baum, "It's Not Easy Being Green … Or Is It? A Content Analysis of Environmental Claims in Magazine Advertisements from the United States and United Kingdom," Environmental Communication, **6**(4), 423–440, 2012, doi:10.1080/17524032.2012.724022.
[66] M.D.T. de Jong, K.M. Harkink, S. Barth, "Making Green Stuff? Effects of Corporate Greenwashing on Consumers," Journal of Business and Technical Communication, **32**(1), 77–112, 2018, doi:10.1177/1050651917729863.
[67] C. Corrado, E. Holt, L. Schambach, "Barriers to Solar Photovoltaic (PV) Adoption on a National Scale in the United States," in 2022 IEEE 49th Photovoltaics Specialists Conference (PVSC), IEEE: 1110–1117, 2022, doi:10.1109/PVSC48317.2022.9938533.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 244-247 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Consideration of Ambiguity in the Analysis Phase of Data Warehouses

Djamila Hammouche*,1, Karim Atif [2]

[1]Department of computer science, Faculty of exact sciences and informatics, Hassiba Benbouali University, Chlef, 02180, Algeria

[2]Department of computer systems, Faculty of computer science, University of sciences and technology Houari Boumediene, Bab Ezzouar, Algiers, 16111, Algeria

ARTICLE INFO

Article history:
Received: 03 August, 2022
Accepted: 05 November, 2022
Online: 20 December, 2022

Keywords:
Ambiguity
Server failure
Data warehouse
Decision making
Fuzzy logic
Membership function

ABSTRACT

We are interested in taking into account ambiguity in the analysis phase of data warehouses, using fuzzy logic. We want to offer decision makers the possibility of using natural language in this phase. We created in a previous work the Baccalaureate fuzzy data warehouse which we were able to query with seven natural language terms to which we created seven membership functions. In this work, we present a fuzzy data warehouse for server failures that we created and for which we used the same terms to which we associated seven membership functions too. And, we carried out a comparison at the end of which we concluded that the definition of the values of the membership function differs according to the context of analysis. Our solution is extensible and can be enriched with new natural terms language. The next step is to design a conversational interface that enables a natural language conversation between the decision maker and the fuzzy data warehouse.

1. Introduction

Decision makers need to use natural language in the analysis phase of data warehouses; this allows them to appreciate existing data as they wish. Our field of application is higher education where decision makers need to analyze the failure rate of servers and easily detect the machine that frequently breaks down, the type of failure, the duration in order to be able to make good forecasts.

No machine is 100% reliable. The fault tolerance criterion of the machine is expressed either in average number of hours between failures, or in number of hours of operation before the end of life of the machine. Except this criterion given by the manufacturer, the users of these machines study server failures that occur on a real time basis. It is in this context that our study is written [1].

In this paper, we present the fuzzy server failure data warehouse that we created using Mondrian environment and MDX queries [1].

We are looking for the answer to the questions:

- Which servers have an average failure rate, lowest highest, etc.?
- How to use the terms: medium, high in queries?

The rest of the paper will be organized as follows: We first explain the context of this study, we present an overview of related works, and then we detail the solution based on fuzzy logic and we present the comparative analysis, finally, we finish with conclusion.

2. The study context

The goal of all our work is to show that the use of natural language facilitates the querying of data warehouses. So, we used Fuzzy logic. We defined the membership functions of the predicates that represent the natural language terms most used by analysts. In this context, we have chosen to experiment our solution on two very distinct areas:

- The field of national education to assess the baccalaureate success rate for which we have published the results [2]-[4].
- And the field of server failures to appreciate the server failure rate that we present in this paper.

In this work, we used the same terms to experiment them in the field of server failures and we defined the corresponding membership functions.

3. Related works

In [5]-[8], the author focuses on fuzzy multidimensional data, presenting a solution based on fuzzy logic and using SQL to take

* Corresponding Author: Djamila Hammouche, d.hammouche@univ-chlef.dz

https://dx.doi.org/10.25046/aj070627

into account imperfect values and vague criteria. ReqFlex is an intuitive user interface to the definition of preferences and the construction of fuzzy queries [9]. Also, the study on the medical data warehouse recording the vital parameters (blood pressure) of patients [10]. In [11], the author adds a new fuzzy layer to the existing model without modifying the data warehouse. Linguistic concepts are integrated for the interpretation of the measurements of the fact table. This separation between the fuzzy concepts and the data warehouse constitutes the strong point of this model. This is the reason why we have retained it in all our work despite the fact that this model has not been tested on a complex system.

4. The solution based on fuzzy logic

There are in classical Boolean logic only two states: TRUE or FALSE. In [12]-[14], the author proposed Fuzzy logic in 1965. This logic makes it possible to express different levels and to describe a phenomenon linguistically, then to represent it by a small number of rules.

A fuzzy term is a natural language word. To model it, we use a trapezoidal function for which characteristic parameters are defined.

4.1. The Fuzzy server failure data warehouse model

We designed the star schema of our data warehouse [15]- [21]. In [1], the model is defined and commented. There is one fact table "FactServerFailure" and four dimension tables: "Dim_Department", "Dim_Section", "Dim_Server" and "Dim_Time".

In fact table "FactServerFailure", there are all the primary keys of all the dimension tables related to the fact table and there is the measurement Failure_rate. It is as the metric to calculate and analyze. It indicates the ratio between the number of failure of a server to the total number of failures of the servers.

Failure_rate=number of server failure*100/total number of server failures.

In the other dimension tables, we find the characteristic attributes of these tables and which constitute axes of analysis such as time, department, etc.

The integration of ambiguous terms is done through two fuzzy meta tables that we call Fuzzy_ct table and Fuzzy_mt table.

Fuzzy Fuzzy_ct table is used to store fuzzy classes associated with linguistic terms (absolutely high, average, absolutely low, etc.) and Fuzzy_mt table is used to store the membership degrees of a value to a fuzzy class and the query result expresses the degree of each failure rate to a fuzzy class. Concrete examples of these degrees of membership are illustrated in the various tables which appear in the section some results.

4.2. The membership functions

We integrated seven fuzzy predicates to qualify the failure rate as absolutely high, rather high, somewhat high, average, somewhat low, rather low and absolutely low and we defined the seven corresponding membership functions.

Table 1: The membership functions

The membership function	Values and intervals
Absolutely high	Y=0 in [0, 7], y=1 in [8, 10] and y=x-7in [7, 8].
Ratherhigh	Y=0 in [0, 6] and [9, 10], y=1 in [7, 8], y=x-6 in [6, 7] and y=-9-x in [8,9].
Somewhathigh	Y=0 in [0, 5] and [9, 10], y=1 in [6,8], y=x-5 in [5, 6] and y=-9-x in [8, 9].
Average	Y=0 in [0, 4] and [7, 10], y=1 in [5,6], y=x-4 in [4, 5] and y=7-x in [6, 7].
Somewhatlow	Y=0 in [0, 6] and [9, 10], y=1 in [7, 8], y=x-6 in [6, 7] and y=-9-x in [8, 9].
Ratherlow	Y=0 in [0, 1] and [5, 10], y=1 in [2, 4], y=x-1 in [1, 2] and y=-x+5 in [4, 5].
Absolutelylow	Y=0 in [3, 10], y=1 in [0, 2] and y=3-x in [2, 3].

4.3. Some results

We present in following some results of the realized system.

In this table appear the servers with their characteristics of the University whose failure rates we have studied since their first commissioning until the year 2020.

Table 2: List of servers studied

Server number	Physical server model	RAM	Storage
1	HP Proliant gen 05	8G	512G
2	HP Proliant gen 05	8G	512G
3	HP Proliant gen 05	8G	512G
4	HP Sauvegarde	8G	512G
5	DELL PowerEdge T300	8G	512G
6	DELL	16G	1.7T
7	HP Z420 Workstation	8G	512G
8	DELL PowerEdge T320	16G	1.7T
9	DELL PowerEdge T320	16G	1.7T
10	DELL PowerEdge T320	16G	1.7T
11	HP Proxy1	12G	900G
12	DELL Proxy2	12G	900G
13	HP server Proliant Gen 10	8G	1T
14	DELL PowerEdge T320	16G	1.7T

In this table appear the servers with failure rates

Table 3: List of servers studied with their failure rates

Server number	Physical server model	Server failure rate
1	HP Proliant gen 05	9.09%
2	HP Proliant gen 05	9.09%
3	HP Proliant gen 05	9.09%
4	HP Sauvegarde	9.09%
5	DELL PowerEdge T300	6.06%

6	DELL	5.05%
7	HP Z420 Workstation	9.09%
8	DELL PowerEdge T320	1.01%
9	DELL PowerEdge T320	5.05%
10	DELL PowerEdge T320	6.06%
11	HP Proxy1	9.09%
12	DELL Proxy2	7.07%
13	HP server Proliant Gen 10	9.09%
14	DELL PowerEdge T320	8.08%

In this table appear the servers that recorded an absolutely high rate of failure. The percentage 9.09% and 8.08% are considered absolutely high failure rates with a total degree of membership (100%) and the percentage 7.07% is considered an absolutely high failure rate with only 07% membership degree. And all the percentages of the other remaining servers are not considered absolutely high failure rates because their membership degree is 0%.

Table 4: Result of query absolutely high

Server number	Server failure rate	Membership degree
1	9.09%	100%
2	9.09%	100%
3	9.09%	100%
4	9.09%	100%
7	9.09%	100%
11	9.09%	100%
13	9.09%	100%
14	8.08%	100%
12	7.07%	07%
05	6.06%	0%
06	5.05	0%
08	1.01%	0%
09	5.05%	0%
10	6.06%	0%

In this table appear the servers that recorded an absolutely low rate of failure. The percentage 1.01% is considered an absolutely low failure rate with a total degree of membership (100%) and all the percentages of the other remaining servers are not considered absolutely low failure rates because their membership degree is 0%.

Table 5: Result of query absolutely low

Server number	Server failure rate	Membership degree
8	1.01%	100%
1	9.09%	0%
2	9.09%	0%
3	9.09%	0%
4	9.09%	0%
5	6.06%	0%
6	5.05%	0%
7	9.09%	0%
9	5.05%	0%
10	6.06%	0%
11	9.09%	0%
12	7.07%	0%
13	9.09%	0%
14	8.08%	0%

4.4. The comparative study

We found that for the same terms the meaning remains the same but the membership functions change significantly. For example, for a very high baccalaureate success rate, the percentage of success is between 90% and 100%, while for server failures the highest rate is around 10% only.

Thus, we have established the following result. The definition of the values of the membership function is closely linked to the field of study. For any natural language term, the term retains its meaning all the time but the definition of the values of the membership function which corresponds to it changes according to the context of study.

We are interested in studying the same terms with other data from similar fields to better frame the definition of values of membership functions and to arrive at a single definition for a field of study. Only multiple experiments will prove it. Also, it is interesting to experiment with these same terms in sub-domains of the same domain like studying the failure rate of computers linked to servers.

Table 6: Results comparison

The membership function	Values in Server failure context	Values in Baccalaureate context[2]
Absolutely high	Y=0 in [0, 7], y=1 in [8, 10] and y=x-7in [7, 8].	Y=0 in [0.80], y=1 in [90,100] and y=0.1x-8 in [80, 90].
Ratherhigh	Y=0 in [0, 6] and [9, 10], y=1 in [7, 8], y=x-6 in [6, 7] and y=-9-x in [8,9].	Y=0 in [0.70], y=1 in [80, 90], y=0.1x-7 in [70, 80] and y=-0.1x+10 in [90,100].
Somewhathigh	Y=0 in [0, 5] and [9, 10], y=1 in [6, 8], y=x-5 in [5, 6] and y=-9-x in [8,9].	Y=0 in [0.50] and in [90,100], y=1 in [60, 80], y=0.1x-5 in [50, 60] and y=-0.1x+9 in [90,100].
Average	Y=0 in [0, 4] and [7, 10], y=1 in [5,6], y=x-4 in [4, 5] and y=7-x in [6, 7].	Y=0 in [0.40] and in [70,100], y=1 in [50,60], y=0.1x-4 in [40,50] and y=-0.1x+7 in [60,70].
Somewhatlow	Y=0 in [0, 6] and [9, 10], y=1 in [7, 8], y=x-6 in [6, 7] and y=-9-x in [8,9].	Y=0 in [0.30] and in [60,100], y=1 in [40,50], y=0.1x-3 in [30,40] and y=-

		0.1x+6 in [50,60].
Ratherlow	Y=0 in [0, 1] and [5, 10], y=1 in [2, 4], y=x-1 in [1, 2] and y=-x+5 in [4, 5].	Y=0 in [0.10], y=1 in [20,40], y=0.1x-1 in [10,20] and y=-0.1x+5 in [40,50].
Absolutelylow	y=1 in [0, 20].	Y=0 in [0.50] and in [90,100], y=1 in [60,80], y=0.1x-5 in [50,60] and y=-0.1x+9 in [90,100].

5. Conclusion

We presented the context of our study which concerns the two data warehouses, that of the Baccalaureate and that of Server failures. We were interested in the analysis phase of data warehouses where we were able to formulate MDX queries with natural terms using fuzzy logic and we explained some returned results. In this paper, we carried out a comparative analysis between the membership functions defined for each data warehouse and we found that even if the term of the natural language integrated into the model retains all its meaning, but the definition of the membership function which corresponds to it differs from one field of study to another, which has enabled us to affirm that the definition of the values of the membership function is strongly linked to the field of study.

Some perspectives emerge from our work, namely the insertion of new terms and the design of new query methods such as an expert system or a chatbot. Also the experimentation of the models in other establishments of the same domains to better test the defined membership functions.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] D. Hammouche, K.Atif, M.Loukam, "An An expert system for fuzzy server failure data warehouse analysis," ICIST'20, Lecee, Italy, 2020, doi:10.1504/IJDSS.2021.113767.

[2] D. Hammouche, K.Atif, "An extended solution to recommend fuzzy MDX queries for decision makers by a collaborative filtering profile," International Journal of Decision Support Systems **4**(3), 257-270, 2021, doi:10.1504/IJDSS.2021.113767

[3] D. Hammouche, M. Loukam, K.Atif, K.W. Hidouci, "Fuzzy MDX queries for taking into account the ambiguity in querying the Baccalaureate Data warehouse," Control, Decision and Information Technologies (CoDIT), 4th International Conference, Barcelone, 179-183, 2017, doi:10.1109/CoDIT.2017.8102587.

[4] D. Hammouche, L. Metchat, K.Atif, K.W. Hidouci, "A solution to recommend fuzzy MDX queries for decision makers by collaborative filtering profile in baccalaureate Data warehouse," In proceedings of the 2nd Conference on Computing Systems and Applications, Algiers,190-199, 2016.

[5] A. Laurent, S. Gangarski, C. Mar-Sala, "Cooperation between a fuzzy knowledge extraction system and a multidimensional database management system, " Francophone Meetings on Fuzzy Logic and its Applications, La Rochelle, France, Cepaduèce éditions, 325-332, 2000.

[6] A. Laurent, "Fuzzy multidimensional databases," Advanced database days, Agadir, Maroc, Hermès, 107-117, 2001, doi: 10.1007/978-3-642-10663-7-4.

[7] A. Laurent, FUB et FUB MINER: deux systèmes pour la représentation, la manipulation et la fouille de données multidimensionnelles floues, Information Interaction Intelligence, **3**, 37-83, 2003.

[8] C. Favre, A. Laurent,Y. Pitarch.,P. Poncelet, "Représentation graphique des hiérarchies contextuelles : modèle avec satellites," EDA 2011, Clermont Ferrand, Revue des Nouvelles Technologies de l'Information, Vol. B-7, 23-37, 2011, corpus ID:38627526.

[9] G. Smits, O. Pivert et T. Girault, "ReqFlex: Fuzzy Queries for Everyone,'' In proceedings of the VLDB Endowment, 1206-1209, 2013, doi: 10.14778/2536274.2536277.

[10] Y. Pitarch, A. Laurent, and P. Poncelet," A conceptual model for handling personalized hierarchies in multidimensional databases," In Proceedings of the International Conference on Management of Emergent Digital EcoSystems, France, 107–111, 2009, doi:10.1145/1643823.1643843.

[11] D. Fasel, K.Shahzad "A data warehouse model for integrating fuzzy concepts in meta table structures" 17th ECBS, Oxford, England, 100-109,2010, doi:10.1109/ECBS.2010.18.

[12] D.Dubois, H. Prade, "Using fuzzy sets in flexible querying: Why and how?," In Proc. Workshop on Flexible Query-Answering Systems, 89-103, 1996, corpus ID:33050805.

[13] D. Perez, M. J. Somodevilla, I. H. Pineda. "Fuzzy spatial data warehouse: A multidimensional model," In Eighth Mexican International Conference on Current Trends in Computer Science, 2007, doi: 10.5772/39389.

[14] D. Dubois and H. Prade, Fundamentals of fuzzy sets, **7**, 1-653, 2000, ISBN: 978-1-4615-4429-6.

[15] A.L. Zadeh , "Fuzzy sets," Journal of Information and control, **8**(3), 338-353,1965, doi:10.1016./S0019-9958(65)90241-X .

[16] M. Golfarelli, S. Rizzi., Data Warehouse Design: Modern Principles and Methodologie, Osborne/McGraw-Hill, 2009.

[17] T. Chikawa, M. Hirakawa, "ARES: a relational database with the capability of performing flexible interpretation of queries," In IEEE Transactions on Software Engineering, 624–634, 1986, doi:101109/TSE.1986.6312958.

[18] R. Agrawal., A. Gupta, "Modeling multidimensional databases," In Proceedings. 13th International Conference on Data Engineering, 232–243, 1997, doi:10.1109/ICDE.1997.581777.

[19] R. Bliujute, S. Saltenis, G. Slivinskas, C. Jensen, "Systematic Change Management in Dimensional Data Warehousing," In IIIrd International Baltic Workshop on Databases and Information Systems, Riga, Latvia, 27–41, 1998, corpus ID:10165485.

[20] R. Agrawal, A. Gupta, A. Sarawagi, "Modeling Multidimensional Databases," ICDE'97, 1997, doi:10.1109/ICDE.1997.581777.

[21] C. W. Holsapple, .K.D. Joshi, "Organizational knowledge resources," Decision Support Systems, **31**, 39-54, 2001, doi: 10.1016/S0167-9236(00)00118-4.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 248-255 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Interference-Aware Nodes Deployment of a LoRa-Based Architecture for Smart Agriculture in the Southern Region of Senegal

El Hadji Malick Ndoye[1,*], Ousmane Diallo[1], Nadir Hakem[2], Emmanuel Nicolas Cabral[1]

[1]*Department of Informatics, University of Assane Seck, B. P. 523 Ziguinchor, Senegal*

[2]*Université du Québec en Abitibi Témiscamingue Rouyn-Noranda, Quebec, Canada*

ARTICLE INFO

Article history:
Received: 01 September, 2022
Accepted: 30 November, 2022
Online: 20 December, 2022

Keywords:
Smart agriculture
LoRa
Gateway
Architecture
Casamance
Interferences

ABSTRACT

In Senegal, agriculture has always been seen as the foundation on which the socioeconomic development of the country rests. However, in the rural world, agriculture remains traditional at a time when the challenges of food self-sufficiency to accompany emergence are launched. In the southern part of Senegal commonly called Casamance, the abundance of rain makes it possible to practice rice cultivation and market gardening research must therefore play a leading role in the introduction of technological innovations, techniques, and decision-support tools to promote productive, competitive, and sustainable agriculture. Therefore, smart agriculture must focus on new solutions for water irrigation, soil quality, and culture monitoring. The emergence of the Internet of Things (IoT) is perceived as a very important lever for successful high-end intelligent agriculture. Indeed, the appearance of increasingly specialized monitoring sensors combined with new wireless communication technologies constitutes good decision-making tools.

The proposal of this paper consists of a new network architecture that can cover a large cultivation area to carry out water irrigation techniques in Casamance. It is, therefore, a question of identifying the best communication technology among new Low-Power, Wide Area Networks (LPWANs) such as Long-Range (LoRa), SigFox, etc which is suited to the environment considered. Also, the choice of the best deployment of sensors for better coverage. The choice of technology must be motivated by the financial costs and the range of transmission. The deployment must fix the optimal distance between the sensors minimizing the interferences according to some parameters specific to the environment. An analytical study is used on the deployment to determine the optimal distance between two gateway nodes to reduce induced interference.

1. Introduction

Today, devices tend to replace humans in their daily work. This induces a permanent interaction between connected devices and humans. Internet of Things (IoT) [1,2] can now be considered one of the most powerful tools for creating, modifying, and sharing countless amounts of information. Indeed, the IoT aims to make objects dialogue with each other and with individuals. It has become an essential means used in many fields such as health [3],[4], agriculture [5], etc. In the case of smart agriculture [6], it is mainly used to modernize the sector. Indeed, the deployment of connected objects allows farmers to have real-time information on the quality of soil, water, plants, etc. So, it offers farmers and producers the opportunity to reduce waste and improve productivity [7–9]. This ranges from the amount of fertilizer used, and the number of trips made by agricultural vehicles, to more efficient use of resources such as water and electricity.

With new IoT tools, farmers can monitor their field conditions for real-time decision purposes. These decisions can be manual or automated according to the data collected such as humidity, light, temperature, crop health, etc. [10]. Innovations in terms of low consumption and connectivity make it possible to extend the scope of IoT applications for smart agriculture and to deploy connected objects in vast environments with strong constraints. Long Range (LoRa) [11] is a radio communication protocol that constitutes the architecture of the system and which allows low-speed but above all long-range data transmission for all your IoT-connected objects. LoRaWAN (Long Range Wide Area Network) [12] is a communication protocol that works over an LPWAN (Low Power Wide Area Network) network, while LoRa technology is the physical layer of the network. Casamance is a particular region that

* Corresponding Author: El Hadji Malick NDOYE, elm.ndoye@univ-zig.sn

https://dx.doi.org/10.25046/aj070628

stands out from other areas of Senegal due to its enormous edaphic, climatic, hydrographic, and biogeographical potential. The high rainfall combined with very dense vegetation in the region makes it sometimes easier to use LPWAN communication technologies such as LoRa rather than 802.15.4 which is a short-range technology. This paper proposes a new LoRa technology-based network architecture suitable for a smart agriculture solution in the Casamance region of Senegal. The contributions can be divided into two major parts:

- A new LoRa network architecture based on the classical one combines short and long-range LoRa radio links for making it possible to carry out smart agriculture in rural areas, particularly in the southern area of Senegal.
- Evaluation of the proposal is conducted through an analytical study used on the deployment to determine the optimal distance between two gateway nodes to reduce induced interference.

The following presents the rest of the paper. Section 2 presents the context and the motivation, and the state of the art on smart agriculture is presented in section 3. Section 4 focuses on the LoRa networks while section 5 presents the proposed architecture of the network. In section 6, we present an analytical study followed by a discussion. Finally, the paper ends with a conclusion in section 7.

2. The context and motivation

Senegal is located at the most western projection of the African continent into the Atlantic Ocean, at the confluence of Europe, Africa, and the Americas, and a crossroads of major maritime and air routes. The climate is dry tropical and characterized by two seasons: a dry season from November to June and a rainy season from July to October. Three types of vegetation: forest in the south, savanna in the center, and steppe in the north. The southern part called Casamance is the rainiest part of the country. Casamance has a very varied range of soils:

- the tropical ferruginous soils leached with concretion and armor;
- weakly ferritic soils;
- undegraded halomorphic soils;
- and moderately organic hydromorphic soils.

The climate is humid tropical with rainfall above 1200mm which decreases towards the east. It is the wettest part of Senegal (rainfall > 800mm) thanks to the presence of the monsoon flow for more than 8 months and we find 20% arable land in the country. The area benefits from a hydrographic network made up of a set of permanent and seasonal waterways called bolongs. The main one is the Casamance River (350 km).

Basse-Casamance is the administrative region of Ziguinchor and is composed of forest areas and agricultural areas (mainly rice and peanuts). The forest areas are characterized by their density, particularly to the southwest of Oussouye covered with a Guinean vegetation cover made up of large palm groves, thousand-year-old cheese groves, lianas, teak, and giant mango trees. Up to the Gambian border, the forests, often state-owned, are dense and protected. The rice fields largely dominate the agricultural landscape of Basse-Casamance as shown in Figure 1.



Figure 1: Rice fields in Casamance

Casamance is the richest agricultural region in Senegal. Agricultural production is also limited by the lack of large-scale mechanization and the inadequacy of agricultural inputs used.

Today, to win the challenge of large-scale agriculture, it is important to develop new digital tools capable of water irrigation, analyzing soils, proposing the best inputs, or making decisions in the face of warning messages, etc. To achieve this, it is important to make better choices on the sensor equipment. This must be motivated by the transmission range of the sensors, the financial cost, etc. In addition, the criteria for placing the sensors must be well studied to deploy the minimum number of sensor nodes for maximum coverage of the area. Otherwise, it is important to set up a network architecture capable of satisfying coverage and especially monitoring needs. A multi-linear architecture with LoRa radio links seems to be the choice to achieve smart agriculture in the specific case of Casamance. The main goal is to propose an end-to-end wireless LoRa architecture to collect the maximum data for irrigation or soil quality detection use.

3. The state of the art

Today, the world faces many phenomena such as climate change, floods, and threatening bush fires exposing certain particularly African populations to famine. Faced with these threats, it is necessary to find new effective methods to meet the challenges of food self-sufficiency. The democratization of the IoT has enabled digitalization in the field of agriculture. The sector is developing new tools to optimize the management of farms and harvests. This then makes it possible to set up intelligent methods of irrigation and water conservation [7,13–15], to measure the quality of the soil or crops [16,17]. Indeed, data and connected objects allow farmers to improve their productivity through network and data management models. The sensors deployed in agricultural areas are diverse and varied according to the characteristics including range, cost, etc. The sensors deployed in agricultural areas are diverse and varied according to the characteristics including range, cost, etc. Regarding the range, there are several types of 802.15.4 short-range wireless technologies or LPWANs such as LoRa, NB-IoT, SigFox, etc. Consequently, many authors focus on different models of networks for smart agriculture. The SAIoT is a model for smart agriculture described in [18]. The aim goal is to develop a real-time monitoring system for soil properties such as temperature, moisture, etc. The model is composed of three levels: the Farm level, the Server level, and the Client level. It implements also

decision support advisory models for Pest & Disease forewarning, Crop Disease identification using image analysis and SMS (Short Message System) based alerts. Connected objects used to collect information use 802.15.4/Zigbee radio links, which therefore limits the range of the solution to small areas. The authors of the paper [19] present another application of smart farming. It consists of an intelligent irrigation system based on 802.15.4-type sensors. The sensor nodes are deployed in a restricted space of over a hundred meters. LPWAN technologies seem to be best suited to large agricultural areas due to the long transmission range. Studies have recently been conducted in this direction to see the contribution of these wireless technologies in smart agriculture. The paper presented in [20] reviews the most used LPWANs in the case of smart agriculture. It presents the Nb-IoT, Sigfox, and LoRa technologies with the characteristics. NB-IoT is an LPWAN solution based on existing cellular networks. In public mode, the principle of LoRa is similar to Sigfox. Transmitters once positioned in an area covered by the network will send their information directly to the Cloud of the LoRa or Sigfox operator. This requires that each sensor be provisioned with a subscription from the operator to be able to communicate with the Cloud. LoRa technology also offers the possibility of operating in private LoRa mode. This means it is possible to build your own LoRa network using gateways for local coverage. No need for subscriptions with operators in private LoRa mode. This mode of operation of LoRa is very interesting in the case of vast agricultural areas devoid of operator network coverage, in particular, called white areas. This explains why a lot of work in the field of smart agriculture uses solutions based on LoRa. An efficient LoRa-based smart agriculture is presented in [21] for humidity and temperature monitoring uses. In [22], the authors present novel watering management based on sensor nodes using LoRa technology for communication. The solution is designed for the specific case of Vietnam and real-time data management. The model describes three types of nodes. The gateway node performs the core functions of data collection from the nodes via LoRa. Node 1 and sensor node 1 has been designed to act as a relay between node 1 and the gateway Nodes 1 and sensor 2 have been designed to act as a relay between node 1 and the gateway while node 2 monitors the quality of the water coming from the pump. The authors present an architecture composed of LoRa nodes for monitoring roles, a gateway for data collection, and a server for and a server for data analysis and decision-making. They evaluate the architecture in terms of humidity and temperature. Another very important factor in agricultural monitoring is the physical topology of the area. Indeed, it can determine the way to deploy the nodes in the zone. Most agricultural areas in southern Senegal are surrounded by waterways for much of the season. In such a configuration it is easier to carry out a deterministic and linear deployment of the nodes. In addition, given the very large size of these areas, it is necessary to opt for equipment with a long communication range. In this context, linear LoRa networks are more suitable for the considered environments. The advantage of linear wireless networks resides above all in their simplicity of deployment. However, in the specific case of linear LoRa networks, it is important to rethink this deployment taking into account the characteristic parameters of LoRa such as range, data rates, propagation model, etc. Much research on the deployment of nodes in wireless networks has been proposed over the years. Very recently, they have turned to LoRa-type wireless networks. Paper [23] presents a LoRa network of campuses deployed on several sites. It describes a method of deploying LoRa sensors to cover communications. The network covers an area of 20,000 meters. Because of the scope of LoRa nodes, it would be more interesting to be able to deploy the nodes over a wider area. In paper [24], it is about covering the city of Southampton in the UK using a LoRa network. The deployment method used in this paper focuses on the optimal distance between gateway nodes that can ensure maximum network coverage. The connectivity assessment tested the efficiency of the deployment with the message delivery rate. Interferences if not well managed can negatively impact the deployment of LoRa nodes. In the paper [25], the authors present a modeling of interferences based on a mathematical model. The analysis of parameters such as the SIR or the SNIR provided a clear idea of the impact of the propagation model on interference in LoRa networks. However, the work carried out was done over a radius of 12 km, which is insufficient in the case of vast agricultural areas. An interference study was conducted from an anechoic chamber to give an idea of large-scale deployment. The evaluation of parameters such as the spread factor [26], and the coding rate [27] sets the conditions for deploying nodes. Other works focus on the deployment of LoRa networks based on criteria different from those mentioned so far as in [28]. This paper evaluates the impact of interference on deployment in terms of BER [29] according to various spread factors, coding rates, and bandwidth

As indicated above, the LoRa technology seems to be the most appropriate in the particular case of the Casamance context. Indeed, the transmission range is around 20 km and the possibility of configuring it in private mode makes it possible to cover vast agricultural areas without an operator network. The other advantage of LoRa is the reduction of financial costs. In the following, we review LoRa technology. This is to explain in detail the physical and logical components of the LoRa architecture.

4. An overview of the LoRa technology

Like LPWAN networks, the LoRaWAN network uses a star topology configuration. The architecture of a LoRa network is essentially composed of four components: end devices, gateways, a network server, and an application server. Figure 2 shows these different elements. The end devices are equipped with LoRa modules to allow them to send and receive radio signals and communicate with the gateways. In addition to being able to exchange with the end nodes, the gateways have an interface that allows them to have TCP/IP-based communication with the network server. The network component that has the most work to do is the network server. It allows the administration of the whole network, in other words, it is the conductor. Its role is to route information from the end nodes to the corresponding application servers and to remove redundant packets.

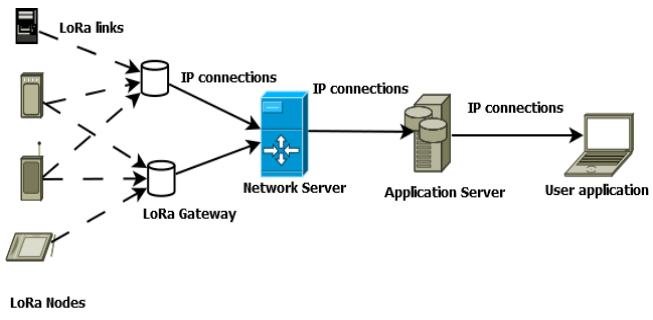


Figure 2: Classic LoRaWAN architecture

4.1 The LoRa physical layer

LoRa is based on spread spectrum modulation which uses the Chirp technique signal in which the frequency increases (up-chirp) or decreases (down-chirp) with time. Spread spectrum is a technique that allows an information signal to be transmitted over a bandwidth several times greater than the minimum bandwidth required. The modulated signals have a constant amplitude with a variable frequency [30]. Spread spectrum modulation reduces energy consumption and increases resistance to interference. In LoRa, signals are modulated in the ISM (Industrial, Scientific, and Medical) frequency band without a license, which varies according to the region (e.g. 868 MHz in Europe, 915 MHz in North America)[30].

The performance of LoRa nodes depends mainly on the following parameters:

The **spreading factor (SF)** is a quantity defined by the ratio between the chip rate and symbol rate. A chip represents pulses of a spread spectrum code and a symbol represents several chips. A chirp contains 2^{SF} bits per symbol. In a LoRaWAN network, the spreading factor is an integer between 7 and 12 [31]. This parameter allows us to vary the transmission range and rate and to send simultaneously on a given channel. A high spreading factor implies a long transmission range and a low data rate (see Table 1).

Table 1: LoRa spreading factor

Spreading factor (for UL at 125 kHz)	Bit rate	Range (Depends on Terrain)	Time on Air (for an 11-bytes payload)
SF10	980 bps	8 km	371 ms
SF9	1760 bps	6 km	185 ms
SF8	3125 bps	4 km	103 ms
SF7	5470 bps	2 km	61 ms

Bandwidth is the difference between the maximum and minimum frequency. High bandwidth values result in high data rates. The transmission time of a symbol (Ts), the spreading factor, and the bandwidth (BW) are related by the following equation :

$$Ts = \frac{2^{sF}}{Bw} [8]$$

Coding rate: LoRa uses the FEC (Forward error correction) technique to detect errors in transmitted frames. Extra bits are added in the frame, more precisely at the payload level. In LoRa these extra bits are controlled by the coding rate (CR). The values that the code rate can take are 4/5, 4/6, 4/7, or 4/8 [30].

4.2 The LoRaWAN protocol

LoRaWAN (Long Range Wide Area Network) is an open protocol proposed by LoRa Alliance (the Organization in charge of the promotion of LoRaWAN). It corresponds to the MAC (Medium Access Control) layer of the OSI (Open Systems Interconnection) reference model. LoRaWAN is designed to connect objects equipped with LoRa modules and powered by battery energy to the Internet. A LoRaWAN network consists of LoRa nodes, gateways equipped with LoRa modules, and an interface for connecting to the Internet, a network server, and an application server. LoRaWAN uses the channel access technique of the ALOHA type [12]. That is, each node in the network can transmit whenever it wishes. According to the LoRaWAN specification [10], Data transmission is performed on multiple channels. Indeed, each device has a maximum of 16 channels that can be preconfigured before a terminal joins a LoRaWAN network. A LoRa frame consists of a preamble, a header, the payload, and an error control field. If the header contains information about the CRC (Cyclic Redundancy Check) and the CR (Coding Rate), the frame is said to be explicit, otherwise, it is implicit.

The LoRaWAN specification defines three modes of operation for end nodes suitable for various IoT applications. Each mode of operation is commonly referred to as a class.

Class A: This is the default class that must be implemented by all end nodes in the LoRaWAN network. In class A, the terminal sends a message to the network server (uplink) through a channel randomly chosen among those configured in the terminal with a data rate, the terminal node opens two reception windows RX1 and RX2 respectively after one second and 2 seconds (default values) following the uplink transmission during which it waits for the server transmission (downlink). If the server does not transmit, then the message is resent. The recommended retransmission limit is eight (8). Due to the short listening window, class A is the most energy efficient.

C**lass B**: Terminals, where class B is enabled, consume more power than those implementing class A because these terminals offer more receive windows. The ping slots are opened at regular time intervals to activate the downlink. They are synchronized through a beacon broadcast in the network to provide the synchronization references. The synchronization time is called BEACON_PERIOD. In addition to these specific features, class B also implements the operating mode of class A.

C**lass C**: Terminals using class C are constantly listening to the channel to receive messages from the network server. In a terminal, class C and class B should not be activated simultaneously. Unlike Class B, which opens only two receive windows, Class C additionally opens an RXC receive window for continuous listening.

As we have already shown, LoRa through its physical and logical components is an ideal candidate for smart agriculture in a country where network coverage is not guaranteed in agricultural areas. Moreover, given the abundance of rainfall and the physical architecture of agricultural areas, it is more judicious to arrange the sensors linearly because of the constraints imposed by the environment. Indeed, arable land is surrounded by waterways. So, for the needs of traceability and economy, it is preferable to deploy them linearly. The works cited above have already shown the interest that researchers have developed in such architecture. It should be noted however that the classical architecture of LoRa as it is presented cannot guarantee good coverage in the particular case of Casamance. The very complex nature of the landscape due to the high rainfall does not guarantee good wired deployment between the gateway nodes and the routers. Indeed, to succeed in such a deployment, specific devices are needed that can protect the cables over long distances, which is likely to be very expensive. The deployment of sensors must take into account several factors such as the extent of the area, the absence of telecommunication operators, and the financial costs. For all these reasons, it is therefore judicious to propose an architecture for deploying sensor

nodes in a linear mode without a wired connection. In other words, the proposed architecture combines LoRa radio links throughout the network with short- and long-range links.

5. The proposed architecture

The contribution of this paper is the implementation of a network architecture based on LoRa for the needs of smart agriculture. Such an architecture must take into account the specific characteristics of the area considered, the financial costs, etc. For all these reasons, the architecture is composed of several types of nodes with different physical characteristics. It is therefore a multi-linear architecture composed of several levels combining short-range and long-range nodes depending on the level as shown in figure 3. The choice of the linear topology is explained for the needs of ease of deployment and location of the nodes, and the complexity of the relief. Indeed, as indicated above, the vegetation in Casamance is very dense and very diversified. Several types of end device nodes can be described according to their role.

- Some sensors measure the rate of wetting on the leaves to anticipate diseases related to leaf humidity and reduce the use of phytosanitary treatments.
- Some sensors measure soil humidity and temperature to control root irrigation and delay the arrival of pests.
- Some measure humidity and air temperature to detect frost, and anticipate the life cycle of insects and diseases.

The end device nodes form the first level and the gateways the second level. Each gateway plays the role of cluster head for all the nodes associated with it. There are mainly two types of links in the network as shown in figure 3:

- Low LoRa links: The Low LoRa links are short-range links, and are used in communications between the nodes of the first level and those of the second level. SF parameters in [7, 8] are used.
- High LoRa links: The high LoRa links are long-range links, and are used in communications between the nodes of the second level in the sink node. SF parameters in this case are in [9, 10, 11].

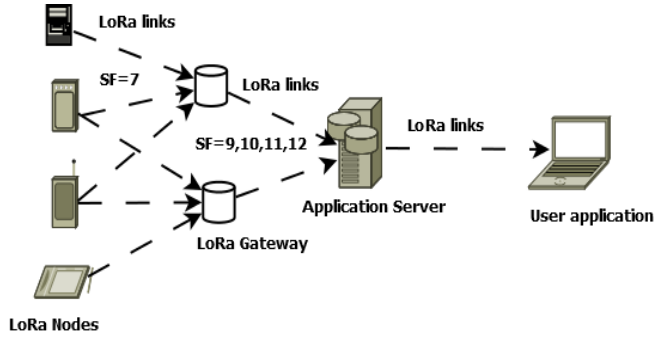


Figure 3: Proposed network architecture

Such architecture faces several challenges. In particular, the challenge of deployment for better optimization of resources such as memories, storage spaces, or computing units. It is well-known that interference in wireless networks can negatively and significantly impact performance. Therefore, in such a network, it is imperative to settle the question of deployment to minimize interference. Figure 4 below gives an illustration of the node's deployment in a rice field in Casamance. In the following, we present an analytical analysis method to define an optimal deployment of network nodes.

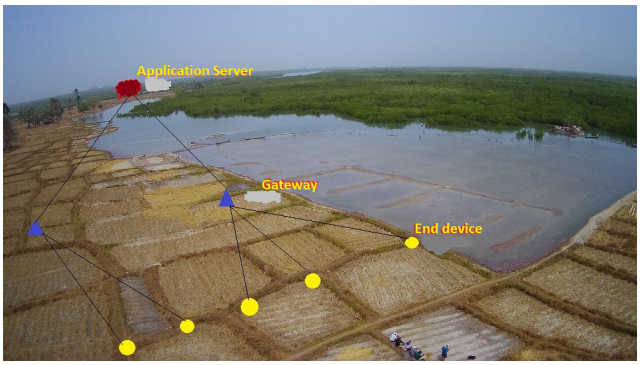


Figure 4: Illustration of node deployment in a rice field

6. Analytical study and discussion

The objective of this study is to define a method for deploying the nodes of the architecture described above. It is therefore a question of defining from certain well-chosen parameters the minimum distance between two gateway nodes which makes it possible to achieve the best deployment. These induce the choice of the spreading factor, the bandwidth, the coding rate, etc. A good interference management policy makes it possible. We can identify the case where there is a risk of interference at the node G_i, interference induced by the activity of the node G_j. We can define the gateway node G_j and the set of nodes j as the interfering nodes for the node G_i. In this condition, an optimal node deployment needs to set a minimum distance between two gateways. In wireless networks, the SIR (Signal-to-Interferences Ratio) [32] is a good indicator of the quality of the received signal. Thus, depending on the desired reception power, it is possible to rely on the SIR to determine the optimal distance between two nodes. In the specific case of this paper, the analysis must determine the optimal distance between two gateways to minimize interference. Indeed, given the proposed architecture, we can consider the gateways as cluster heads. Thus, the distance between the gateway nodes has a significant impact on the deployment.

The cluster i is formed by the gateway G_i which is the cluster head and the set of nodes n_i. In the same way we define cluster j as shown in figure 5. To determine the minimum distance between G_i and G_j, we set:

- T_i: the transmission power of node i
- T_j: the transmission power of node j
- T_G: the transmission power of node G_j
- d_i: the distance between node i and gateway G_i
- d_j: the distance between node j and gateway G_i
- d_{ij}: the distance between gateway i and an interfering gateway
- d: the distance between two adjacent gateways

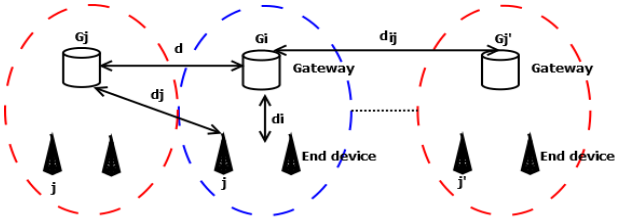


Figure 5: Distance between gateway illustration

The propagation model represents the signal and therefore the interference induced during transmission. To take into account the environmental fluctuations of the medium, Log Normal Shadowing is required in this study because of its random nature. The path loss at an arbitrary distance $d > d_0$ from the transmitter is given by the following equation (1):

$$PL_{dB}(d) = PL_{dB}(d_0) + 10n\,log(\frac{d}{d_0}) \pm X_\sigma \qquad (1)$$

where n is the path loss attenuation factor, $PL_{dB}(d_0)$ is the path loss at a reference distance d_0, and $X\sigma$ is a zero-mean Gaussian distributed random variable. The standard deviation σ is expressed in dB. So, for a given transmission power P_T, the received power is provided by the equation (2):

$$P_{R\sigma} = P_T - PL_{dB}(d) \qquad (2)$$

The received signal at the gateway G_i from node i is the sum of multiple terms: the first, is the desired signal received from node i; the second is the interferences caused by the set of gateways G_j around G_i, the third is the interference generated by all the sets of nodes j associated to the gateways G_j. Let be K' the set of node j associated with a gateway Gj. We can define the SIR [20] as:

$$SIR_{dB} = S(i) - \sum_{j=1}^{x} S(G_j) + \sum_{K\prime \epsilon G_j} \sum_{k\prime=1}^{MAX} s(k') \qquad (3)$$

$$S(y) = T_y - PL_{dB}(d_0) - 10 * n * \log\left(\frac{d_y}{d_0}\right) \pm X_\sigma(y) \quad (4)$$

where *y* can be a gateway or an end device node, *x* is the number of interfering gateways for the gateway G_i and MAX is the number of nodes associated with a given gateway.

The coexistence of several radio links of different ranges is a particularity of the architecture. Short radio links of 2 km are used at the first level while those of long-range over 4 km are at the second level of the architecture. As shown in figure 3, we note various spreading factors according to the level. Thus, the first level spreading factor is set to 7 while it is in the range {9,10,11} in the second level. Moreover, for the sake of simplicity, it is assumed that a gateway communicates with only one of its nodes at a given time. So, for *x* interfering gateways, there are *x* interfering nodes at a specific time. In addition, the transmitting power for the end devices nodes is assumed to be the same value.

So, the *SIR* becomes:

$$SIR_{dB} = S(i) - \sum_{j=1}^{x} S(G_j) - \sum_{j=1}^{x} S\,(j) \qquad (5)$$

$$SIR_{dB} = T_i - PL_{dB}(d_0) - 10 * n * log\left(\frac{d_i}{d_0}\right) \pm X_\sigma(i) - \sum_{j=1}^{x}(T_G - PL_{dB}(d_0) - 10 * n * log\left(\frac{d_{ij}}{d_0}\right) \pm X_\sigma(Gj)) - \sum_{j=1}^{x}(T_i - PL_{dB}(d_0) - 10 * n * log\left(\frac{d_j}{d_0}\right) \pm X_\sigma(j)) \qquad (6)$$

Environmental fluctuations in the considered area for coverage issues are supposed to be homogeneous. Environmental conditions are known to be very stable in Senegal according to the season. Thus, $X_\sigma(i), X_\sigma(G_j)$, and X_σ(j) follow the same Gaussian distributed random variable $N\sim(0,\sigma)$. According to these considerations we have:

$$SIR_{dB} = (2x+1) * X_\sigma + (1-x) * T_i + (2x-1) * PL_{dB}(d_0) \quad (7)$$

$$-x * T_G - 10 * n * log\left(\frac{d_i}{d_0}\right) + 10 * n * \sum_{j=1}^{x} log\left(\frac{d_{ij}}{d_0}\right) + 10 * n * \sum_{j=1}^{x} log\left(\frac{d_j}{d_0}\right) \qquad (8)$$

We assume that the deployment is deterministic and uniform. Therefore, the relationship between d_{ij} and d is given by the formula (9):

$$d_{ij} = j * d \qquad (9)$$

where *j* represents the number of interfering gateways for a given gateway *i*.

In addition, for all the nodes j associated with the same gateway, we can affirm that the distance d_j is almost equivalent to the distance d_{ij}. Thus, from equation (8) we have:

$$SIR_{dB} = (2x+1) * X_\sigma + (1-x) * T_i + (2x-1) * PL_{dB}(d_0) - x * T_G - 10 * n * log\left(\frac{d_i}{d_0}\right) + 10 * n * \sum_{j=1}^{x} log\left(\frac{d * j}{d_0}\right) + 10 * n * \sum_{j=1}^{x} log\left(\frac{d * j}{d_0}\right) \qquad (10)$$

The simplified formula gives:

$$SIR_{dB} = (2x+1) * X_\sigma + (1-x) * T_i + (2x-1) * PL_{dB}(d_0) - x * T_G - 10 * n * log\left(\frac{d_i}{d_0}\right) + 20 * n * x * \log\frac{d}{d_0} + 20 * n * \log\prod_{j=1}^{x} j \qquad (11)$$

Let be:

$Z = (2x+1) * X_\sigma$ and

$$K = (1-x) * T_i + (2x-1) * PL_{dB}(d_0) - x * T_G - 10 * n * log\left(\frac{d_i}{d_0}\right) + 20 * n * x * \log\frac{d}{d_0} + 20 * n * \log\prod_{j=1}^{x} j \qquad (12)$$

Z is so a gaussian distributed random variable $N\sim(0,(2x+1)*\sigma^2)$.

Let be

P_{succes} the probability that $SIR \geq SIR_{Thresold}$.

$$P_{succes}[Z + K \geq SIR_{Threshold}] = p \quad (13)$$

$$\Leftrightarrow P_{succes}[Z \leq SIR_{Threshold} - K] = 1 - p \quad (14)$$

This is equivalent to:

$$P_{succes}\left[\frac{Z}{\sigma*\sqrt{(2x+1)}} \leq \frac{(SIR_{Threshold}-K)}{\sigma*\sqrt{(2x+1)}}\right] = 1-p \qquad (15)$$

$$\Leftrightarrow P_{succes}\left[H \leq \frac{(SIR_{Threshold}-K)}{\sigma*\sqrt{2(x+1)}}\right] = 1-p \qquad (16)$$

where $H = \frac{Z}{\sigma * \sqrt{(2x+1)}}$ is a gaussian random variable $N \sim (0,1)$.

According to the repartition function of $N \sim (0,1)$, it is assuming that:

$$\frac{(SIR_{Threshold} - K)}{\sigma * \sqrt{(2x+1)}} = F^{-1}(1-p) \quad (17)$$

where F^{-1} is the inverse function of *F*.

We deduce from this equation the value of K.

$$K = SIR_{Threshold} - \sigma * \sqrt{(2x+1)} * F^{-1}(1-p) \quad (18)$$

Finally, the distance between two adjacent gateways is given according to equations (12) and (18):

$$\log d = \left(A + B * x + C * \sqrt{2x+1} + D * log \prod_{j=1}^{x} j\right) \Big/ 20 * n * x$$

where:

$$A = SIR_{dB} - T_i + PL_{dB}(d_0) + 10 * n * \log \frac{d_i}{d_o}$$

$$B = T_i - 2 * PL_{dB}(d_0) + T_G + 20 * n * \log d_0$$

$$C = -\sigma * F^{-1}(1-p)$$

$$D = -20 * n$$

The mathematical analysis makes it possible to unambiguously determine the optimal deployment of gateway nodes minimizing interference. This allows us to define a reusable deployment pattern to cover a large agricultural area. The choice of pattern takes into account the characteristics of the environment considered as environmental fluctuations due to rainfall, the nature of the soil, etc. The pattern is defined as the set formed by a gateway G_i and the interfering gateway nodes G_j which guarantee a given reception threshold. The minimum distance d as a function of the size of the pattern is proposed in the following.

Evaluation is done according to the conditions described in [33]. It should be noted that in Senegal the work on smart agriculture is currently in a purely experimental phase [34,35]. Therefore, the work does not focus on the propagation model. Hence the choice of parameters for an environment similar to that of Senegal such as in Lebanon's environment. Path loss parameters are shown in Table 2.

Table 2: The path loss parameters

Parameters	Values
T_{ip}	14 dB
T_{Gj}	20 dB
d_i	2000 m
n	4,179
σ	7,2 dB
Reception probability	0.9
$PL_{db}(d_0)$	102,86

It is shown according to the paper [22] that for a spreading factor of 12 and a reception rate of around, the ISR threshold is equal to 6 dB.

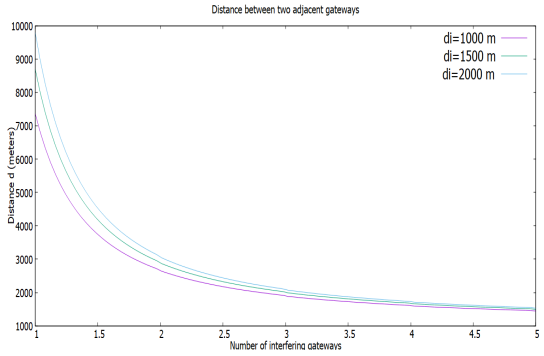


Figure 6: Number of interfering gateways as a function of distance

Figure 6 represents the minimal distance between two adjacent gateways. It shows the influence of the distance between an end device node and its gateway on the deployment. Indeed, it shows the maximum number of interfering gateways for a given distance d_i. The curve shows that the number of interfering gateways is inversely proportional to the distance. Indeed, when the distance is excellent, the interference is less and therefore the interference can be limited to 1 gateway. On the other hand, when the distance is reduced (less than 3000 m), the interferences are very significant. In this case, the number of interferents increases without exceeding a limit of 5. It should also be noted that the minimum distance does not exceed the defined distance d_i.

In the specific case of the rice fields of Casamance, it is preferable to minimize the equipment that makes up the architecture because of the presence of water around the fields. Therefore, it is more advisable to choose the deployment of gateways over long distances (example d = 6Kms).

7. Conclusions and future work

Smart agriculture is seen today as a very important lever for winning the challenge of food self-sufficiency in a world where resources such as water are becoming increasingly scarce. Indeed, the new techniques of irrigation, monitoring of soil quality, of temperature make it possible to improve productivity and maintain cultivable soils. In the particular case of Casamance in Senegal, the practice of smart agriculture must take into account the very dense rainfall in this environment. This implies the choice of the most appropriate sensors and wireless communication technology. Moreover, it is necessary to have an optimal deployment of the sensors to have good coverage of the zone. This induces the study of a network architecture that ensures this purpose.

A new LoRa network architecture based on the classical one is presented in this paper. This architecture combining short and long-range LoRa radio links make it possible to carry out smart agriculture in rural areas, particularly in the southern area of Senegal. It focuses on the specific case of the rice field for irrigation system purposes. An analytical study on the deployment helps us to define a model of deployment of the nodes particularly the gateways. This study is based on factors such as the SIR, the reception probability, the distance between nodes, etc. In future work, we first plan to simulate the network in Omnet++ to confirm the analytical study. This involves conducting a performance study in terms of throughput, end-to-end delay, and energy to analyze the advantages and limits of the model. Then, it is about deploying LoRa sensors and gateways in a real environment.

References

[1] P. Sethi, S.R. Sarangi, "Internet of Things: Architectures, Protocols, and Applications," Journal of Electrical and Computer Engineering, **2017**, 1–25, 2017, doi:10.1155/2017/9324035.

[2] L. Atzori, A. Iera, G. Morabito, "The Internet of Things: A survey," Computer Networks, **54**(15), 2787–2805, 2010, doi:10.1016/j.comnet.2010.05.010.

[3] K.N. Swaroop, K. Chandu, R. Gorrepotu, S. Deb, "A health monitoring system for vital signs using IoT," Internet of Things, **5**, 116–129, 2019, doi:10.1016/j.iot.2019.01.004.

[4] D.S.R. Krishnan, S.C. Gupta, T. Choudhury, "An IoT based Patient Health Monitoring System," in 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), IEEE: 01–07, 2018, doi:10.1109/ICACCE.2018.8441708.

[5] T.A. Khoa, M.M. Man, T.-Y. Nguyen, V. Nguyen, N.H. Nam, "Smart Agriculture Using IoT Multi-Sensors: A Novel Watering Management System," Journal of Sensor and Actuator Networks, **8**(3), 45, 2019, doi:10.3390/jsan8030045.

[6] M.R.M. Kassim, "IoT Applications in Smart Agriculture: Issues and Challenges," in 2020 IEEE Conference on Open Systems (ICOS), IEEE: 19–24, 2020, doi:10.1109/ICOS50156.2020.9293672.

[7] D. Thakur, Y. Kumar, S. Vijendra, "Smart Irrigation and Intrusions Detection in Agricultural Fields Using I.o.T.," Procedia Computer Science, **167**, 154–162, 2020, doi:10.1016/j.procs.2020.03.193.

[8] D.M. Andeme Bikoro, S. Fosso Wamba, R.M. Etoua, "Effective Contribution of Internet of Things (IoT) in Smart Agriculture: State of Art," Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, **443 LNICST**, 219–233, 2022, doi:10.1007/978-3-031-06374-9_14.

[9] M. Cicioğlu, A. Çalhan, "Smart agriculture with internet of things in cornfields," Computers & Electrical Engineering, **90**, 106982, 2021, doi:10.1016/J.COMPELECENG.2021.106982.

[10] G.S. Nagaraja, A.B. Soppimath, T. Soumya, A. Abhinith, "IoT Based Smart Agriculture Management System," in 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), IEEE: 1–5, 2019, doi:10.1109/CSITSS47250.2019.9031025.

[11] J.P.S. Sundaram, W. Du, Z. Zhao, "A Survey on LoRa Networking: Research Problems, Current Solutions and Open Issues," 2019.

[12] J. Haxhibeqiri, E. de Poorter, I. Moerman, J. Hoebeke, "A Survey of LoRaWAN for IoT: From Technology to Application," Sensors, **18**(11), 3995, 2018, doi:10.3390/s18113995.

[13] E.A. Abioye, M.S.Z. Abidin, M.S.A. Mahmud, S. Buyamin, M.H.I. Ishak, M.K.I.A. Rahman, A.O. Otuoze, P. Onotu, M.S.A. Ramli, "A review on monitoring and advanced control strategies for precision irrigation," Computers and Electronics in Agriculture, **173**, 2020, doi:10.1016/j.compag.2020.105441.

[14] L. García, L. Parra, J.M. Jimenez, J. Lloret, P. Lorenz, "IoT-based smart irrigation systems: An overview on the recent trends on sensors and iot systems for irrigation in precision agriculture," Sensors (Switzerland), **20**(4), 2020, doi:10.3390/S20041042.

[15] R.K. Kodali, M.S. Kuthada, Y.K. Yogi Borra, "LoRa Based Smart Irrigation System," in 2018 4th International Conference on Computing Communication and Automation (ICCCA), IEEE: 1–5, 2018, doi:10.1109/CCAA.2018.8777583.

[16] V. Saiz-Rubio, F. Rovira-Más, "From Smart Farming towards Agriculture 5.0: A Review on Crop Data Management," Agronomy, **10**(2), 207, 2020, doi:10.3390/agronomy10020207.

[17] P. Sumathi, R. Subramanian, V. Karthikeyan, S. Karthik, "Retracted: Soil monitoring and evaluation system using EDL-ASQE: Enhanced deep learning model for IoT smart agriculture network," International Journal of Communication Systems, **34**(11), 2021, doi:10.1002/dac.4859.

[18] K.A. Patil, N.R. Kale, "A model for smart agriculture using IoT," in 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), IEEE: 543–545, 2016, doi:10.1109/ICGTSPICC.2016.7955360.

[19] E.M. Hasiri, A. Asniati, M.A. Suryawan, R. Rasmuin, "The Implementation of Smart Farming Application Based on the Microcontroller and Automatic Sprinkler Irrigation System of Agricultural Land," Advances in Science, Technology and Engineering Systems Journal, **5**(2), 174–179, 2020, doi:10.25046/aj050222.

[20] H. Klaina, I.P. Guembe, P. Lopez-Iturri, M.Á. Campo-Bescós, L. Azpilicueta, O. Aghzout, A.V. Alejos, F. Falcone, "Analysis of low power wide area network wireless technologies in smart agriculture for large-scale farm monitoring and tractor communications," Measurement, **187**, 110231, 2022, doi:10.1016/J.MEASUREMENT.2021.110231.

[21] S.J. Suji Prasad, M. Thangatamilan, M. Suresh, H. Panchal, C.A. Rajan, C. Sagana, B. Gunapriya, A. Sharma, T. Panchal, K.K. Sadasivuni, "An efficient LoRa-based smart agriculture management and monitoring system using wireless sensor networks," International Journal of Ambient Energy, 1–4, 2021, doi:10.1080/01430750.2021.1953591.

[22] T.A. Khoa, M.M. Man, T.-Y. Nguyen, V. Nguyen, N.H. Nam, "Smart Agriculture Using IoT Multi-Sensors: A Novel Watering Management System," Journal of Sensor and Actuator Networks, **8**(3), 45, 2019, doi:10.3390/jsan8030045.

[23] G. Ferre, F. Rivet, R. Tajan, E. Kerherve, "Design and Deployment of an IoT Network Based on LoRa," in 2017 27th EAEEIE Annual Conference (EAEEIE), IEEE: 1–5, 2017, doi:10.1109/EAEEIE.2017.8768583.

[24] P.J. Basford, F.M.J. Bulot, M. Apetroaie-Cristea, S.J. Cox, S.J. Ossont, "LoRaWAN for Smart City IoT Deployments: A Long Term Evaluation," Sensors, **20**(3), 648, 2020, doi:10.3390/s20030648.

[25] L. Irio, R. Oliveira, "Modeling the Interference caused to a LoRaWAN Gateway due to Uplink Transmissions," in 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), IEEE: 336–340, 2019, doi:10.1109/ICUFN.2019.8806142.

[26] B. Reynders, S. Pollin, "Chirp spread spectrum as a modulation technique for long range communication," in 2016 Symposium on Communications and Vehicular Technologies (SCVT), IEEE: 1–5, 2016, doi:10.1109/SCVT.2016.7797659.

[27] G. Baruffa, L. Rugini, V. Mecarelli, L. Germani, F. Frescura, "Coded LoRa Performance in Wireless Channels," in 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), IEEE: 1–6, 2019, doi:10.1109/PIMRC.2019.8904298.

[28] V.R. Stoynov, V.K. Poulkov, Z. v. Valkova-Jarvis, "Interference Management in LoRaWANs — Overview and Simulation Study," in 2018 Advances in Wireless and Optical Communications (RTUWO), IEEE: 251–256, 2018, doi:10.1109/RTUWO.2018.8587904.

[29] T. Elshabrawy, J. Robert, "Analysis of BER and Coverage Performance of LoRa Modulation under Same Spreading Factor Interference," in 2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), IEEE: 1–6, 2018, doi:10.1109/PIMRC.2018.8581011.

[30] G. Ferre, A. Giremus, "LoRa Physical Layer Principle and Performance Analysis," in 2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS), IEEE: 65–68, 2018, doi:10.1109/ICECS.2018.8617880.

[31] J.-T. Lim, Y. Han, "Spreading Factor Allocation for Massive Connectivity in LoRa Systems," IEEE Communications Letters, **22**(4), 800–803, 2018, doi:10.1109/LCOMM.2018.2797274.

[32] G.P. Efthymoglou, P.S. Bithas, A.G. Kanatas, "Exact SNR and SIR analysis in Poisson wireless networks," Electronics Letters, **53**(5), 356–358, 2017, doi:10.1049/el.2016.2378.

[33] R. el Chall, S. Lahoud, M. el Helou, "LoRaWAN Network: Radio Propagation Models and Performance Evaluation in Various Environments in Lebanon," IEEE Internet of Things Journal, **6**(2), 2366–2378, 2019, doi:10.1109/JIOT.2019.2906838.

[34] C.S.M. Babou, B.O. Sane, I. Diane, I. Niang, "Home Edge Computing Architecture for Smart and Sustainable Agriculture and Breeding," in Proceedings of the 2nd International Conference on Networking, Information Systems & Security - NISS19, ACM Press, New York, New York, USA: 1–7, 2019, doi:10.1145/3320326.3320377.

[35] L. Ndiaye, K. Gueye, S. Ouya, "Proposal for an alert and monitoring solution based on the SIP and IoT protocol: the case of agriculture," in The 4th International Conference on Networking, Information Systems amp Security., ACM, New York, NY, USA: 1–6, 2021, doi:10.1145/3454127.3457617.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 256-261 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Optimizing Sensors Locations for Tsunami Warning System

Mikhail Lavrentiev[*,1], Dmitry Kuzakov[1], Andrey Marchuk[1,2]

[1]*Institute of Automation and Electrometry SB RAS, Novosibirsk, 630090, Russia*

[2]*Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, 630090, Russia*

ARTICLE INFO

Article history:
Received: 06 August, 2022
Accepted: 08 December, 2022
Online: 20 December, 2022

Keywords:
Initial sea surface displacement
Tsunami source parameters
Part of wave profile

ABSTRACT

To reduce the time necessary for determination of tsunami source parameters it is proposed to optimize the location of sensors system and to use only a part of the measured wave profile. Based on computation, it is possible to balance the number of sensors in use and the time period after the earthquake needed to obtain tsunami wave parameters. Numerical experiments show that even a part of the wave period (compared to ¼ of the entire period) provides enough information to get the wave amplitude within the well-known concept of calculation in advance. This is due to the application of Fourier theory in the form of orthogonal decomposition of the measured wave profile. It is important that the proposed algorithm requires only a few seconds using regular personal computer. Using the real depth profile offshore Japan we compute the time required to calculate the wave amplitude (with 10 percent accuracy) in case of one, two and three sensors. Optimization could be performed in terms of minimal time required to get the wave profile. It is also possible to calculate the sensor network design, which provides the maximal time between wave parameters determination and the wave approaching nearest coast. The new feature here observed is that optimal positions of sensors are different if one needs minimizing time to detect tsunami wave or maximazing time it takes the wave to approach the nearest cost after recovering the wave parameters at source. This may require to rearrange decision making at tsunami warning centers.

1. Introduction

This paper is an extension of the work originally presented in the conference OCEANS 2021: San Diego – Porto [1].

As a rule, the numerical modeling modulus of tsunami warning system should contain the following three major parts: wave generation (earthquake magnitude and hypocenter location recalculated in terms of the initial seabed displacement), wave propagation and inundation of the dry land. For example, our reference code, the MOST (Method of Splitting Tsunami) software package (see [2,3]), has all these parts. Alternative codes have similar structure [4,5]. It is natural to use "initial disturbance" at tsunami source as the "input data" to calculate tsunami wave propagation from the source to the coastal area. Displacement of the sea bed (due to the earthquake) could be evaluated on the basis of knowledge of the Earth crust structure and the location of the earthquake epicenter. So, regardless the underlying reasons, in any case the so-obtained initial displacement is nothing but approximation. We propose to use instead the sea surface displacement at tsunami source, obtained by processing the measured profile of the real tsunami wave. These measured data are available due to the rather well-developed system of bottom pressure sensors, located at sea bed around Pacific Ocean, given in Figure 1 from [6].

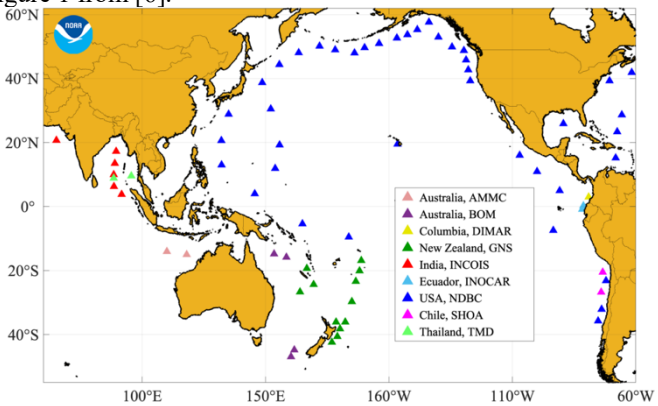


Figure 1. Location of Dart buoys around the Pacific Ocean [6]

* Corresponding Author: Mikhail Lavrentiev, mmlavrentiev@gmail.com

https://dx.doi.org/10.25046/aj070629

In addition, there exist systems of bottom sensors, connected to the dry lend (processing centers) by a cable, for example DONET (Dense Oceanfloor Network system for Earthquakes and Tsunamis) [7] and S-net (Seafloor observation network for earthquakes and tsunamis) [8] being developed in Japan. Figure 2, taken from [8], shows the location of S-net deep-water stations eastward of Honshu Is.

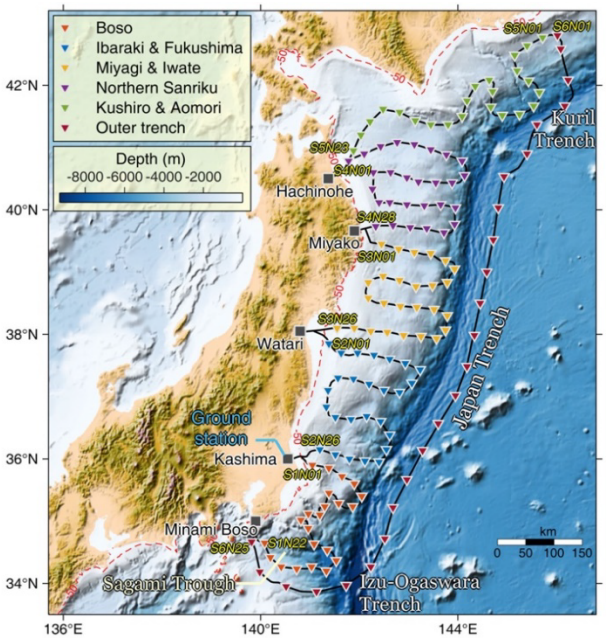


Figure 2. Configuration of the S-net observation system along the Japan trench [8]

In some publications, the optimal positioning of the sensor network refers to the location of sensors that allow detection of a tsunami wave in the shortest time after its occurrence (see, for example, [9] and the references therein). The present article, however, deals with the configuration of a network of deep-water recorders, not for the earliest detection of the tsunami occurrence, but for determining the water surface displacement profile with the required accuracy in the shortest possible time.

In [10] the S-net observation data are used reconstruct the source parameters by the sequential multiple linear regression method.

The original idea of calculation in advance was used in [11]. It consists in introduction of so-called Unit Sources (UnSs). These are nothing but rectangles 50x100 km – typical size of the sea bed displacement in case of 7.5 M earthquake. It is possible to simulate numerically tsunami wave propagation initiated by a given displacement form. Covering the particular subduction zone with such UnSs (along with the typical for this zone shape of the initial sea bed displacement) one can create "in advance" the database of the calculated tsunami wave time series (synthetic mareograms). Suppose that we have a sea bottom pressure sensor (a number of those are installed over the Pacific Ocean) which is able to report the parameters of the tsunami wave passing over in a real time mode. Having such a measured wave profile we now approximate it as a linear combination of several calculated wave profiles, each being initiated by one of the UnSs [12]. In fact, such a database exists and is available for use. System of UnSs covers major subduction zones around Pacific Ocean [13].

As was noted in [14], even a part of the measured wave profile is enough to reconstruct the main tsunami wave parameters within the calculation in advance strategy. We believe that 10% error is acceptable to decide if a particular wave is dangerous in a given location. Here we show how the necessary time to determine the wave amplitude at source depends of the number of sensors one uses. This is done by the example of three different locations of the "composed" tsunami source within a particular water area offshore Japan.

The rest of this paper is composed as follows. We first describe the "calculation in advance" data inversion strategy. Then the idea of orthogonal decomposition method is given. The method has a very low computational cost and makes it possible to obtain an approximation of tsunami source parameters by using only a part of the measured wave profile. Digital bathymetry offshore the central part of Honshu Island (Japan) is then introduced. Setting up of numerical experiments is then described, including location of model tsunami sources and artificial water level measurement sensors. The obtained numerical results are given in Section 3. Finally, these results are briefly discussed.

2. Setup of Numerical Experiments

2.1. Problem Statement

As part of the general problem of estimating the tsunami hazard in the shortest possible time, let us determine the role of the location of the system of deep-water sensors. We will assume that, at the cost of allowing 10% error in obtained wave amplitude, the source parameters are determined from data from a single sensor, the first one the wave has reached. Moreover, we will use only part of the wave period, which is about 1/4 of this first period (having the positive phase first). The exact value of this NPWP (Necessary Part of Wave Period) is determined automatically, on the basis that the amplitude values practically stop growing as the wave passes over the sensor [14]. Let us consider the ratio of the number of sensors to the time required to determine the approximate values of the wave parameters in the tsunami source.

Following [3], we will assume that the epicenter of the earthquake is located within a certain zone (subduction zone). This assumed zone is covered by 8 rectangles of size 50x100 km - the typical size of the seafloor deformation zone during an earthquake of magnitude M=7.5 (the accepted threshold value of magnitude for the formation of tsunami waves). In each of these rectangles (Unit Sources - UnSs) is placed a perturbation with a shape characteristic of seafloor deformation resulting from earthquakes in this subduction zone. The seafloor deformation during a stronger earthquake will be approximated by a linear combination of four of these UnSs with some coefficients. The problem of determining the source parameters comes down to determining the set of these amplification coefficients.

2.2. Orthogonal Decomposition Method

To elaborate a fast and robust algorithm for the determination of the above amplification coefficients, Fourier series theory was used, as it deals in particular with optimal approximation of a given function with the linear combination of functions being orthogonal and normalized.

Let *f(t)* be the measured time series of the wave profile (marigram). One can consider the data from DART buoy or nay alternative sensor. By $f_k(t)$, $k=1,\ldots,n$, let us denote the calculated wave profiles, obtained (by direct numerical solutions to the linear or nonlinear shallow water system [15]) at the same point of sensor location. The parameter k means that the source of this wave is

located at the *k*-th UnS having "standard" shape. As was shown numerically, these functions, $f_k(t)$, could be considered linearly independent. To describe better the measured wave, one should determine the coefficients, b_k, for the linear combination of these functions, for which the below difference in L_2 norm is the smallest:

$$\int_0^T \left(\sum_{k=1}^N b_k f_k(t) - f(t)\right)^2 dt \rightarrow \min \quad (1)$$

In such a statement the problem of determination of the wave parameters at source is reduced to the one of the optimal approximation of a given function *f(t)* by the linear combination of functions $\{f_k(t)\}$.

The Fourier series theory states that the coefficients of such optimal approximation (1) are exactly the Fourier coefficients of expansion of *f(t)* in a series with respect to $\{f_k(t)\}$ (see, for example, [16]), provided that the system $\{f_k(t)\}$ is orthogonal and normalized:

$$\left(f_i(t), f_j(t)\right) = \int_{t_0}^{T} f_i(t) f_j(t) dt = 0, \quad (2)$$

$$\left(f_i(t), f_i(t)\right) = 1. \quad (3)$$

So, the Fourier series theory was used in [17] to design the algorithm to determine coefficients for optimal approximation of the measured tsunami wave. This algorithm, tested in [1,14,18], consists of several stages.

At the first stage the "marigrams" $\{f_k(t)\}$ obtained from the unit sources should be recalculated to meet the requirements (2)-(3).

Then, the function *f(t)* (tsunami wave profile) should be expanded to the Fourier series with respect to recalculated synthetic marigrams. At the last stage the obtained Fourier coefficients are transformed into the sought coefficients in (1).

Note that the described algorithm has a very low computational cost as mostly simple algebraic operations are involved. We consider calculation of integrals (scalar products and L_2 norms as simple operations, too.) During the performed numerical experiments it takes less than a second using regular personal computer.

2.3. Domain for Numerical Study – Bathymetry

All numerical studies were arranged at the water area offshore Kanto Region, central part of the Honshu Island (Japan). JODC database (see [19]) was used to create the gridded bathymetry for the computation area, given in Figure 3. The area under study is bounded in East-West direction by 137.0º and 143.0º East Longitude, and in North-South direction by 32.0º and 37.0º North Latitude. Spatial grid steps are 223 m (East-West direction) and 273 m (North-South direction. Array of computational nodes has size of 2000x2500.

2.4. Description of Numerical Studies Setup

Tsunami sources in subduction zones usually extend along deep-water troughs. In this case, the source is most effectively reconstructed by sensors located in the direction of the short axis of such sources. In the area under consideration the deep-water trench is oriented from North to South (Figure 3), so the possible sources are extended in this direction, and the virtual sensors are located along the meridian, which, in our opinion, allows reconstructing effectively the initial displacement of the water surface using the data of a minimum number of sensors. Along the northeastern coast of Honshu Island, where S-net sensors are installed (Figure 2), the possible tsunami sources are extended in the direction of Japan Trench, which allows an optimal choice among the already available deep-water recorders. This is not yet possible in the area of the Kanto Peninsula.

We consider offshore water area at the Izu-Bonin subduction zone. Suppose that tsunami sources are expected within the strip covered by eight UnSs (having size 50x100 km each), indicated as white rectangles, see the Figure 1. In total the 100x400 km zone is considered. Simulating a rather strong earthquake, the Composed Sources (CSs) are studied. These are nothing but linear combinations of four neighboring USs with certain amplification coefficients. We choose the following values for the coefficients *(0.7, 0.8, 1.2, 1.3)*. It means that for each of the CS_i, *i=1,2,3*, (indicated in Figure 3) the upper-left UnS has the coefficient *0.7* (in the linear combination, which determines our Composed Source), the lower-left – *0.8*, upper-right – *1.2* and lower-right – *1.3*. The amplitude of sea surface indignation of each of these CS_i does not exceed *+150* cm and its shape is given in Figure 4. Form of such a CS is typical in the area of study. Our task is to determine these amplification coefficients.

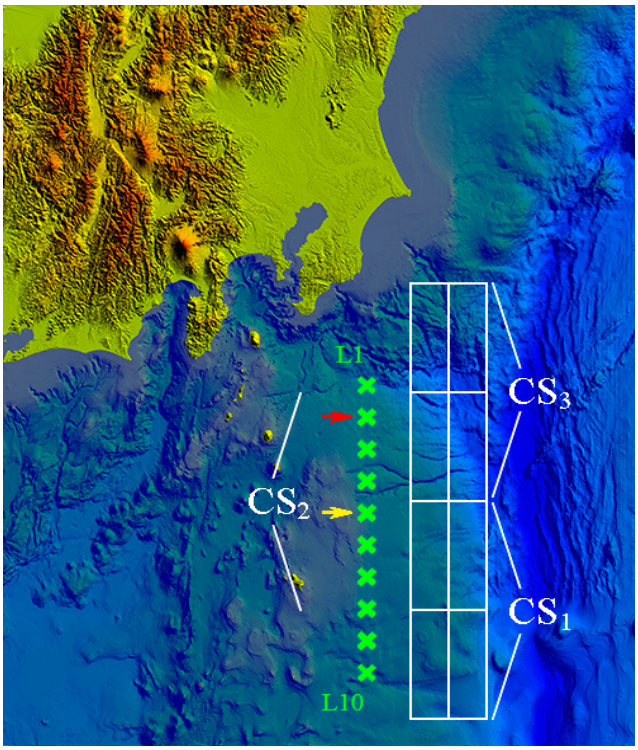


Figure 3: Digital bathymetry of water area under study. Artificial tsunami sources are indicated as CS_i. Green crosses indicate artificial sensors

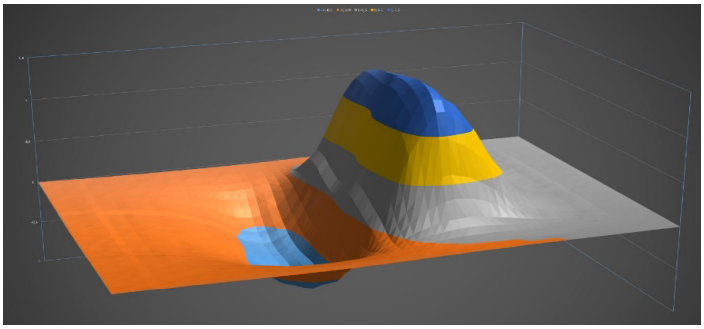


Figure 4: Visualization of CS_i – water level in the artificially composed tsunami source having size 100 x 200 km

A number of virtual tsunami detection sensors are located close to the subduction zone. Exact positions of these 10 "tsunameters" are indicated by green crosses. Numbering is arranged from up (sensor *L1*) to down, the "last one is indicated as *L10*. These "virtual tsunameters" *Li* could coincide with the existing sensors from S-net and DONET, the other could suggest possible positions of DART buoys.

3. Numerical Results

A series of numerical calculations of tsunami propagation from each of the 8 UnSs under consideration was performed to obtain synthetic tsunami profiles, on the basis of which the source parameters recovery algorithm was constructed. For numerical modeling, the MOST algorithm [2,3], which correctly describes tsunami propagation in a sufficiently deep-water area, was used. This method uses difference scheme based on spatial direction splitting difference scheme to approximate the system of nonlinear shallow water differential equations. Calculations can result in wave series at all 10 tsunami recorder locations from each UnS.

A series of computational experiments were carried out to assess the efficiency of reconstructing a composed sources *CSi*, *i=1,2,3*, by using the data from deep-water recorders located at different points in the region. The gridded bathymetry for these calculations is described in Section 2.2, and the field of water surface displacement in the basis (Unit) Source, constructed using the seismic source model in elastic half-space [12], is presented in Figure 5.

Based on the results of numerical modeling of tsunami propagation generated by each of *8* such UnSs, a database of synthetic mareograms (wave time series) at the locations of 10 virtual deep-water detectors (given in Figure 3) was created. Wave series calculated at these virtual tsunameters location were used to reconstruct the source composed of *4* UnSs with different amplification coefficients. To estimate the quality of restoration of composed sources *CSi*, *i=1,2,3,* the results of tsunami modeling generated by each *CSi* at different sensors were used. The time needed for recovery and deviation estimate from the exact wave profile were determined and are presented in the form of tables.

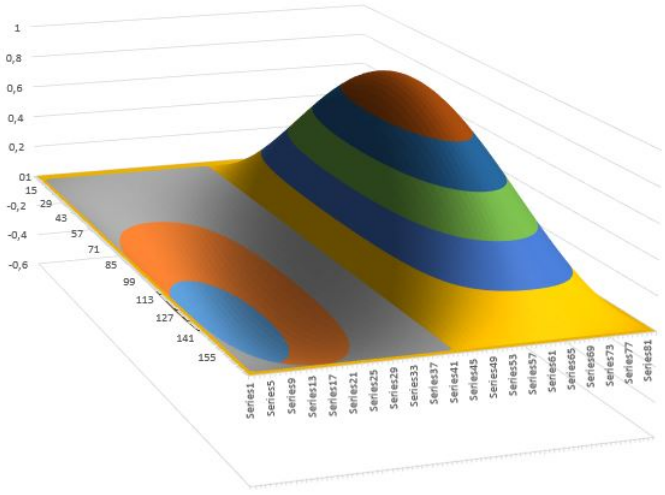


Figure 5. Water surface displacement in the Unit Source having size 50 x 100 km

Results of numerical simulation described in [14] are given in Tables 1-3 below.

The following "critical" time designations are used in the tables:

NPWP – necessary part of the wave profile, measured in seconds, to determine correctly (within 10% error) the coefficients in approximation (1);

T1 (sec) – time moment in which the wave first maximum appears at the given virtual sensor;

T2 (sec) – the time it takes the wave to approach the nearest coast after the moment where the coefficients in (1) are correctly determined.

Table 1: Source CS1 (see Figure 1). Wave travel time to the nearest coast is equal to 1260 sec.

No Rec.	NPWP, sec.	T1, sec	T2, sec
L1	1031	740	229
L2	794	630	466
L3	697	525	563
L4	552	435	708
L5	478	375	782
L6	469	358	791
L7	383	367	877
L8	385	372	875
L9	348	336	912
L10	317	306	943

Table 2: Source CS2 (Figure 1, the middle one). Wave travel time to the nearest coast is equal to 886 sec.

No Rec.	NPWP, sec.	T1, sec	T2, sec
L1	443	343	443
L2	358	285	528
L3	262	266	624
L4	319	303	567
L5	354	340	532
L6	389	373	497
L7	382	368	504
L8	473	366	413
L9	502	391	384
L10	550	443	336

Let us consider three scenarios of a possible tsunami. The combined sources of the form shown in Figure 4 have three possible positions *CSi* (*i=1,2,3*), see Figure 3. The analysis of Tables 1-3 shows that if we have only one working sensor at our disposal, it should be placed at point *L5* (yellow arrow in Figure 1). The guaranteed detection time of the source parameters will be *478* seconds (case of *CS3* source). In the case of any other location of the sensor in at least one scenario under consideration, this time will be longer. However, if we want to have the longest time available before the wave arrives on shore after determining the source parameters, we should choose sensor position *L2* (red arrow in Figure 3). This time will be *466* sec (again for the source *CS3*). With a different location of the sensor, we will find a case where this time will be shorter.

If we have two sensors at our disposal, they must be chosen differently. For the fastest detection time of the source parameters in the "worst" of three cases it will be sensors *L3* and *L10* (shown by yellow arrows in Figure 6), and this time will be *317* seconds (source *CS3*). That is, the addition of one sensor allows reducing the required time by almost 1.5 times. If we want to achieve the longest possible time before the wave arrives on the shore, we should place the sensors at the points *L1* and *L4* (shown by red arrows in Figure 6). This time will be *488* seconds (source *CS1*), which is slightly longer than in the case of one sensor.

Table 3: Source CS3 (the upper one at Figure 1). Wave travel time to the nearest coast is equal to 719 sec.

No Rec.	NPWP, sec.	T1, sec	T2, sec
L1	231	218	488
L2	242	233	477
L3	296	284	423
L4	314	304	405
L5	394	333	325
L6	500	412	219
L7	626	503	93
L8	782	615	0
L9	935	740	0
L10	1049	849	0

For the convenience of further analysis, the data about the optimal location of one and two sensors are collected in Tables 4 and 5.

Table 4: Optimal location of 1 and 2 sensors for fastest determination of source parameters

	Source parameters detected, sec		
	CS1	CS2	CS3
L5	394	354	478
L3+L10	296	262	317

It is clear from Table 4 that by introducing an additional sensor to the monitoring system it is possible to reduce the time for source parameters determination. The "reduction rate" for the necessary time is similar for all considered scenarios *CSi, i=1,2,3*. We may assume that doubling the number of sensors leads to the reduction of time to detect the source parameters compared to 30-40%. Additional numerical experiments are needed to get the exact numbers, which should definitely depend on a particular subduction zone.

Table 5: Optimal location of 1 and 2 sensors to have larger travelling time for the wave to approach nearest dry land after source parameters determination

	Wave travelling time to approach dry land after coefficients determination, sec		
	CS1	CS2	CS3
L2	477	528	466
L1+L4	488	567	708

In case we are interested in time one has after obtaining the source parameters before the wave approaches nearest shore, Table 5 shows, that this time increases after taking into consideration an additional sensor. However, the observed gain is practically negligible for *CS1* scenario and is compared to 40 sec (about 8%) for *CS2* scenario. Note that in these two scenarios we are able to determine source parameters rather fast, see Table 4.

In case of the most unfavorable *CS3* with the largest time required for the source parameters determination, wave traveling time to the nearest dry land increases valuably according to Table 5. This fact leads to a conclusion, that it is far nontrivial to choose the optimality criteria for the monitoring system.

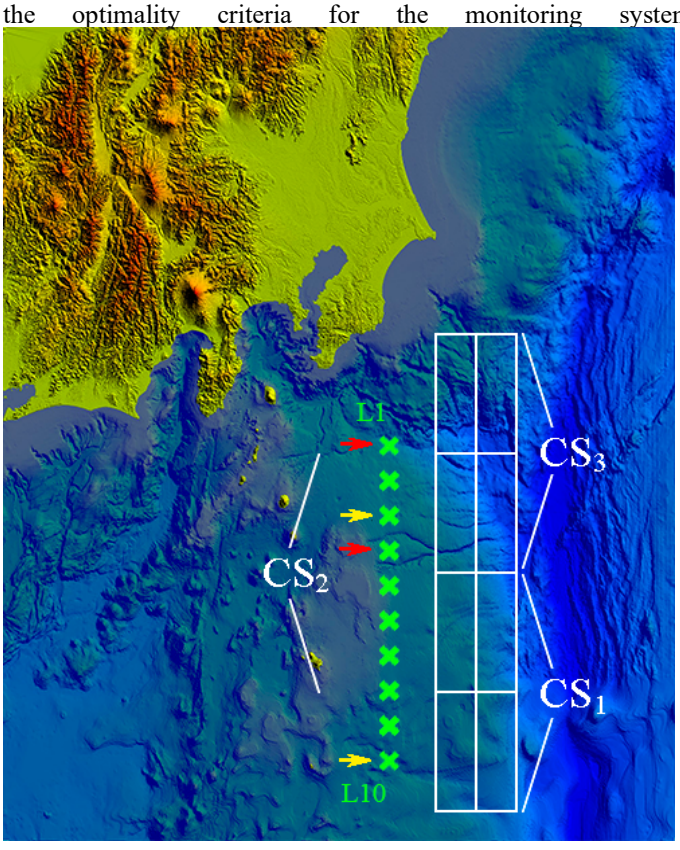


Figure 6: Visualization of the digital bathymetry. Model Composed tsunami sources are indicated as *CSi*. Green crosses indicate artificial sensors

4. Discussion

Most studies on the organization of the tsunami detection system (location of sensors) to provide data to tsunami warning services have the objective of determining the wave parameters at the source as quickly as possible. By doing so, the distribution of maximum wave heights along the coastline can be predicted more quickly. However, numerical results show that a different location of the sensors would provide more time before the wave arrives at the nearest shore. Considering the fairly densely populated coastline of Japan, this parameter seems to be at least as important as the fastest wave detection. This should be taken into account when planning the location of additional sensors.

The optimal location of the surveillance system can be calculated (not just being randomly suggested) according to the selected criterion. The more sensors are used, the better the value of the selected criterion will be – the source parameters will be determined faster or after determining the source parameters there will be more time until the wave reaches the nearest shore.

Thus, when switching from one to two sensors in the model case under consideration, the time for which the algorithm determines the source parameters decreases from *478* to *317* sec (Table 4), that is, practically one and a half times (source *CS3*). Let us note that in other considered scenarios as well (sources *CS1* and *CS2*) this time decreases from *394* to *296* sec (*1.33* times less time needed) for the *CS1* case and from *354* to *262* sec (*1.35* times less) for the *CS2* source.

For the criterion “maximum time remaining until the wave reaches the nearest shore”, this time increases from *466* to *488* sec when an additional sensor is added (Table 4). This is a comparative increase, but for the *CS3* source scenario, the increase in time to analyze and make decisions to evacuate the population is *242* sec (from *466* to *708* sec). This certainly may save lives.

Thus, when designing a monitoring system for tsunami warning centers, it is advisable to choose a criterion for the optimality of the system, taking into account various factors (including the cost of creation) and to optimize the location of sensors.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research was carried out under state contract with IAE SB RAS (121041800012-8) and with ICMMG SB RAS (0315-2019-0004) and under state contract with ICMMG SB RAS (0315-2019-0004).

References

[1] D. Kuzakov, M. Lavrentiev, An. Marchuk, “Toward the optimization of measurement system for tsunami warning,” in 2021 International Conference Global Oceans 2021: San Diego - Porto, In Person Virtual, September 20-23, 2021, Town and County San Diego, doi: 10.23919/OCEANS44145.2021.9705912

[2] V.V. Titov, F.I. Gonzalez, “Implementation and testing of the method of splitting tsunami (MOST) model,” in NOAA Technical Memorandum ERL PMEL-112, USA,1977.

[3] E. Gica, M. Spillane, V.V. Titov, C.D. Chamberlin, J.C. Newman, “Development of the forecast propagation database for NOAA's Short-term Inundation Forecast for Tsunamis (SIFT),” in NOAA Tech. Memo. OAR PMEL-139, NTIS: PB2008-109391, 2008.

[4] X. Wang, W.L. Power, COMCOT: A Tsunami Generation Propagation and Run-Up Model, GNS Science, 2011.

[5] A.C. Yalciner, B. Alpar, Y. Altinok, I. Ozbay, F. Imamura, "Tsunamis in the Sea of Marmara: Historical Documents for the Past, Models for Future," Marine Geology, **190**, 445-463, 2002.

[6] DART (Deep-ocean Assessment and Reporting of Tsunamis). Available online: URL http://nctr.pmel.noaa.gov/Dart/ (accessed on 20.07.2022).

[7] DONET system concept. Available online: URL https://www.jamstec.go.jp/donet/e/ (accessed on 31.07.2022).

[8] I.E. Mulia, K. Satake, "Synthetic analysis of the efficacy of the S-net system in tsunami forecasting," Earth Planets Space **73**(36), 2021. https://doi.org/10.1186/s40623-021-01368-6

[9] A. Ferrolino, R. Mendoza, I. Magdalena, J.E. Lope, "Application of particle swarm optimization in optimal placement of tsunami sensors," Peer J Comput. Sci. **6**:e333, 2020. DOI 10.7717/peerj-cs.333

[10] L. Yao, K. Goda, "Hazard and Risk-Based Tsunami Early Warning Algorithms for Ocean Bottom Sensor S-Net System in Tohoku, Japan, Using Sequential Multiple Linear Regression," Geosciences 12.9: 350, 2022.

[11] D.B. Percival, D.W. Denbo, M.C. Eble, E. Gica, H.O. Mofjeld, M.C. Spillane, L. Tang, V.V. Titov, “Extraction of tsunami source coefficients via inversion of DART® buoy data,” Nat. Hazards, **58**(1), 567–590, 2011, doi: 10.1007/s11069-010-9688-1.

[12] V.K. Gusiakov, "Residual displacements at elastic halfspace surface," in Conditionally correct problems of mathematical physics of interpretation geophysical surveys, Eds. Computing Center SB RAS, Novosibirsk, 1978, 23-51. (In Russian).

[13] National Data Buoy Center. Available online: URL https://www.ndbc.noaa.gov/obs.shtml (accessed on 20.07.2022).

[14] M. Lavrentiev, D. Kuzakov, An. Marchuk, "Fast Determination of Tsunami Source Parameters," Advances in Science, Technology and Engineering Systems, **4**(6), 61-66, 2019, doi: 10.25046/aj040608

[15] J.J. Stoker, Water Waves. The Mathematical Theory with Applications, Interscience publishers: New York, NY, USA, 1957.

[16] D.R. Kincaid, E.W. Cheney, Numerical analysis: mathematics of scientific computing, **2**, American Mathematical Soc., 2002.

[17] A. Romanenko, P. Tatarintsev, “Algorithm for reconstruction of the initial surface disturbance at the tsunami epicenter”, NSU J. of Information Technologies, **11**(1), 113-123, 2013.

[18] D. Kuzakov, M. Lavrentiev, An. Marchuk “Reconstruction of tsunami source by a part of wave time series before the first maximum” Journal of Physics: Conference series, 2019, 1268 012042, p 1-9

[19] J-DOSS|JODC [Electronic resource]: http://jdoss1.jodc.go.jp/vpage/depth500_file.html (accessed 18.07.2022).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 262-268 (2022)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Transfer and Ensemble Learning in Real-time Accurate Age and Age-group Estimation

Anh-Thu Mai [*,1,2], Duc-Huy Nguyen[1,2], Thanh-Tin Dang[1,3]

[1]*Vietnam National University Ho Chi Minh City (VNU-HCM), Vietnam*

[2]*Training Program of Excellent Engineers in Vietnam (PFIEV), Ho Chi Minh city University of Technology (HCMUT) , Vietnam*

[3]*Faculty of Electrical and Electronics Engineering, Ho Chi Minh city University of Technology (HCMUT), Vietnam*

ARTICLE INFO

Article history:
Received: 22 August, 2022
Accepted: 04 December, 2022
Online: 20 December, 2022

Keywords:
Age Estimation
Convolutional Neural Networks
Transfer Learning
Ensemble Learning

ABSTRACT

Aging is considered to be a complex process in almost every species' life, which can be studied at a variety of levels of abstraction as well as in different organs. Not surprisingly, biometric characteristics from facial images play a significant role in predicting human's age. Specifically, automatic age estimation in real-time situation has begun to affirm its position as an essential process in a vast variety of applications. In this paper, two approaches are addressed as solutions for such application: prediction of accurate age and age group by using the two most fundamental techniques in the domain of deep learning – convolutional neural networks (CNNs) and deep neural networks (DNNs). In summary, this work can be split into two main key contributions. By applying a novel hierarchical aggregation built on the base of neural network developed from the training dataset, in the first stage, features extraction, the convolutional activation features are extracted from the captured facial image. As soon as this part is done, the features classification step is performed, in which Softmax Regression (SR) and majority vote classifiers are applied to predict accurate age and age group respectively. The effectiveness of the designed model was showed satisfactorily in the experimental results, which emphasizes the promising of the solution and indicates another direction for future development of algorithms and models in the field of machine learning.

1. Introduction

Automatic age prediction from face images can be employed in a huge number of scenarios, from video surveillance [1], health-care system, health management and statistics purposes [2], performance of individuals recognition [3], advertisement according to users' age, or even criminal investigations [4]. The problem is the reliability of the results since the acquired images could be in unconstrained conditions because of lighting or pose, or when the photographed individuals do not aware of the cameras' presence. Thanks to the advent of computational approaches, these constraints become less and less stressful for automatic computer vision algorithms. Furthermore, this analysis must be completed in a limited time interval of milliseconds, owing to the fact that it is performed in real time condition. Consequently, it is necessary for the solution of this task to be both fast and accurate.

In order to achieve this goal, first of all, face detection is performed using a tool in real-time computer vision's deep learning based-face detector, OpenCV. Particularly, this method is built mainly on ground of Single Shot Detector (SSD) and ResNet backbone for the purpose of increasing the accuracy and accelerating the estimation process, which could be considered to be a compensation for some acceptable major setbacks such as having unconscious biases in the training set or detecting darker-skinned people less accurately than lighter-skinned ones when comparing with Haar-likes cascades [5] or histogram of oriented gradients (HOG) [6]. After the previous step has been fulfilled, by training with an abundant dataset, an age prediction model is created, thanks to which the features corresponding to each age are extracted using a structure of convolutional neural networks (CNNs). Meanwhile, a binary large object (blob) constructor pre-processes the input images and extracts the face region of interest (ROI) so as to calibrate the convolutional architecture for fast feature embedding (caffe) based model that is utilized in this work.

*Corresponding Author: Anh-Thu Mai, Email: thu.mai.k18@hcmut.edu.vn

Table 1: Outline of Some Outstanding Studies on Age Prediction

Publication	Feature representation	Face database/Database size	Algorithm	Evaluation protocol	Performance/Accuracy
[14]	2D shape, raw pixel values	Private/500	Regression	500 train, 65 test	MAE 4.3
[15]	Anthropometric model	HOIP	Classification	Leave-one-out	CS 57.3%(M), 54.7%(F)
[16]	Ages pattern subspace (AGES)	FG-NET/1002 MORPH	Regression	Train on FG-NET, Test on MORPH	MAE 8.83; CS 70%
[9]	Active appearance model	FG-NET/1002	Hybrid	Leave-one-person-out	MAE 4.97; CS 88%
[17]	BIF	MORPH II/55,132	Regression	50% train, 50% test	MAE 4.2
[11]	Active appearance model, label distribution	FG-NET/1002 MORPH II/55132	Classification	Leave-one-person-out (FG-NET), 10-fold cross-validation	MAE 4.8(FG-NET), 4.8(MORPH II)
[18]	Kullback-Leibler/raw intensities	LFW/150,000 MORPH II/55,132 FG-NET/1002	Classification (CNN)	N/A	MAE 2.8(FG-NET), 2.78(MORPH)

The remaining of this paper is arranged as follow. In the next section, we will briefly review some researches inspiring our work. The architecture of the model will be explained precisely in section 3. The overviewed algorithms and the detailed experimental results will be discussed in section 4 and 5 respectively. The last section provides a summary of this work as well as offers a conclusion and possible future development.

2. Related works

Since the early 2000s, automatic age prediction has drawn researcher's attention in the field of machine learning and has undergone a renewed interest some years later as a consequence of the availability of larger databases having more real-annotated date than in the past. Some noticeable researches are shown in Table 1 to represent the evolution of the methods. Overall, the process of this task involves two main phases: features extraction to pick out particular aging characteristics and features classification to divide these features.

Feature extraction stage. The use of a classifier based on support vector machines (SVMs) and support vector regression (SVR) is widely used in several distinct works, whereas other researchers, who concentrated on achieving a faster and better performance, relied mainly on the combination of textural and local appearance. Facial images analysed by those methods will be capture the variation in shape and intensity. Nevertheless, the applied of deep learning algorithm, especially CNNs and Deep Neural Networks (DNNs) in age estimation has become more and more popular [7][8]. Particularly, the work of using ordinal regression and multiple output CNNs for this task by Niu. has demonstrated superior performance in comparison with other methods, and acts as motivation factor for the development our the deep learning approach.

Pattern classification. It has been proposed, including, among others, to use SVMs and SVR [9], Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA) [10], along with their regularized and kernelized versions, neural networks and their variant of Conditional Probability Neural Network [11]. In multi-task classification issues, Softmax Regression (SR) [12] shows that it outperforms other methods in terms of performance. A detailed examination on these aging classification approaches has been shown in [13].

3. Deep learning based-face detector

According to the aforementioned technique that is used in this paper, the first step that is performed is face detector using OpenCV. More precisely, the performed face detector is established on a SSD following a ResNet-like features which allow the result to be both faster and more accurate. Besides, this method uses DNNs module barked into OpenCV's library without adjusting parameters and build a lightweight output model. Since the face detection task is much more accurate for lighter-skinned people, we make a decision of calibrating the pre-trained face detector with a dataset of muti-ethnic so as to regain the balance among different skin's colors.

4. Accurate age prediction

Before building the model from pre-trained weights, compiling it with some functions and then training, the facial images from the dataset were pre-processed. Figure 1 shows the overall approach.

4.1. IMDB-WIKI dataset

Among online available database, e.g. MORPH, FG-NET, Kaggle, etc. we chose public celebrities' images from IMDB-WIKI dataset, which offers approximate 0.5 million of images ranging from 0 to 100 years old. The age distribution of this dataset is shown in Figure 2. Furthermore, in this process, several images were randomly ejected so as to each age has the same amount of images. Therefore, about 62,000 images remain for training our CNNs.

4.2. Model's construction using VGG-like pre-trained weights:

In purpose of strengthen the quality of distinguished features, prior to training, a histogram equalization is applied for facial images. This pre-processing stage has claimed its importance from the received results.

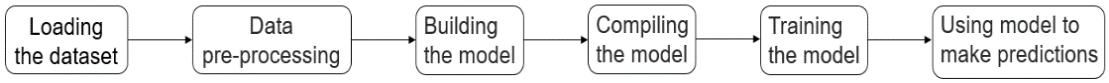


Figure 1: Ensemble process

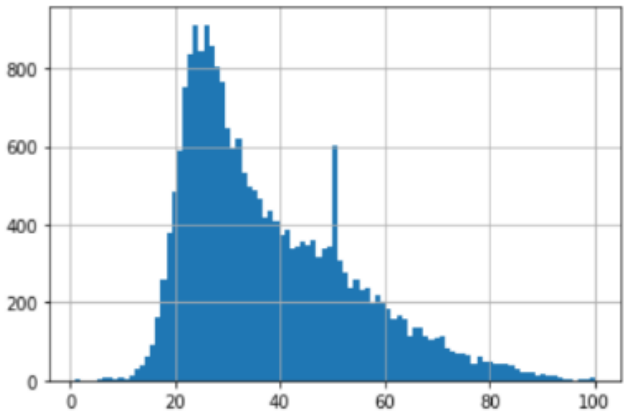


Figure 2: Age distribution histogram from the IMDB-WIKI dataset

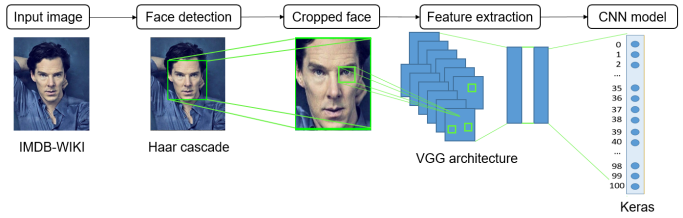


Figure 3: A general description of our procedure for developing CNN architecture in Keras

In this paper, we use VGG-like pre-trained weights and build our VGG model to deal with the experiments, which grants us the opportunity to fine-tune only the weights of the final layer instead of building and training from the scratch. The essential stage in the compilation process is believed to be selecting the optimizer and loss function. We employ "Adam", i.e. adaptive moment estimation, to adjust the learning rate in the performance of the optimization step. A lower learning rate would result in more accurate weights but in the interim it will stretch the handling time for other weights. Whereas for our loss function and accuracy calculating apparatus for the data use for validating, we utilize Categorical Cross Entropy [19], the foremost common choice for classifying patterns. As specified, transfer learning scheme was applied to train the model to assure that we obtain an output of the overall accurate age combined from 101 layers. Consequently, we might have an outcome with a superior achievement on human face's biometric characteristics.

When processing, an image with a fixed size 224x224 RGB was inputted into this network. The model has five convolutional blocks with a max-pooling at the end of each block. The number of filters are orderly 64, 128, 256, 512 and 512. A block with three dense layers is inserted at the end of the model. The final dense layer acts as a softmax function. This model type used is Sequential which is the most straightforward way to construct such model layer by layer in Keras, which totally matches with the VGG neural network architecture [20]. The process of building such model's substructure architecture is clearly shown step-by-step in Figure 3.

In order to address a multi-class classification task, as aforementioned, the classifiers that show the finest results are usually a CNN which has SR as the last layer. The previous layers are a combination of Convolutional layers and nonlinear activation functions and pooling. These layers can be considered as a tool to help extract features of the data, the last layer is SR, a transparent but exceptionally effective linear classifier. By this way, we are able to assume that a collection of one-vs-rest classifiers is trained simultaneously, supporting each other, so it will result in better performance than training separately each classifier. The adequacy of CNNs, is that the both models use for feature extraction and classification will be trained together. This classifier allows us to find feasible coefficients which pair with the found feature vector, while the role of the feature extraction model is to fine-tune the coefficients of the convolutional layers in order to receive a linear, consistent feature result in comparison with the classifier in the last layer. We also make a checkpoint to monitor model over iterations and prevent it from overfitting. The iteration which has the least validation loss value will incorporate with the optimum weight coefficients so the validation loss will be monitored and only the best result will be saved.

5. Age-group prediction

5.1. Adience dataset

For age group estimation, Adience benchmark dataset, one of the newest databases which is designed for this kind of task, is used. Over 16000 unconstrained facial images are divided into 8 different age groups: 0-2; 4-6; 8-13; 15-20; 25-32; 38-43; 48-53; 60-.

5.2. Model using Caffe-based

Fistly, the input data images will have to go through a pre-processed phase, in which they will be randomly resized and cropped, subtracted mean values, scaled values with a scaling factor, switched the channel Blue to Red and vice versa by a function from OpenCV. As a result, we achieve a 4-dimensional main matrix with data formatted in batch size, channel, height, and width. The whole detailed process can be seen below in Figure 4.

The pre-trained caffe-based model, which used for detecting ages will go through a calibrating process. In this designed network, the system is built with three convolutional layers, each taken after by a rectified linear unit (ReLU) and followed by a pooling layer. The primary two layers are applied normalization by the local response normalization [21]. Constructed right after these convolutional blocks, a collection of 512 neurons will create a fully-connected layer and there are two such kind of layers.

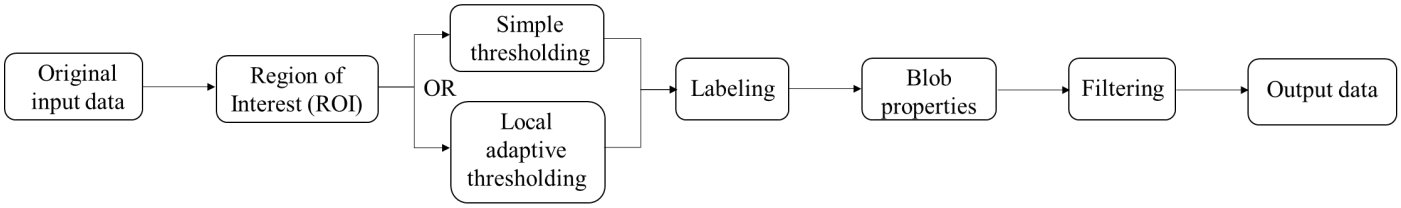


Figure 4: Data pre-processing pipeline

We tested with the method of using the designed model so as to predict age-group for novel faces. In fact, each 256×256 human's facial image will produce an assemblage of four 227×227 which are cropped from each edge as well as another one which takes the center position of the original image as the center. The network is represented by all five of these images in combination with their horizontal reflections, and the final estimation is chosen to be the average value over all of the variants. Nonetheless, there are numerous challenges with the Adience dataset's images, like occlusions or motion blur, which can cause minor misalignments and have a noticeable effect on the accuracy of the calculation. The over-sampling method is designed for this situation. By bypassing the requirement to improve alignment quality and forcusing on immediately supplying the network with many translated versions of the same face, it can correct these misalignments.

5.3. Transfer Learning on VGG Architecture for age-group prediction

In Figure 5, we illustrate our pipline for the proposed architecture that is designed and realized in this work. The convolutional layers are common and follow the pattern of those of VGG16 architecture. The fully connected layers, however are reduced because of our eight target classes. Hence, fc6, fc7 and fc8 has 512, 512 and 8 neurons, respectively.

Transfer learning: The weights of the convolutional network are loaded from VGG model trained on VGG-Face dataset. The model is then fine-tuned. Because there are so few training data, this is a vital procedure. This also provides quicker convergence and reduces overfitting for the model. The network is trained using data from the OUI-Adience Face Image Project. The main steps are outlined in Figure 5, where a pre-trained VGG-Face model based on the two datasets: IMDB-WIKI and VGG-Face CNN descriptor is loaded, froze superfluous weights, and added a customized classifier with pointed layers of trainable parameters to improve the original configuration. This customized classifier is then added to solve the multi-class classification task by training classifier layers on targeting training data, adjusting hyper-parameters, and unfreezing some additional layers if needed.

5.4. Ensemble learning

The next phase involves putting both of the aforementioned models into practice simultaneously under fresh conditions and selecting the superior model by evaluating the outcomes. The category that receives the most votes becomes the main age when classifying an image, which gives the impression that it is an election. As a result, we decide to combine the results using both the weighted majority voting and the soft voting. A summary of the process is shown in Figure 6.

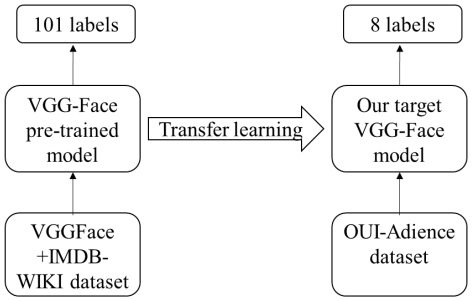


Figure 5: Overview of the Keras-based pipeline for CNN models

6. Overviewed algorithm

6.1. Accurate age prediction algorithm from front view

Because the change of face appearance is less changing and contains more information in frontal view than in profile views with the same rotation in depth, we only construct an age predictor for frontal view with a slight rotation.

When making a prediction from a new face image, the outputs will be a column tensor in which each element corresponds to a class's score. These scores, through softmax, will be normalized into probabilities and softmax classifier has a probabilistic interpretation [22] which will assign the probabilities to each class.

An input x has a probability a_i that it would fall into class i-th. For the age-set problem, y_i is a label, f_j represents the j-th element of the vector of class scores f, which connects the raw image through some rules to the class scores. By combining them, we get the following functions: the softmax function (1) and the cross-entropy loss function (2) [23]:

$$a_i = e^{f_{y_i}} \Big/ \sum_j e^{f_j} \tag{1}$$

$$L_i = -\log\left(e^{f_{y_i}} \Big/ \sum_j e^{f_j}\right) = -\log(a_i) \tag{2}$$

The result of this calculated prediction would be inaccurate and unreliable if we choose the one with highest possibility as the output, as there are a plenty of different classes in this task to estimate accurate age. In order to make this prediction become more reliable, in other words convert this task into a regression classification task, we calculate the probability of each class having an accurate age. By multiplying each output by its corresponding label and computing the sum of those values, we arrive at a summed result. Using the softmax expected value, E, as

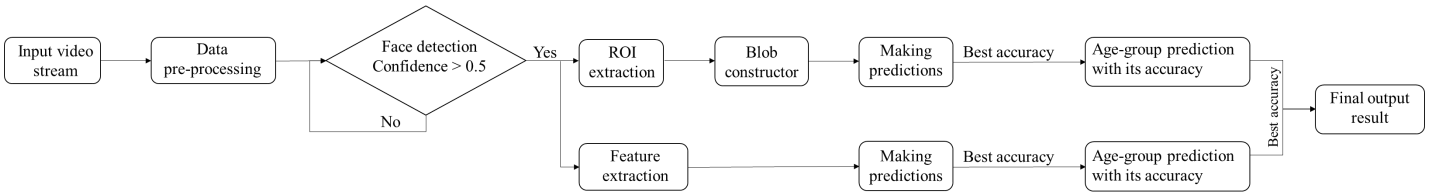


Figure 6: Our overall method

the basis for our apparent age prediction, we calculate the apparent age prediction as follows:

$$E(A) = \sum_{i=0}^{100} y_i a_i \tag{3}$$

6.2. Majority vote classifiers

Age-group predictions rely primarily on majority votes. In terms of results fusion, among all hard voting and soft voting are the most common schemes.

Hard voting, also known as weighted majority voting, is defined based on the assumption that every image x has the classification C(x) of a model i-th, whose weights w_i, with the total of $w_1,...,w_m$ is equal to 1, C is from an collection of n labels along with I(.) as an indicator function. Each classifier will vote for a class, and the majority class ŷ with most votes [24]:

$$\hat{y} = \arg\max_{j \in n} \sum_{i}^{m} w_i I\left(C_i(x) = j\right) \tag{4}$$

Soft voting, on the other hand, equalizes all classifiers to provide its proper probability for a particular target class. Supposing an image x receive the predicted result ŷ, with m is the number of models and n is the number of classes, then the probability is p_{mn}. The label which possesses the highest sum of weighted probability will represent as the output for this predicting method:

$$\hat{y} = \arg\max_{j \in n} \frac{1}{m} \sum_{i}^{m} p_{ji}\left(x\right) \tag{5}$$

6.3. Real-time application

Face detection is performed by using the Feature Base Approach method. It locates faces by extracting structural features from some facial regions like eyes, nose, mouth etc., which help distinguish between a face and many other objects and then uses them to detect a face. In our paper, we implement a feature-based approach by using OpenCV.

Instead of creating and training the model from scratch, we use one of those trained Haar Cascade models in OpenCV which are available as XML files. In developing this application, we applied our model for real-time situation. Due to the fact that a real-time video stream is an assemblage of images, face detection is then applied for each frame taken by computer. By combining this model with age prediction, we then carry out the estimation step and complete the whole model.

7. Experimental results

In Figure 7 and Figure 8 below, we have illustrated two real-time test cases where each consists of three cases which are Age-group prediction only (left), Accurate age prediction only (center) and Simultaneous age prediction (right): single face and standard condition, and the other, multiple faces and unfavorable condition. Based on this typical case, we discover that our final model gives the most accurate and stable results with the actual age of the sample compared to the two monotone models that we originally built individually. In fact, the efficiency of the age prediction model is its association algorithm, because no matter how good the dataset is, it cannot wipe out all input situations (within the collection ability of our group). Therefore, we choose the optimal development of the algorithm rather than building the perfect dataset from scratch.

7.1. Accurate age prediction

The selective dataset was split into around 16000 train instances and then 6000 or thereabouts test instances. Our final train loss and accuracy were 2.892846 and 0.239473, respectively. Based on 6000 images of the test instances, our target model's mean absolute error (MAE) reaches the value of 4.09 years.

There are some typical scenarios in Table 2 in which we test our modes in real-time conditions and then evaluate their performance. We showed the summary of 50 example cases and each accurate age was tested by 5 different people. The average accuracy is 91.97%, age difference is 2.7 years, with the MAE for age groups under 18, 20-50 and over 60 are 1.2, 2.5 and 5.3, respectively

7.2. Age-group prediction

For the VGG age-group detector, we used a quarter of the dataset for validation and the rest for training. A 100-epoch training process took approximately 6 hours to complete. Its validation and training accuracy are respectively 74.3% and 72.95%. Besides, a one-off accuracy, i.e., the prediction of one more or one less than the actual age group, is 92.11% for this model.

This method outperforms some of the previous studies on age prediction in Table 1 when it comes to performance of a caffe-based age-group detector. A high level of accuracy can be achieved through the oversampling method, which provides an average accuracy of 50.27% with MAE 5.13 age groups and an average one-off accuracy of 84.55%. Furthermore, the display of real-time accuracy with predicted age group also allows us to establish the current stability of the result.

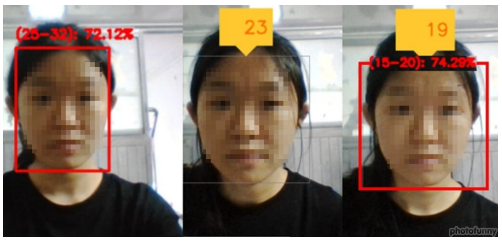


Figure 7: Age-group prediction only (left), Accurate age prediction only (center) and Simultaneous age prediction (right) in single face and standard condition of the 20-year-old girl.

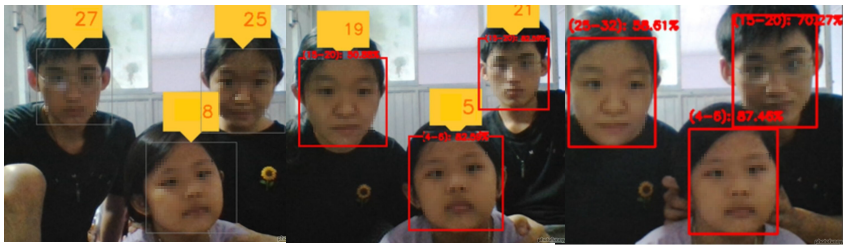


Figure 8: Age-group prediction only (left), Accurate age prediction only (center) and Simutaneous age prediction (right) in multiple faces and unfavorable condition of 3 people whose real age from left to right is 20, 5 and 21.

Table 2: Some Experimental Cases

Age-group test	8-13	8-13	15-20	15-20	25-32	25-32	38-43	38-43	60-	60-
Accurate age test	11	13	16	20	27	30	42	46	65	68
Real age	9,9,7,11, 10	12,12,12, 14, 13	14,16,14, 15,14	20,21,20, 20,20	29,29,30, 25,26	27,27,30, 32,32	37,36,37, 40,38	50,50,47, 51,51	71,71,70, 63,60	76,71,77, 63,72
Condition	Single face, Unfavora -ble	Single face, Standard	Multiple faces, Standard	Single face, Standard	Multiple faces, Standard	Single face, Unfavora ble	Multiple faces, Unfavora -ble	Multiple faces, Standard	Multiple faces, Standard	Multiple faces, Unfavora -ble
Age error, accuracy	1.8 83.64%	0.8 93.85%	1.4 91.25%	0.2 99%	2 92.59%	2 93.33%	4.4 90.24%	3.8 91.74%	4.8 92.62%	5.8 91.47%

Table 3: Confusion Matrix of Caffe Model

	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60-
0-2	**0.70**	0.26	0.03	0.00	0.01	0.00	0.00	0.00
4-6	0.15	**0.57**	0.22	0.02	0.03	0.01	0.00	0.00
8-13	0.03	0.17	**0.55**	0.08	0.14	0.02	0.00	0.01
15-20	0.01	0.02	0.15	**0.24**	0.51	0.06	0.01	0.01
25-32	0.01	0.01	0.09	0.11	**0.61**	0.15	0.02	0.01
38-43	0.01	0.01	0.07	0.06	0.46	**0.29**	0.06	0.05
48-53	0.01	0.01	0.06	0.05	0.26	0.34	**0.15**	0.13
60-	0.01	0.01	0.06	0.03	0.11	0.27	0.17	**0.36**

Table 4: Confusion Matrix of VGG Model

	0-2	4-6	8-13	15-20	25-32	38-43	48-53	60-
0-2	**0.86**	0.14	0.00	0.00	0.00	0.00	0.00	0.00
4-6	0.40	**0.83**	0.10	0.00	0.02	0.01	0.00	0.00
8-13	0.00	0.05	**0.76**	0.04	0.13	0.01	0.00	0.00
15-20	0.00	0.01	0.07	**0.33**	0.53	0.05	0.01	0.00
25-32	0.00	0.00	0.01	0.02	**0.79**	0.17	0.01	0.00
38-43	0.00	0.01	0.01	0.00	0.27	**0.64**	0.06	0.01
48-53	0.00	0.01	0.01	0.01	0.07	0.39	**0.40**	0.11
60-	0.00	0.00	0.01	0.00	0.02	0.24	0.30	**0.43**

We attribute the performance of our selective approach to the mutual support of both models. To be more specific, the performance of the caffe-based model for children under 13 is just above average, while VGG-based results are more reliable. But in return, due to confusion with other age groups, the over 60 years old perform less well in terms of efficiency. Moreover, the confusion matrixes in Tables 3 and 4 show that the caffe-based age predictor is heavily biased toward the 25-32 age group. According to the two shown tables, the VGG-Face model estimates age-group labels with greater accuracy than the other, however, the confusion rate (i.e. the difference between the main age label and its adjacent groups) of the caffe-model is smaller than that of the VGG-model in most age groups.

7.3. Comparison between two models

As a result, we find that our age prediction model is slightly influenced by external factors, e.g., real-time lighting conditions and the number of faces within the camera range especially when the boundary overlap of faces occurs. Since our free dataset contains a high number of noisy images, the quality of our prediction fluctuates between some adjacent results. In Table 1, we found that our ensemble method performs better in classification than other methods based on deep neural networks trained on public datasets such as MORPH [14-16] by focusing on the MAE index and the number of divided classes. Furthermore, our method achieved much better results than most deep neural network approaches [15-17], CNN weights are used as feature extractors, while ensemble learning algorithms and majority vote classifiers are applied to generate the most reliable results.

8. Conclusion

A technique using pre-trained fine-tuned VGG-Face weights and Caffe-based to predict accurate age and age-group was described and developed in this paper. The built model showed satisfying results with an acceptable level of reliability, besides the speed and quality of the performances. Although under some unfavorable conditions the accuracy was not as expected, the model presented promising achievement and indicated the orientation to develop and optimize the algorithm in particular and the whole model in general. The contribution of this paper is to demonstrate a combination of methods that are not overcomplicated but still highly effective. From there we want to show that in addition to investing completely in the dataset, we can take advantage of the strengths of the existing well-performed ones and devote our time to develop and optimize the algorithm.

We intend to put under scrutiny other pre-trained weights as well as datasets and thereby create more ethnically balanced datasets which will be very helpful for more efficient feature extraction in order to evaluate and select the most appropriate model with the best performance. Additionally, a completed model for information collection like face recognition can be developed, is a challenge that authors of this work plan to overcome.

Conflict of Interest

All authors have no conflicts of interest to disclose.

Acknowledgment

This research was supported in terms of equipment and expertise by the Faculty of Electrical and Electronic Engineering, Ho Chi Minh City University of Technology (HCMUT), Vietnam National University, Ho Chi Minh City (VNU-HCM).

References

[1] A. Dantcheva, P. Elia and A. Ross, "What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics," IEEE Transactions on Information Forensics and Security, **11**(3), 441-467, 2016, doi: 10.1109/TIFS.2015.2480381.

[2] Q. Zhang, D. Zhou and X. Zeng, "HeartID: A Multiresolution Convolutional Neural Network for ECG-Based Biometric Human Identification in Smart Health Applications," IEEE Access, 2017, doi: 10.1109/ACCESS.2017.2707460.

[3] R. D. Labati, A. Genovese, V. Piuri and F. Scotti, "Weight estimation from frame sequences using computational intelligence techniques," IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA) Proceedings, 2012, doi: 10.1109/CIMSA.2012.6269603.

[4] E. Eidinger, R. Enbar and T. Hassner, "Age and Gender Estimation of Unfiltered Faces," IEEE Transactions on Information Forensics and Security, **9**(12), 2170-2179, 2014, doi: 10.1109/TIFS.2014.2359646.

[5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, doi: 10.1109/CVPR.2001.990517.

[6] T. Surasak, I. Takahiro, C. Cheng, C. Wang and P. Sheng, "Histogram of oriented gradients for human detection in video," 5th International Conference on Business and Industrial Research (ICBIR), 2018, doi: 10.1109/ICBIR.2018.8391187.

[7] P. Kruachottikul, N. Cooharojananone, G. Phanomchoeng, T. Chavarnakul, K. Kovitanggoon and D. Trakulwaranont, "Deep learning-based visual defect-inspection system for reinforced concrete bridge substructure: a case of Thailand's department of highways" J Civil Struct Health Monit, 2021, doi: 10.1007/s13349-021-00490-z

[8] P. Kruachottikul, N. Cooharojananone, G. Phanomchoeng, T. Chavarnakul, K .Kovitanggoon, D. Trakulwaranont, and K. Atchariyachanvanich, "Bridge Sub Structure Defect Inspection Assistance by using Deep Learning," IEEE 10th International Conference on Awareness Science and Technology (iCAST), 2019, doi: 10.1109/ICAwST.2019.8923507.

[9] G. Guo, Y. Fu, C. R. Dyer and T. S. Huang, "Image-Based Human Age Estimation by Manifold Learning and Locally Adjusted Robust Regression," IEEE Transactions on Image Processing, **17**(7), 1178-1188, 2008, doi: 10.1109/TIP.2008.924280.

[10] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," CVPR, 2011, doi: 10.1109/CVPR.2011.5995404.

[11] X. Geng, C. Yin and Z. -H. Zhou, "Facial Age Estimation by Learning from Label Distributions," IEEE Transactions on Pattern Analysis and Machine Intelligence, **35**(10), 2401-2412, 2013, doi: 10.1109/TPAMI.2013.51.

[12] F. Li, "Linear classification: Support Vector Machine, Softmax," 2021. [Online]. Available: https://cs231n.github.io/linear-classify/. [Accessed 9 June 2021].

[13] I. Huerta, C. Fernández, A. Prati, " Facial Age Estimation Through the Fusion of Texture and Local Appearance Descriptors," Computer Vision - ECCV 2014 Workshops, doi: 10.1007/978-3-319-16181-5_51

[14] A. Lanitis, C. J. Taylor and T. F. Cootes, "Toward automatic simulation of aging effects on face images," IEEE Transactions on Pattern Analysis and Machine Intelligence, **24**(4), 442-455, April 2002, doi: 10.1109/34.993553.

[15] H. Takimoto, Y. Mitsukura, M. Fukumi, and N. Akamatsu, "Robust gender and age estimation under varying facial pose." Electronics and Communications, 2018, doi: 10.1002/ecj.10125

[16] X. Geng, Z. -H. Zhou and K. Smith-Miles, "Automatic Age Estimation Based on Facial Aging Patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence, **29**(12), 2234-2240, 2007, doi: 10.1109/TPAMI.2007.70733.

[17] G. Guo and G. Mu, "Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression," CVPR, 2011, doi: 10.1109/CVPR.2011.5995404.

[18] Z Z. Hu, Y. Wen, J. Wang, M. Wang, R. Hong and S. Yan, "Facial Age Estimation With Age Difference," IEEE Transactions on Image Processing, **26**(7), 3087-3097, 2017, doi: 10.1109/TIP.2016.2633868.

[19] A. Kumar, "Keras – Categorical Cross Entropy Loss Function," 2020. [Online]. Available: https://vitalflux.com/keras-categorical-cross-entropy-loss-function/. [Accessed 2 June 2021].

[20] Gilbert Tanner, "Introduction to Deep Learning with Keras," 2019. [Online]. Available: https://towardsdatascience.com/introduction-to-deep-learning-with-keras-17c09e4f0eb2. [Accessed 10 June 2021]..

[21] K. Ricanek and T. Tesafaye, "MORPH: a longitudinal image database of normal adult age-progression," 7th International Conference on Automatic Face and Gesture Recognition (FGR06), 2006, doi: 10.1109/FGR.2006.78.

[22] "Pandas DataFrame," 2019. [Online]. Available: https://www.geeksforgeeks.org/python-pandas-dataframe/. [Accessed 2 June 2021].

[23] F.-F. Li, "Linear classification: Support Vector Machine, Softmax,," 2021. [Online]. Available: https://cs231n.github.io/linear-classify/. [Accessed 9 June 2021].

[24] James, Gareth Michael, "Majority vote classifiers: theory and applications," Stanford University, 1998.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 6, 269-282 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Mobility Intelligence: Machine Learning Methods for Received Signal Strength Indicator-based Passive Outdoor Localization

Fanchen Bao[1], Stepan Mazokha[1], Jason O. Hallstrom[*,1]

[1]*Florida Atlantic University, Institute for Sensing and Embedded Network Systems Engineering (I-SENSE), Boca Raton, FL, 33431, USA*

ARTICLE INFO

Article history:
Received: 29 August, 2022
Accepted: 27 November, 2022
Online: 20 December, 2022

Keywords:
Passive outdoor localization
RSSI
WiFi probe requests
Mobility intelligence

ABSTRACT

Knowledge of pedestrian and vehicle movement patterns can provide valuable insights for city planning. Such knowledge can be acquired via passive outdoor localization of WiFi-enabled devices using measurements of Received Signal Strength Indicator (RSSI) from WiFi probe requests. In this paper, which is an extension of the work initially presented in WiMob 2021, we continue the work on the mobility intelligence system (MobIntel) and study two broad approaches to tackle the problem of RSSI-based passive outdoor localization. One approach concerns multilateration and fingerprinting, both adapted from traditional active localization methods. For fingerprinting, we also show flaws in the previously reported area-under-the-curve method. The second approach uses machine learning, including machine learning-boosted multilateration, reference point classification, and coordinate regression. The localization performance of the two approaches is compared, and the machine learning methods consistently outperform the adapted traditional methods. This indicates that machine learning methods are promising tools for RSSI-based passive outdoor localization.

1 Introduction

City planning initiatives can benefit from insights into mobility patterns of pedestrians and vehicles. Given the prevalence of WiFi-enabled devices carried by individuals and their continuous emission of WiFi probe requests, such mobility patterns can be revealed by localization of the devices based on measurements of WiFi probe requests via the recorded received signal strength indicator (RSSI). **Active** and **passive** localization are the two standard modes of RSSI-based localization [1]. Active localization occurs on the target device with the help of a custom app that may or may not require cooperation (e.g., data exchange); it answers the question "Where am I?" Passive localization occurs without the device's knowledge in an anonymous manner; it answers the question "Where are you?" Due to the apparent difficulty of large-scale mobile app adoption for localization purposes, passive localization is the more realistic choice.

High variability in RSSI measurements due to fading (path-loss), shadowing (temporary obstruction between a sender and a receiver), and interference (overlap of WiFi channels) is a significant challenge in RSSI-based localization [2]. To mitigate RSSI variability, some active localization methods leverage device cooperation to obtain additional location-sensitive information, such as signal angle of arrival [3, 4], round-trip time [5], device orientation [6], and device prior location [7]. However, passive localization cannot benefit from these methods, as it does not have the luxury of device cooperation.

Other active localization methods achieve good results without device cooperation [8]–[13]. Since the operations of these methods share similarities with passive localization, they might be transferable. However, fundamental differences exist between the two. For instance, missing data are unlikely in active localization because the probing signal is emitted by the access points in large numbers. In passive localization, the probing signal is initiated by the target device in an uncontrolled and sporadic manner. Consequently, the active localization methods may have data requirements unachievable in passive localization.

Previous work has also explored methods tailored to passive localization. However, they typically suit applications on a very small (indoor) or very large scale (entire city), making them inadequate for applications on a medium scale (within the span of a city street) (see Section 2.2 for details).

Therefore, investigation into methods suitable for RSSI-based passive localization in urban areas at the scale of a city street is necessary. The investigation can be divided into two parts. One focuses on the feasibility of adapting the methods used in cooperation-free active localization (hereinafter referred to as "traditional active localization methods") to passive localization; the other focuses on new approaches not fully explored by previous research on passive localization.

To facilitate the investigation, we have developed a privacy-

[*]Corresponding Author: Jason O. Hallstrom, Email: jhallstrom@fau.edu

centric mobility intelligence system (MobIntel) consisting of sensors and cloud infrastructure to collect and analyze RSSI measurements from anonymized WiFi probe requests. MobIntel has been tested in downtown West Palm Beach, Florida, to count the number of WiFi-enabled devices in a sensor's visible range and to calculate the percentage of devices that have persisted over time [14].

In this paper, which is an extension of the work initially presented in WiMob 2021 [15], we discuss a significant augmentation to the analytical facilities of MobIntel, exploring methods to conduct passive localization in a controlled environment, with the goal of identifying methods that are performant at the scale of a city street (1.4 km). The specific contributions of the paper include:

1. We conduct a systematic performance evaluation of two adaptations of traditional active localization methods—path-loss-based multilateration and fingerprinting—to support passive localization. We examine performance impacts under varying training and testing data sets, including (1) training and testing on individual RSSI measurements, (2) training and testing on averaged RSSI measurements, and (3) training on averaged, but testing on individual RSSI measurements. We also compare the performance of fingerprinting across three implementation methods, including vector-based Euclidean distance, Gaussian-based area-under-the-curve, and Gaussian-based tail probability.

2. For Gaussian-based fingerprinting in particular, we discover that the previously reported area-under-the-curve method, based on the arithmetic mean of the area-under-the-curve in a Gaussian distribution [7], does not perform well when there is large mismatch between the unknown RSSI measurements and the fingerprint of a reference point. We propose an alternative method, which uses the Gaussian tail probability as the likelihood of the unknown RSSI measurements originating from a particular reference point. We show that our proposed method achieves better localization performance.

3. We carry out a systematic evaluation of the accuracy impact of integrating machine learning algorithms—multi-layer perceptron, support vector machine, and K-nearest neighbors—with traditional active localization methods. We demonstrate that the machine learning-based methods outperform their traditional counterparts, regardless of which model is incorporated, or which training or testing set is used.

This paper is organized as follows. Section 2 provides a brief overview of related work in RSSI-based active and passive localization. Section 3 describes how we adapt the localization method from a traditional active context to a passive context and how we incorporate machine learning methods to boost performance. Section 4 presents details of data collection, including experimental design, hardware configuration, and data collection procedures. Section 5 discusses the preparation of various data sets for analysis. Section 6 reports the localization performance on methods adapted from traditional multilateration and fingerprinting. Section 7 describes the model training process for the machine learning methods and reports their localization performance. Section 8 compares localization performance among the best models of each method investigated. Finally, Section 9 concludes the paper by presenting limitations and directions for future work.

2 Related Work

Prior work in RSSI-based localization has been conducted in active (localization done on the device) and passive (device has no knowledge of localization) settings. Here, a brief review of related work in both areas is presented. Emphasis is given considering applicability in urban areas.

2.1 Active Localization

Multilateration is a common method for active localization. It involves estimating distances from the target device to all nearby access points (APs) using an RSSI-distance model, and then applying linear least squares regression to pinpoint the device's location based on geometric analysis. Most work in multilateration focuses on building a better or simpler RSSI-distance model. In [12], the authors use the RSSI path-loss model as a base, treat the path-loss exponent as a random variable, and apply maximum likelihood estimation to derive a closed-form distance estimator. In [13], the authors use linear regression to fit RSSI-distance data to a generic path-loss model without estimating the path-loss exponent.

Relying solely on RSSI is usually not sufficient to deliver high accuracy. Thus, some studies incorporate additional location-sensitive information to mitigate RSSI variability and improve model performance. For example, in [5], the authors leverage the newly added fine timing measurement frame in WiFi protocols to compute the RF signal round-trip time between device and AP. Round-trip time is then fused with RSSI using Kalman filter to form a new parameter that provides better distance estimation than RSSI alone.

Other studies, as detailed in the review [4], drop RSSI completely and use signal angle-of-arrival, time-of-arrival, time-difference-of-arrival, and other more consistent measures to estimate device-to-AP distance. These techniques yield higher accuracy than RSSI-based multilateration.

The other commonly used method in active localization is fingerprinting. Fingerprinting has offline and online stages. In the offline stage, a test device roams around a target area surrounded by numerous APs. The APs emit RF signals constantly, which are captured by the test device at various reference points along its path. The RSSI measurements obtained at each reference point form a vector, or fingerprint, which describes the RSSI pattern representative of the reference point in the AP network. Together, the offline RSSI fingerprints form a **radio map** that is used during the online stage to match an unknown vector of RSSI measurements to the most likely reference point.

Much research has focused on the offline stage to improve the quality of fingerprints in the radio map. A common strategy is to incorporate additional location-sensitive information, such as azimuth data collected from a device's magnetometer [6] and a device's prior location and speed data [7, 16], such that an enhanced radio map can be created. These studies have shown that an enhanced radio map yields better localization performance.

One challenge in fingerprinting is the scalability of creating a high quality radio map outdoors. It requires data collection at a large number of reference points and deployment of more APs, both of which are daunting tasks. To resolve these issues, the authors of [17] propose crowdsourcing to build radio maps at large scales. Incentivized volunteers obtain RSSI measurements from public WiFi hotspots and upload the RSSI and GPS data to a server for radio map construction. Their work demonstrates the robustness of crowdsourced data for localization purposes, yet falls short of implementing an actual crowdsourcing solution.

Since passive localization is largely independent of the device, any method requiring additional location-sensitive data from the device cannot be readily adapted. Crowdsourcing might be adaptable because it does not require device cooperation. However, crowdsourcing in passive localization is different from active localization, because the participants are APs instead of mobile devices. In theory, these APs can sniff WiFi probe requests, record their RSSI measurements, and upload the recorded data. Yet, in practice, configuring these APs to serve as crowdsourcing endpoints, which are scattered around private businesses and homes, could be more challenging than simply generating the radio map manually. Thus, crowdsourcing is also not readily applicable to passive localization. In summary, the only methods that are readily adaptable from traditional active localization to passive localization are purely RSSI-based multilateration and fingerprinting.

2.2 Passive Localization

The authors in [18] were the first to propose the concept of device-free passive localization. It is based on the idea that an entity can be detected and tracked in a network of APs that are constantly sending beacons. Due to shadowing effects on the beacon's signal, it is possible to construct a radio map based on RSSI disturbances, and use fingerprinting to localize the entity even if it has no radio frequency capability. In [19], the authors push this idea further by creating a high density AP network, achieving not only localization, but radio tomographic imaging for one or two entities in the AP network. In [20], the authors present device-free localization models based on relevance vector machines and show good performance with single entity tracking in a cluttered environment.

Despite the progress in device-free passive localization, there are two major limitations that preclude its use in an urban environment. First, creating a sufficiently sensitive AP network requires a high density of APs ([19] use 28 APs and [20] use 24 APs, both in a 6 m by 6 m area); this is not scalable to city streets. Second, the number of devices localizable within the AP network is limited because if too many devices are present, the RSSI disturbances would be too chaotic to characterize. This limitation means that the device-free approach cannot handle crowded city streets.

The other approach in passive localization involves the target device emitting a signal spontaneously, such as a WiFi probe request, but without actually communicating with the APs. Probe request sensors are typically deployed at strategic positions to capture and process probe request data. In [21], the authores use this approach to localize static devices in an intersection. A target device is localized to the same position as the sensor that has captured the device's probe request. If multiple sensors detect a device, the position of the sensor receiving the highest RSSI measurements is used. However, this method has low resolution if the number of sensors deployed is small (i.e., many devices would be localized to the same sensor if that is the only sensor nearby). This can be tolerated if the resolution requirement is low—for instance, at the scale of an entire city, a resolution level of intersections is acceptable—but it is not sufficient to reconstruct movement patterns on a street. One can improve resolution by deploying more sensors at smaller intervals, but this is not scalable.

In [22], the authors also use a qualitative method, but they improve resolution by deploying GPS-enabled mobile sensors carried by volunteers, who roam the monitored area. Since the sensors are mobile, localization is not static. If a mobile sensor and a target device have similar movement patterns, localization resolution can be high. However, if the mobile sensor moves in the opposite direction as the target device, localization resolution is compromised. Another drawback is that continuous human-based monitoring is infeasible over a long period of time. The paper suggests using drones or robots as alternatives, yet the scalability issue remains.

In [1], the authors propose a quantitative method to localize a stationary device using a single mobile sensor, without the need for a radio map. Their method requires the sensor to move either in a straight or right-angled path close to the target device's potential location, taking RSSI measurements from the device's WiFi probe requests at distinct points along the path. Using the RSSI path-loss model, they estimate the distance from the measurement points to the target device. Using these distance measurements, along with the known distance between the consecutive measurement points, they are able to estimate the location of the target device. The advantages of this method include low cost and ease of implementation. However, the method is limited by the requirement that the target device be stationary when the sensor is moving on the data collection path. It also suffers from the scalability issue associated with mobile sensors, if multiple targets are to be localized simultaneously.

Overall, the methods described above for passive localization fall short for the requirements of monitoring mobility patterns on city streets. The device-free methods are appropriate only at small scales, whereas the qualitative methods are only sensitive enough for large scales with low resolution requirements. The methods based on mobile sensors, though capable of delivering higher resolutions, present scalability issues in urban areas. Therefore, to achieve passive localization on city streets, a different approach is needed that leverages device participation (but not cooperation), uses static sensors, and is quantitative in nature.

3 Our Methods

This section describes two general localization approaches that satisfy the abovementioned requirements. The first approach involves adapting traditional active localization methods, including multilateration and fingerprinting. The second leverages machine learning, including machine learning-boosted multilateration, classification of reference points, and regression of reference points' coordinates.

3.1 Multilateration

The concept of multilateration in active localization is described in Section 2.1 and a visual representation is available in [23]. Since

this method is independent of device cooperation, its adaptation to passive localization is straightforward. Instead of using the RSSI measurements obtained by the target device from nearby APs, it relies on RSSI measurements acquired by the sensors from the target device.

The RSSI measurements must be converted to distance values between the target device and each sensor. This can be achieved using the basic RSSI path-loss model shown in (1), where R is the RSSI measurement, d is the distance (non-zero) between the sensor and the target, A is the RSSI measured at a reference distance from the target (e.g., 1 m), and n is the mean, environment-specific path-loss index that describes the speed of signal loss with increasing distance [24]. In a stable environment, A and n are constants. Therefore, the path-loss model can be treated as a negative linear relation between R and $\log(d)$, which can be fitted using linear regression to produce an RSSI path-loss model.

$$R = A - 10n \log(d) \tag{1}$$

After estimating device-to-sensor distances, we use least squares optimization to search for the coordinates $(\hat{x}, \hat{y})$ that minimize the residue in (2). In the equation, d_i is the estimated distance between the target device and the i-th sensor; (x_i, y_i) represents the coordinates of the i-th sensor; and $(\hat{x}, \hat{y})$ represents the estimated coordinates of the target device.

$$\text{Residual} = \sum_{i=1}^{N} \left(d_i - \sqrt{(x_i - \hat{x})^2 + (y_i - \hat{y})^2} \right)^2 \tag{2}$$

3.2 Fingerprinting

The concept of fingerprinting in active localization is described in Section 2.2. Similar to multilateration, the adaptation from active to passive localization only entails changing the manner in which RSSI is obtained, i.e., instead of measuring RSSI on the target device from signals sent by nearby APs, we measure RSSI on the nearby sensors from the signal sent by the target device.

We employ two methods to establish RSSI fingerprints. The first involves *vector-based fingerprinting*. We use RSSI vectors as fingerprints, in which each value is the mean of all RSSI measurements obtained by a specific sensor at a given reference point [11]. For example, given N sensors S_1, S_2, ..., S_N and M reference points P_1, P_2, ..., P_M, we denote R_{ij1}, R_{ij2}, R_{ij3}, ..., as the RSSI measurements emitted from reference point P_i $(1 \le i \le M)$ and captured by sensor S_j $(1 \le j \le N)$. Let μ_{ij} be the mean of R_{ij1}, R_{ij2}, R_{ij3}, ... The vector-based fingerprint of P_i can be written as $F_i^V = [\mu_{i1}, \mu_{i2}, ..., \mu_{iN}]$.

When a target device generates a vector of RSSI measurements $U = [R'_1, R'_2, ..., R'_N]$ at an unknown location, the penalty of matching U to F_i^V can be approximated by their Euclidean distance. The smaller the distance, the more likely the target device originates from reference point P_i.

The second method is *Gaussian-based fingerprinting*. It uses a set of Gaussian distributions as RSSI fingerprints, where each distribution corresponds to the RSSI measurements obtained by a specific sensor at a given reference point [7]. Extending the example above, *Gaussian-based fingerprinting* computes both the mean (μ_{ij}) and standard deviation (σ_{ij}) of RSSI measurements collected by sensor S_j with regard to reference point P_i. The Gaussian distribution is expressed in (3).

$$g_{ij}(x) = \frac{1}{\sigma_{ij}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu_{ij}}{\sigma_{ij}})^2} \tag{3}$$

Considering a reference point P_i, and N sensors S_1, S_2, ..., S_N, P_i's Gaussian-based fingerprint can be expressed as a set of distributions $F_i^G = \{g_{i1}(x), g_{i2}(x), ..., g_{iN}(x)\}$.

Given a set of RSSI measurements $U = [R'_1, R'_2, ..., R'_N]$ from the target device, the penalty of matching U to F_i^G can be estimated in two ways. The first is the area-under-the-curve (AUC) method, proposed by [7]. It computes the AUC (denoted as A_{ij}) for each R'_j under the Gaussian distribution $g_{ij}(x)$, as shown in (4). The penalty of matching U to F_i^G is approximated by the mean of all A_{ij}. The smaller the mean over A_{ij}, the more likely the target device originates from reference point P_i.

$$A_{ij} = \left| \int_{\mu_{ij}}^{R'_j} g_{ij}(x) dx \right| \tag{4}$$

Our initial localization results show that *Gaussian-based fingerprinting* with the AUC method does not perform well, especially when there is significant misalignment between some R'_j and μ_{ij}. This is most likely due to the flaw of the penalty structure in the AUC method (see Section 6.2 for details). To offer a better alternative, we propose the **tail method**, which uses tail probabilities to represent the likelihood that an individual R'_j originates from $g_{ij}(x)$. The product of all such individual likelihoods is the overall likelihood that U matches F_i^G. Concretely, let L_{ij} be a likelihood estimator, which is the smaller of the two AUCs from the cumulative density functions extending from R'_j to either tail of the distribution. Figure 1 illustrates the definition of L_{ij}. In the figure, R'_j is placed in a Gaussian distribution $g_{ij}(x)$, with mean μ_{ij} and standard deviation σ_{ij}. L_{ij} is the AUC from R'_j to the right tail. The larger the value of L_{ij}, or the closer R'_j is to μ_{ij}, the more likely R'_j originates from $g_{ij}(x)$.

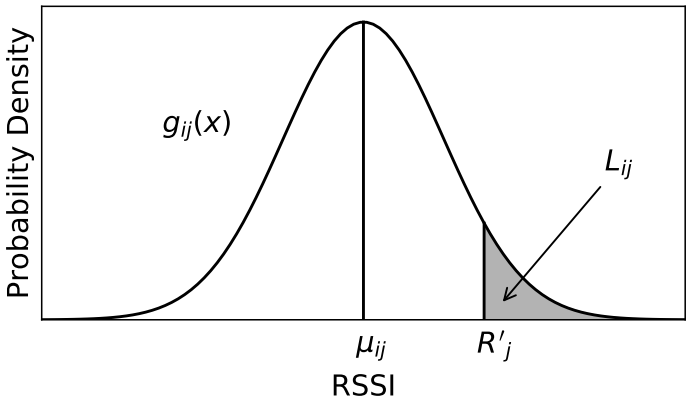


Figure 1: Tail Probability of R'_j in a Gaussian Distribution $g_{ij}(x)$

L_{ij} can be written as (5). The penalty of matching U to F_i^G can be approximated by the product over all L_{ij}, which is then converted to the sum of the negative logarithm of all L_{ij} to facilitate

computation. The smaller the sum, the more likely the target device originates from reference point P_i.

$$L_{ij} = \int_0^{\min(R'_j,\ 2\mu_{ij}-R'_j)} g_{ij}(x)dx \tag{5}$$

Once the penalty is obtained from the target device to all reference points using either the vector- or Gaussian-based method, K-nearest neighbors is used to find the top k best-match reference points. The exact value of k is determined empirically via a validation set. The average location of the top-k reference points is the predicted location of the target device.

3.3 Machine Learning-boosted Multilateration

We use the RSSI path-loss model in (1) to estimate distance based on RSSI measurement. However, (1) does not capture all the error terms [10]. While it is possible to create more complex path-loss models with add-on error terms, we can train a standard machine learning model to "learn" the relationship between RSSI measurements and distances, with error implicitly included. Machine learning may capture relationships not explicitly described by a physical model. Thus, it is likely that a machine learning model will perform better than the idealized physical model in (1). We use three standard machine learning algorithms—multi-layer perceptron (MLP), support vector machine (SVM), and k-nearest neighbors (K-NN)—to train regression models from RSSI measurements and distance at each sensor. Each model is then used to conduct multilateration, as described in Section 3.1.

3.4 Machine Learning Classification of Reference Points

In fingerprinting, a set of RSSI measurements are passed to a fingerprinting algorithm as inputs, and the most likely reference point is returned as the localization output. This process is similar to a machine learning classification model, where input features are RSSI measurements, and output labels are reference points. The difference between the two is that the former uses pre-specified rules to determine how RSSI measurements match reference points, whereas the matching is "learned" from training data in the latter. We apply the same machine learning algorithms (MLP, SVM, and K-NN) to train classifiers that map a set of RSSI measurements to the best-match reference point.

3.5 Machine Learning Regression of Coordinates

In [9], the authors offer a simple approach to performing active localization. They train a machine learning regressor that takes RSSI measurements and propagation delays of transmitted signals as input, and directly predicts the GPS coordinates of a target device. Despite not having access to the propagation delay data, we can still adapt this approach to passive localization using the RSSI measurements alone. We follow the same setup as in Section 3.4, but instead of training a classifier, we train two regressors that take RSSI measurements as input and predict a reference point's x- and y-coordinate, respectively. With the two regressors, we can directly estimate a target device's location based on the RSSI measurements.

4 Data Collection

To evaluate the passive localization methods discussed in Section 3, RSSI measurements from mock probe requests, along with the coordinates of reference points, were collected in a semi-controlled environment. In the following sections, we present the experimental setup, experimental design, hardware details (sensor and emitter), field preparation, and data collection procedure.

4.1 Experimental Design

Figure 2(A) shows the layout of the testbed. Each cell is a 1 m by 1 m square, making the entire field 15 m by 15 m. Four sensors (red dots) s_1, s_2, s_3, and s_4 are placed at coordinates (0, 15), (15, 15), (0, 0), and (15, 0), respectively. During data collection, an emitter is placed on each data collection point (blue dots) and allowed to send approximately one-thousand mock probe requests. Each contains a Media Access Control (MAC) address that is traceable to a specific data collection point, distinguishable from a genuine probe request, and unique. After a mock probe request is captured by a sensor, the data is uploaded and stored in a database hosted on Amazon Web Services (AWS). Since four sensors are present in the testbed, at most four RSSI measurements are collected per probe request.

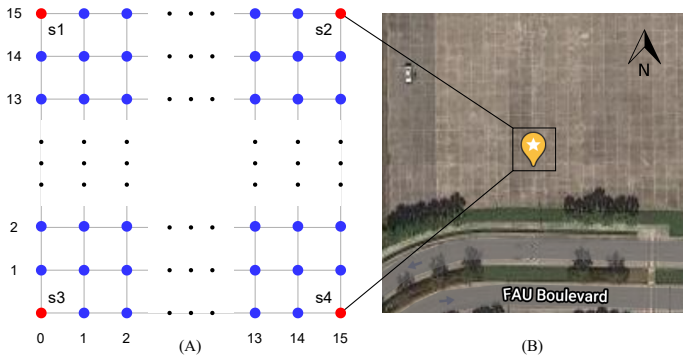


Figure 2: Experimental design and satellite view of the testbed

4.2 Hardware - Sensor

We used the same sensor described in [14]. It captures WiFi probe requests from the environment, removes duplicate MAC addresses at each 30-second window, and uploads aggregated data to AWS. In this paper, we added a new feature to differentiate mock probe requests from genuine probe requests, which allows the sensor to upload mock probe requests to a database dedicated to the experiment. The uploaded data include MAC address, time of capture, and RSSI measurement.

4.3 Hardware - Emitter

We used the same stress testing device described in [14] as the emitter of mock probe requests. It emits WiFi probe requests with a custom MAC address, on a specified channel, at a specified rate.

MAC address customization ensures traceability. The first four letters of the address are synced to the time of emission (timezone EST) in the format "HH:MM". Since the timestamp at each data

collection point is recorded, a time-stamped MAC address encodes the location where the mock probe request was emitted. To ensure mock probe requests are distinguishable from genuine probe requests, the fifth and sixth positions of the address are fixed as "00". A "HH:MM:00" prefix guarantees few, if any, collisions with genuine probe requests. Finally, to ensure that mock probe requests are unique among themselves, the remaining six positions are filled with random hexadecimals. As an example, a mock probe request emitted at 3:04 PM EST would have a MAC address similar to "15:04:00:1A:2B:3C".

The MAC address customization is only traceable at the granularity of one minute. If two mock probe requests are emitted from different data collection points within the same minute of an hour, their locations cannot be distinguished (i.e. the first four letters of their MAC addresses would be identical). To avoid such complications, the emitter remains at each data collection point for at least a full minute before moving to the next. In the experiment, the emitter uses *cronjobs* to send approximately 1,000 mock probe requests in the first 50 seconds of each minute, and zero in the remaining 10.

4.4 Testbed Preparation

An empty outdoor area (GPS: 26.381488, -80.099640) with no nearby obstacles was used as the testbed to avoid complications in RSSI measurement. Figure 2(B) shows an overhead view of the field. Figure 3 provides a detailed view after the 15 m by 15 m grid has been prepared with spray chalk. The southwestern corner is chosen as the origin point of the field, with coordinates $(0, 0)$.

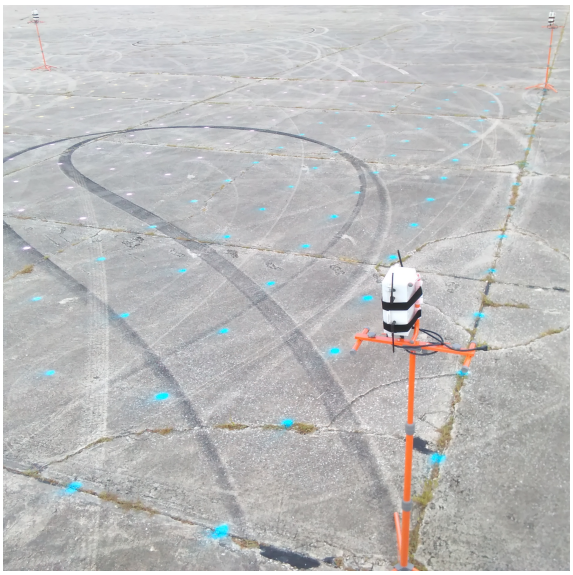


Figure 3: Mounted sensors on the testbed with grid drawn

The sensors were mounted on tripods, with an elevation of approximately 1.8 m above the ground. The four tripods were positioned at s_1, s_2, s_3, and s_4, as illustrated in Figure 2. Three of the tripods are visible in Figure 3.

4.5 Data Collection Procedure

Data are collected by placing the emitter on each data collection point. The emitter is mounted on the same type of tripod as the sensor, but with an elevation of only 1 m. This mimics the height at which a WiFi-enabled device is likely to be carried by a pedestrian. During the first 50 seconds of each minute, the emitter is left alone on the data collection point, without anyone present within the boundaries of the testbed. In the remaining 10 seconds, a researcher enters the field and moves the emitter to the next point. This setup ensures that signal transmissions are not affected by transient obstructions. At each data collection point, the coordinates and the current timestamp are recorded. This procedure is repeated until all data collection points are visited.

Data collection for this manuscript was conducted from 15:00 to 20:00 on 2020-07-16.

5 Data Preparation

The raw probe request data contains 221,553 observations, one per mock probe request. Each observation contains five parameters, corresponding to the RSSI measurement (one per sensor) and timestamp. Observations with missing measurements are discarded. The resulting **cleaned data set** contains 171,537 observations. To further reduce the computation cost of model training, and to balance the number of observations across data collection points, 100 observations are randomly selected from each data collection point to form the **sampled data set**.

The sampled data set is one of the two data sets used in this paper. It contains 25,200 rows and 11 parameters. Each row corresponds to an individual mock probe request. The parameters correspond to four RSSI measurements (R_1 through R_4), coordinates of the data collection points (x and y), distances to each sensor (d_1 to d_4), and unique labels of the data collection points (*label*). Table 1 shows the structure of the sampled data set.

Table 1: Structure of the sampled data set

$R_1 ... R_4$	x	y	$d_1 ... d_4$	label
-44 ... -41	0.0	1.0	14.0 ... 15.0	16
-43 ... -41	0.0	2.0	13.0 ... 15.1	32
...	...	...	...	...

The other data set used in this paper is the **mean data set**, derived from the cleaned data set by taking the mean and standard deviation of RSSI measurements from each sensor at each data collection point. The mean data set contains 252 rows and 15 parameters. Each row corresponds to a data collection point. The parameters correspond to four mean RSSI measurements (μ_1 to μ_4), coordinates of the data collection points (x and y), distances to each sensor (d_1 to d_4), four RSSI standard deviations (σ_1 to σ_4), and unique labels of the data collection points (*label*). Table 2 shows the structure of the mean data set.

Table 2: Structure of the mean data set

μ_1 ... μ_4	x	y	d_1 ... d_4	σ_1 ... σ_4	label
-44.3, ..., -42.8	0.0	1.0	14.0 ... 15.0	0.93 ... 1.15	16
-44.1, ..., -40.7	0.0	2.0	13.0 ... 15.1	1.60 ... 0.64	32
...	...	...	...	...	...

After random shuffling, the sampled data set is split into a **sampled training set** (80%, 20,160 observations) and a **sampled testing set** (20%, 5,040 observations) using `scikit-learn` [25]. The mean data set cannot undergo the same splitting procedure due to having only one observation per data collection point. Hence, the entire mean data set is treated as the **mean training set** (252 observations). Finally, the mean and standard deviation of RSSI measurements by each sensor at each data collection point from the sampled testing set form the **mean testing set** (252 observations).

The training sets are used exclusively for model training and validation. In particular, the sampled training set is used in all methods except fingerprinting; the models thus trained are called **sampled models**. The mean training set is used in all methods; the models thus trained are called **mean models**. The difference between sampled and mean models is that the former is tuned using the individual observations in a training set, whereas the latter is tuned by their mean values. It is worth noting that averaging RSSI measurements seems to have both positive and negative impacts on model performance, as it simultaneously reduces random error [2] and training size. It is thus interesting to study how the performance of the mean model compares to that of the sampled model.

All data preparation procedures were conducted using the `pandas` Python library [26]. Figure 4 illustrates the relationships among the derived data sets, where their names, lineage, and shapes (rows × columns) are shown.

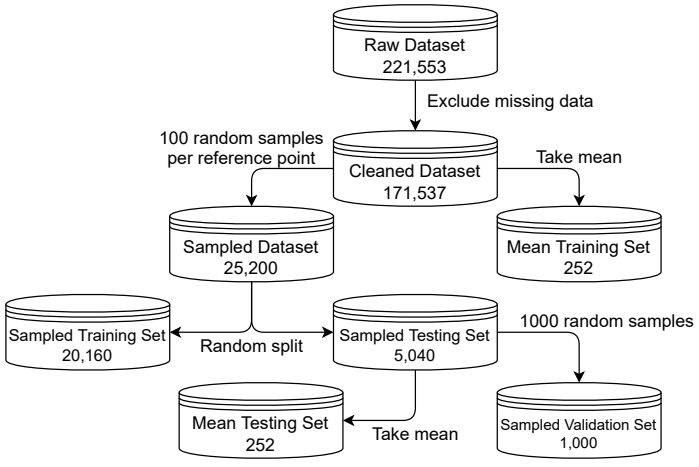


Figure 4: Relationship among all data sets

6 Traditional Methods

In this section, we present the adapted multilateration and fingerprinting methods for passive localization and consider their performance. Here, localization performance is visualized by the cumulative distribution function (CDF) curve of localization error. Localization error is defined as the Euclidean distance between the predicted and the true coordinates of a test observation. The larger the area-under-the-curve of a CDF curve of localization error, i.e., the closer the curve gets to the top and left corner, the better the localization performance.

6.1 Multilateration

Two RSSI path-loss models, one using the sampled training set and one using the mean training set, are created by linear regression according to (1) for each sensor. The performance of the two models, evaluated on the sampled testing set, is shown in Table 3. The R^2 score describes the amount of variance explainable by the path-loss model, and the mean absolute error (MAE) describes the error between the estimated and actual distance. Within each sensor, the two models have almost identical performance. This means that averaging RSSI measurements does not alter model performance. Across the models of different sensors, s_4 has the best performance, whereas s_1 has the worst. This may be explained by a consistent source of interference from the northwest corner of the testbed. Since s_1 is closest to the northwest corner, it was affected the most. s_4 is furthest from the interference source, so it was affected the least. However, the validity of this explanation cannot be verified until the same performance pattern repeats over time in future research.

Table 3: Metrics of the path-loss model

Sensor	Sampled		Mean	
	R^2	MAE	R^2	MAE
s_1	0.61	4.62	0.61	4.63
s_2	0.65	4.06	0.65	4.07
s_3	0.61	4.03	0.61	4.03
s_4	0.71	3.44	0.71	3.46

Two multilateration models are generated, one based on the sampled RSSI path-loss model and the other based on the mean RSSI path-loss model. We first convert RSSI measurements to distances, and then pass the distance data to `scipy`'s least squares optimization API to estimate the target device's coordinates [27]. Figure 5(A) demonstrates the localization performance of the traditional multilateration method. The x-axis denotes localization errors, and the y-axis denotes the probability that a model's localization error is smaller or equal to a selected x-axis value. The blue (sampled-sampled), orange (mean-sampled), and green (mean-mean) curves represent the localization performance under three scenarios: the sampled model tested on the sampled testing set, the mean model tested on the sampled testing set, and the mean model tested on the mean testing set, respectively.

The sampled-sampled curve does not show on the figure because it overlaps with the mean-sampled curve. This is expected because the sampled and the mean RSSI path-loss models have similar distance estimation performance. The mean-mean scenario performs slightly better than the other two. However, this is likely an artifact because the mean training and testing sets are very similar. Recall that the cleaned data set is the source of the mean training set, and

the sampled testing set is the source of the mean testing set (Figure 4). Since the sampled testing set is a subset of the cleaned data set, the mean testing set shares information with the mean training set. The similarity between the mean testing and training sets inflates the measured performance of the mean-mean scenario.

6.2 Fingerprinting

The radio map for vector-based fingerprinting is a 252×4 matrix prepared from the mean training set, where each row represents a reference point, and each column represents the mean RSSI measurements from a sensor. As mentioned in Section 3.2, we empirically determine k, the number of top-match reference points that produces the best coordinate estimation. A series of fingerprinting models parameterized with different values of k are used to perform localization on a validation set. The k leading to the least validation error is selected. Usually, the validation set is withheld from the training set. However in fingerprinting, the mean training set, which is also the radio map, has no redundancy to supply a validation set. Therefore, we set aside 1,000 observations from the sampled testing set as the validation set (Figure 4). The observations in the validation set are not included in evaluating model performance. Validation error is defined as the root mean squared localization error, following [11].

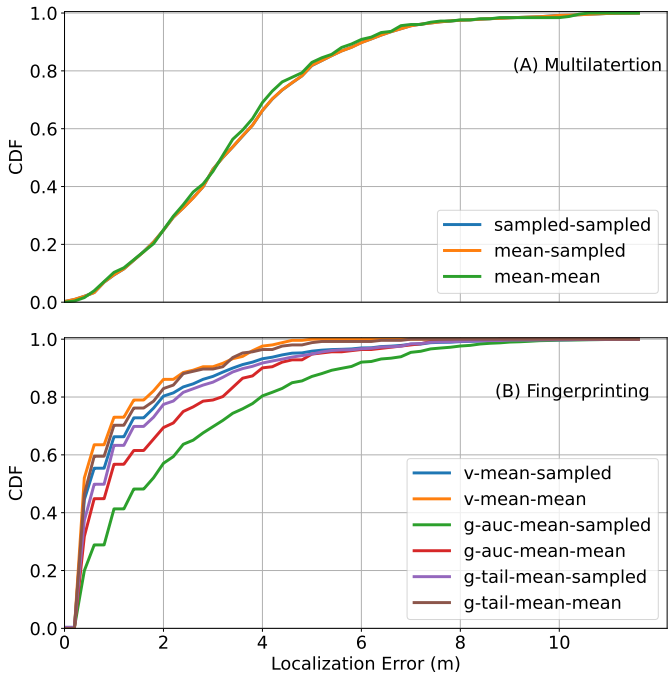


Figure 5: CDF curves of error in traditional methods

We use `scikit-learn`'s K-nearest neighbor API to search for the optimal k. The API is called with default settings, except `algorithm` is set to *brute*, and `n_neighbors` is set to k, which ranges from 1 to 20 for each validation run. For the data sets collected, we find that the validation error reaches its minimum when $k = 2$. Using this k, we evaluate the performance of vector-based fingerprinting; results are shown in Figure 5(B). The axis layout is the same as described earlier. The blue (v-mean-sampled) and orange (v-mean-mean) curves represent localization performance when vector-based fingerprinting is tested on the sampled (excluding the validation set) and mean testing sets, respectively. Similar to traditional multilateration, the performance of v-mean-mean is better than that of v-mean-sampled. Again, this is most likely an artifact due to information overlap between the mean testing and training sets.

The radio map for Gaussian-based fingerprinting is a 252×4 matrix prepared from the mean training set, where each row represents a reference point, and each column represents the Gaussian distribution, modeled from the mean and standard deviation of the RSSI measurements from the corresponding sensor. We follow the same procedure described above to find the best k, except that in the K-nearest neighbor API, (4) and the mean over A_{ij} are used for the AUC method, whereas (5) and the sum over negative logarithm of L_{ij} are used for the tail method. The result of the search shows that the minimum validation error is reached when $k = 2$ for both the AUC and the tail methods. Figure 5(B) shows the localization performance of Gaussian-based fingerprinting based on the optimal k value. The green (g-auc-mean-sampled) and red (g-auc-mean-mean) curves represent the performance under the AUC method, while the magenta (g-tail-mean-sampled) and brown (g-tail-mean-mean) curves represent the performance under the tail method. Both the g-auc-mean-sampled and g-tail-mean-sampled curves are based on the sampled (excluding the validation set) testing set, whereas both the g-auc-mean-mean and g-tail-mean-mean curves are based on the mean testing sets. The g-auc-mean-mean and g-tail-mean-mean curves again exhibit better performance than the g-auc-mean-sampled and g-tail-mean-sampled curves, respectively, which is likely an artifact.

Comparing performance across all curves (excluding the ones under mean-mean scenario), the vector-based fingerprinting (v-mean-sampled) performs the best. However, its lead over Gaussian-based fingerprinting under the tail method (g-tail-mean-sampled) is very modest.

Comparing the two methods within Gaussian-based fingerprinting, the tail method consistently outperforms the AUC method. This is most likely due to the difference in penalty structure, where the AUC method, unlike the tail method, does not penalize misalignment proportionally to its severity. Figure 6(A) shows an example of how the penalty structure of the AUC method would fail to match a set of unknown RSSI measurements to the most probable fingerprint. In the figure, a Gaussian-based fingerprinting radio map is visualized, which contains two reference points (P_1 and P_2), each fingerprinted by two sensors (S_1 and S_2). The two blue distributions represent the RSSI fingerprint of P_1, denoted by the means $[\mu_{11}, \mu_{12}]$. Similarly, the two orange distributions represent the RSSI fingerprint of P_2, denoted by the means $[\mu_{21}, \mu_{22}]$. For simplicity, all distributions are assumed to have the same standard deviation. The set of unknown RSSI measurements, $U = [R'_1, R'_2]$, are illustrated by two black dashed lines. Finally, the areas-under-the-curve used in the AUC method are labeled as A_{11} and A_{12} for P_1, and A_{21} and A_{22} for P_2.

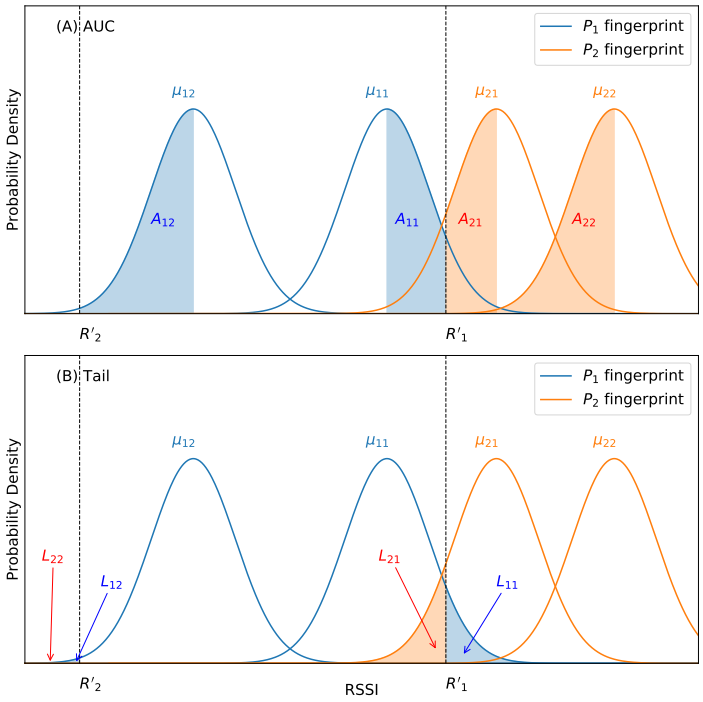


Figure 6: Example of matching under AUC or tail methods

Although it is clear that P_1 is a better match for U than P_2, it is not apparent in Figure 6(A). According to the AUC method described in Section 3.2, the match is selected as the smaller of $A_{11} + A_{12}$ and $A_{21} + A_{22}$. Due to the large misalignment between R'_2 and both μ_{12} and μ_{22}, we have $A_{12} \approx A_{22}$. Thus, the actual comparison is between A_{11} and A_{21}, which does not yield a clear winner at first glance. However, the decision should not have been left to a comparison between A_{11} and A_{21} in the first place. The extreme misalignment between R'_2 and μ_{22} should have already disqualified P_2. The AUC method is unable to recognize this because its penalty structure is inversely proportional to the severity of RSSI misalignment (i.e., the amount of penalty per unit of misalignment decreases as misalignment becomes more severe). R'_2 demonstrates how this penalty structure fails. While it is expected that a match between R'_2 and μ_{22} is penalized much more than a match between R'_2 and μ_{12}, the actual penalty associated with the former (A_{22}) is almost the same as the penalty associated with the latter (A_{12}). Thus, the failure to proportionally penalize misalignment makes the AUC method ill-equipped for Gaussian-based fingerprinting.

On the contrary, Figure 6(B) illustrates that the tail method can readily match U to P_1 as expected. The layout of Figure 6(B) is the same as Figure 6(A), except that the areas-under-the-curve are redefined and labeled as L_{11} and L_{12} for P_1, and L_{21} and L_{22} for P_2. According to the tail method described in Section 3.2, the match is selected as the smaller of $-lg(L_{11}) - lg(L_{12})$ and $-lg(L_{21}) - lg(L_{22})$. While L_{11} and L_{21} remain similar, the difference between L_{12} and L_{22} is large, with L_{22} being virtually 0. As a result, $-lg(L_{11})-lg(L_{12})$ is easily smaller than $-lg(L_{21}) - lg(L_{22})$, which matches U to P_1. The reasons why the tail method works are two-fold. First, its penalty structure is proportional to the severity of RSSI misalignment, which ensures that larger misalignment receives exponentially more penalty. Second, multiplication of tail probabilities (or the sum of their negative logarithms) further amplifies the misalignment penalty. In an extreme case, imagine that R'_1 is perfectly aligned with μ_{21} in Figure 6(B), and thus receiving the least amount of penalty. Yet, the severe misalignment between R'_2 and μ_{22} would still overwhelm the overall penalty for P_2 and readily make it a worse match than P_1. Therefore, the tail method is better-suited for Gaussian-based fingerprinting than the traditional AUC method; the difference in their localization performance is thus unsurprising.

7 Machine Learning Methods

Three methods—machine learning-boosted multilateration, machine learning classification of reference points, and machine learning regression of coordinates—are evaluated for RSSI-based passive localization. Each method is approached using three standard machine learning algorithms, MLP, SVM, and K-NN. Each model is trained on sampled and mean training sets to produce sampled and mean models, respectively.

During model training, we conduct two rounds of cross-validation to tune hyperparameters. For the sampled model, standard five-fold cross-validation is used. For the mean model, multi-fold cross-validation is not possible because the mean training set lacks sufficient redundancy. Thus, the same validation set discussed in Section 6.2 is used. In the first round of hyperparameter tuning, a wide range of values are provided, such that an optimal range can be identified for each hyperparameter. In the second round, values within each optimal range are used to find the exact hyperparameter values that minimize cross-validation error.

The finalized hyperparameter values are used to re-train the model, which is then subject to performance evaluation. The outcome of the evaluation is visualized as a CDF curve of the localization error in the same manner as described in Section 6. The sampled model is evaluated on the sampled testing set only, producing a sampled-sampled curve. The mean model is evaluated on both the sampled and mean testing sets, producing mean-sampled and mean-mean curves, respectively. As mentioned before, the validation set is excluded from evaluating the localization performance of the mean models.

All machine learning and hyperparameter tuning algorithms are executed using the corresponding APIs in `scikit-learn`.

7.1 Machine Learning-boosted Multilateration

In machine learning-boosted multilateration, the RSSI path-loss model is replaced by a machine learning regressor to estimate distance based on RSSI measurements. The advantage of this approach is that it assumes no prior knowledge of the complex RSSI-distance relationship and conducts regression without the constraints of a physical model.

Table 4: Metrics of machine learning RSSI-distance regressors

	MLP				SVM				K-NN			
	Sampled		Mean		Sampled		Mean		Sampled		Mean	
Sensor	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE
s_1	0.53	2.48	0.52	2.50	0.53	2.45	0.53	2.47	0.52	2.48	0.53	2.48
s_2	0.64	2.05	0.64	2.05	0.63	2.05	0.63	2.06	0.63	2.12	0.64	2.07
s_3	0.57	2.39	0.56	2.40	0.56	2.35	0.55	2.36	0.57	2.39	0.56	2.44
s_4	0.68	1.92	0.67	1.93	0.67	1.89	0.66	1.92	0.67	1.89	0.67	1.92

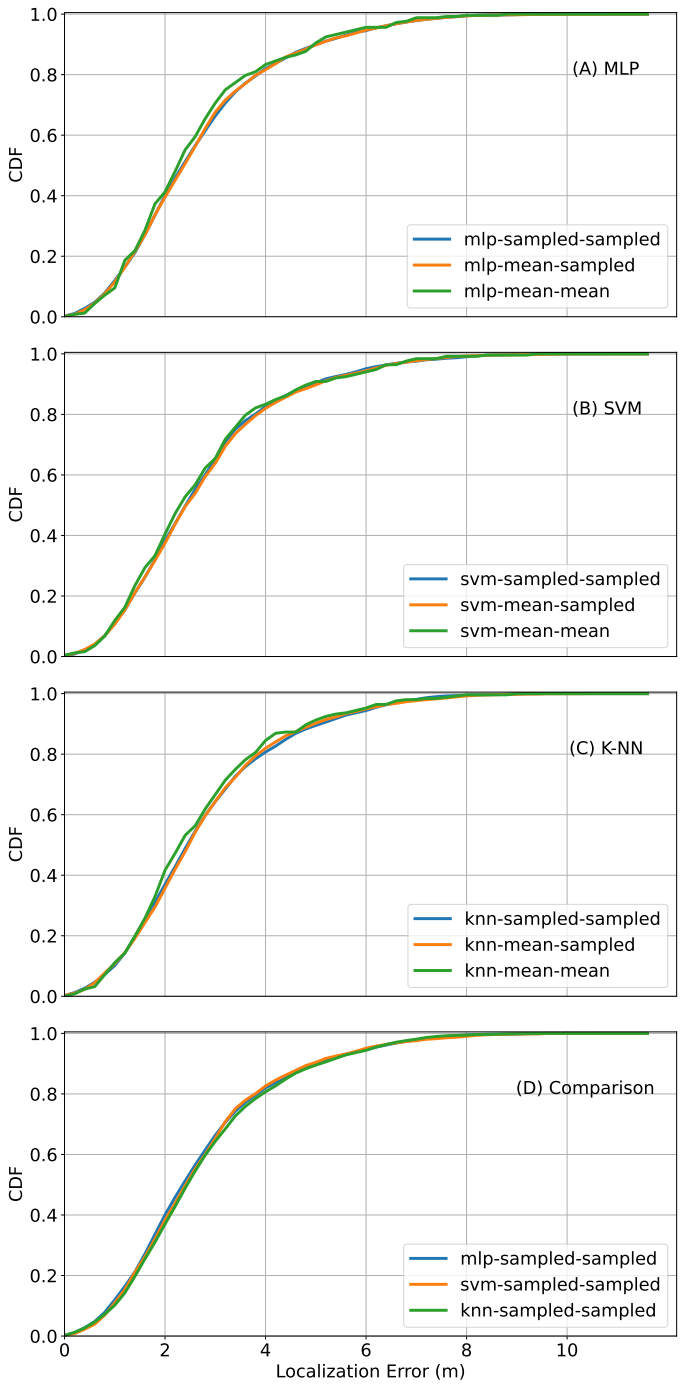


Figure 7: CDF curves of error in machine learning-boosted multilateration

The hyperparameters tuned for the MLP regressor are `hidden_layer_sizes`, `activation`, and `solver`; `C`, `kernel`, and `gamma` for the SVM regressor; and `n_neighbors`, `weights`, `algorithm`, and `leaf_size` for the K-NN regressor. The maximum number of tuning iterations is 1,000 for the MLP, and 20,000 for the SVM and K-NN regressors.

Table 4 lists the R^2 and MAE values for the machine learning RSSI-distance regressors after two rounds of hyperparameter tuning. Across machine learning algorithms, the sampled and mean models exhibit similar performance. Across sensors, s_4 exhibits the lowest error, whereas s_1 exhibits the highest error. This follows the results from the RSSI path-loss model. Comparing the performance between the machine learning and the RSSI path-loss models, we observe that the former consistently outperforms the latter in terms of MAE.

We use each machine learning RSSI-distance regressor to generate its corresponding multilateration model as described in Section 6.1. Figures 7(A), (B), and (C) demonstrate their localization performance. Across all three machine learning algorithms, the models have similar performance, which is not surprising, given the similar performance in RSSI-distance regression mentioned above. The mean-mean curve performs slightly better, which, as described in Section 6, could be an artifact.

A comparison is also made among the best CDF curves across the three methods (mean-mean curve excluded), as shown in Figure 7(D). There is little difference in localization performance among them.

7.2 *Machine Learning Classification of Reference Points*

Machine learning classification is an alternative to the traditional fingerprinting method. A trained machine learning classifier takes four RSSI measurements from a target device as input and produces a predicted label as output. The label can be mapped to a reference point, which is considered the estimated location of the target device.

The hyperparameters tuned for the MLP classifier are `hidden_layer_sizes`, `activation`, `solver`, and `alpha`; `C`, `kernel`, `gamma`, and `decision_function_shape` for the SVM classifier; and `n_neighbors`, `weights`, `algorithm`, and `leaf_size` for the K-NN classifier. The maximum number of tuning iterations is 1,000 for all classifiers.

Table 5: True and adjacent accuracy of machine learning classifiers

	MLP		SVM		K-NN	
Accuracy	Sampled	Mean	Sampled	Mean	Sampled	Mean
True	0.63	0.55	0.70	0.68	0.69	0.68
Adjacent	0.81	0.78	0.84	0.83	0.84	0.80

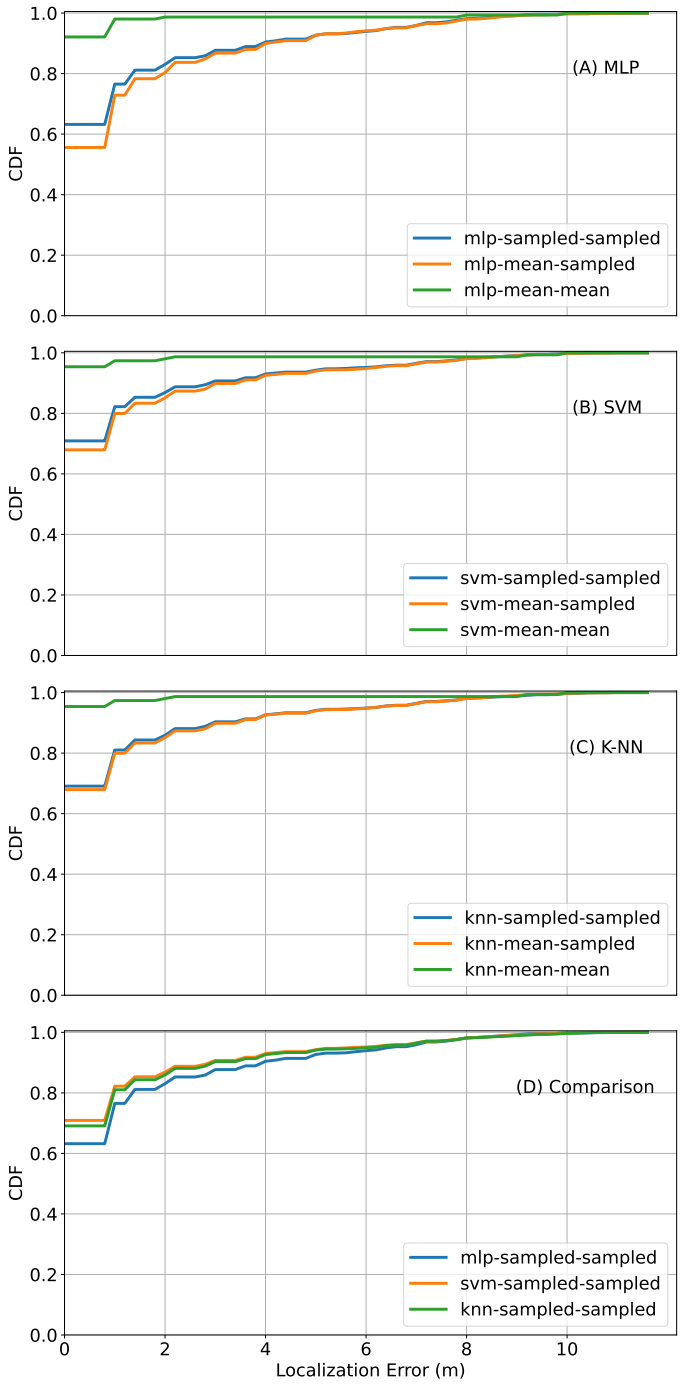


Figure 8: CDF curves of error in machine learning classification

Table 5 lists the true and "adjacent" accuracy of each classifier after two rounds of hyperparameter tuning. True accuracy represents the proportion of testing observations with correctly predicted labels. *Adjacent* accuracy denotes the proportion of testing observations with predicted labels that either match the true labels or one of the eight surrounding labels. The purpose of including adjacent accuracy is to reveal the degree of inaccuracy when a wrong prediction is made. Due to this intentional leniency, adjacent accuracy is always better than true accuracy. Across machine learning algorithms, the sampled model performs slightly better than the mean model. MLP performs the worst, whereas SVM and K-NN have similar performance.

Figures 8(A), (B), and (C) illustrate the localization performance of all three algorithms. Across machine learning algorithms, the sampled-sampled curve exhibits slightly better performance than the mean-sampled curve, but the mean-mean curve tops them both, with zero meter localization error more than 90% of the time. As discussed earlier, the exceptional performance exhibited by the mean-mean curve is most likely an artifact. However, the exaggeration in classification is much more pronounced than in multilateration. This is due to the way we compute localization errors for classifiers. Since we consider a correct match to a reference point as being equivalent to zero meter localization error, we accumulate many zero-error instances when evaluating the localization performance of classifiers. These zero-error instances inflate the CDF curves. In contrast, multilateration methods never exhibit zero meter localization errors.

We also compare the best CDF curves across the three machine learning algorithms, with the mean-mean curves excluded. Figure 8(D) shows the results, where SVM performs slightly better than the others.

7.3 Machine Learning Regression of Coordinates

Machine learning regression of coordinates trains two RSSI-coordinate regressors based on the x- and y-coordinates, respectively. It allows direct quantitative estimation of a target device's position. Each regressor takes four RSSI measurements from an observation as input and produces a pair of estimated coordinates as output.

The hyperparameters tuned for the MLP regressor are `hidden_layer_sizes`, `activation`, `solver`, and `alpha`; `C`, `kernel`, and `gamma` for the SVM regressor; and `n_neighbors`, `weights`, `algorithm`, and `leaf_size` for the K-NN regressor. The maximum number of iterations is 1,000 for the MLP regressor, and 20,000 for the SVM and KNN regressors.

Table 6: Metrics of machine learning RSSI-coordinate regressors

	MLP				SVM				K-NN			
	Sampled		Mean		Sampled		Mean		Sampled		Mean	
Coordinate	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE
x	0.84	1.24	0.83	1.30	0.91	0.71	0.82	1.29	0.91	0.71	0.88	0.97
y	0.86	1.16	0.85	1.24	0.90	0.74	0.85	1.21	0.91	0.71	0.90	0.80

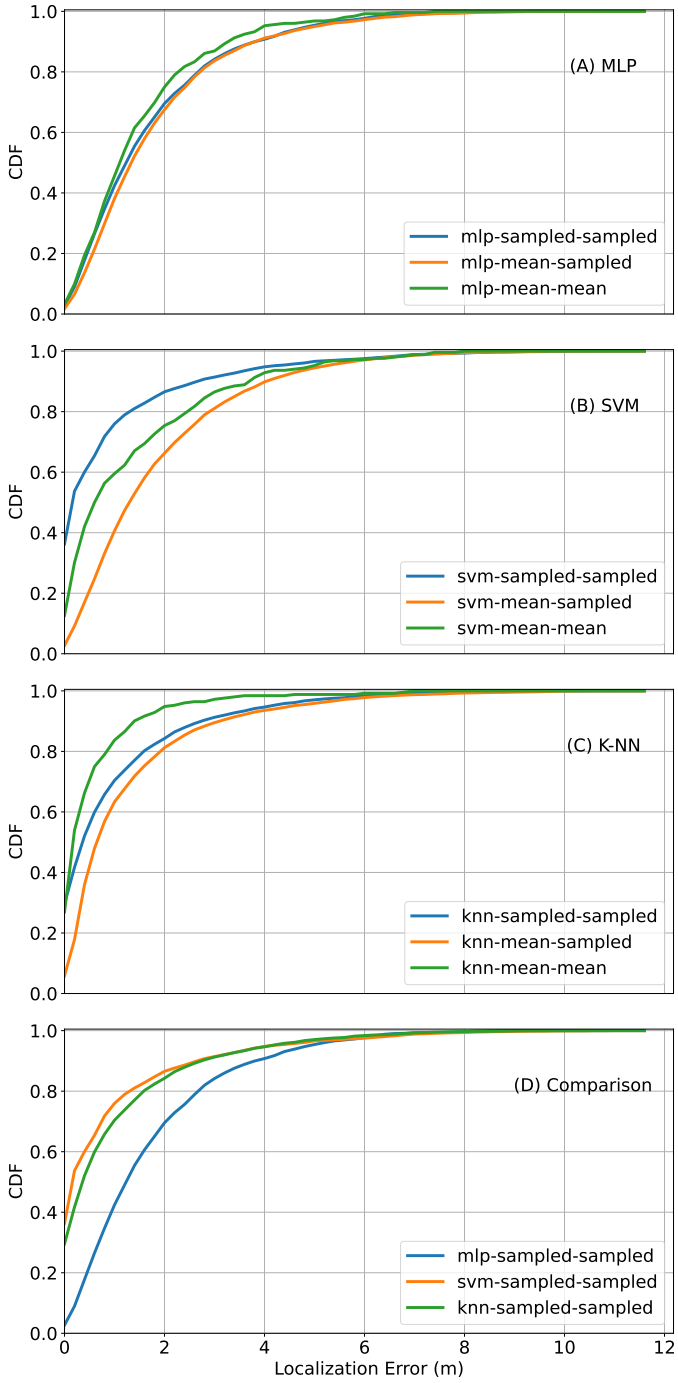


Figure 9: CDF of error with machine learning RSSI-coordinate regression

Table 6 lists the R^2 and MAE values for the machine learning RSSI-coordinate regressors after two rounds of hyperparameter tuning. Across all regressors, the sampled model has better regression performance than the mean model, which is likely due to the difference in training size. However, it is worth noting that the difference in training set size does not affect the performance of the RSSI-distance regressors (Section 7.1). A possible explanation of why the RSSI-coordinate regressor is more sensitive to training set size is that it is a more complex model (four features as input) than the RSSI-distance regressor (one feature as input). Across algorithms, the sampled models of SVM and K-NN perform slightly better than that of MLP, while the mean model of K-NN outperforms that of MLP and SVM. Between the two coordinates, y-coordinates generally have lower estimation error than x-coordinates. We speculate that this is the result of how our testbed is prepared, where there is more measuring variability in the x- versus y-axis.

Figures 9(A), (B), and (C) summarize the localization performance of all three algorithms. Within MLP and K-NN, the sampled-sampled curve performs slightly better than the mean-sampled curve, but the mean-mean curve outperforms both. This aligns with our earlier hypothesis that indicates performance exaggeration in the mean-mean curve. However, in the SVM models, the situation is surprisingly reversed. The sampled-sampled curve shows better performance than the exaggerated mean-mean curve. This is most likely due to the superior RSSI-coordinate regression performance of the SVM sampled model compared to the SVM mean model.

We then compare the performance of the best CDF curves across the three machine learning algorithms (excluding the mean-mean curves), as shown in Figure 9(D). The results show that the sampled-sampled SVM model performs the best.

8 Royal Rumble

To find the best-performing model overall, we juxtapose the best model within each method (excluding the mean-mean curves) in the same plot. If multiple models show equally good performance within a particular method, a model is arbitrarily selected. The result is shown in Figure 10. The axes are the same as before. The blue curve corresponds to the best performant model in traditional multilateration (sampled-sampled); the orange curve corresponds to the best in traditional fingerprinting (v-mean-sampled); the green curve corresponds to the best in machine learning-boosted multilateration (mlp-sampled-sampled); the red curve corresponds to the best in machine learning classification of reference points (svm-sampled-smapled); and the purple curve corresponds to the best in machine learning regression of coordinates (svm-sampled-sampled). Several observations can be made.

First, the best machine learning models outperform the best

traditional models. Specifically, machine learning-boosted multilateration performs better than its traditional counterpart. Similarly, machine learning classification performs better than traditional fingerprinting. This is not surprising because machine learning methods can train regressors and classifiers based on RSSI measurements without the constraints imposed by an underlying physical model. It allows them to capture patterns that might be ignored in the traditional methods.

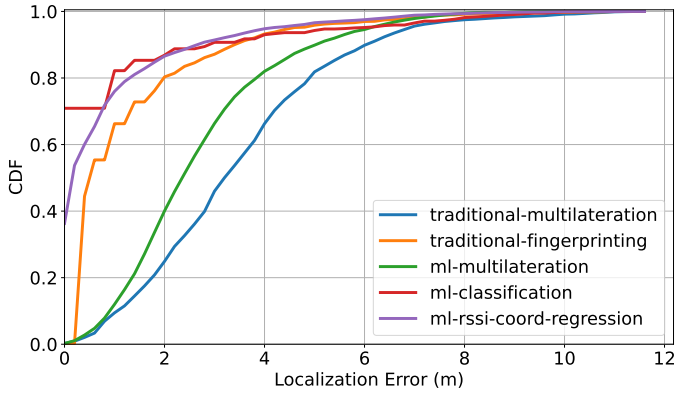


Figure 10: The best CDF curves of error among all methods

Second, the best non-multilateration models outperform the best multilateration models by a wide margin. This is most likely due to the nature of the multilateration methods, where two rounds of estimation are involved: RSSI to distance and distance to coordinates. Since each round of estimation incurs error, the multilateration methods accumulate more error than the other methods, where only one round of estimation is involved.

Third, except for fingerprinting, where there is no sampled-sampled option, all of the CDF curves used for comparison in Figure 10 are sampled-sampled curves. In other words, given the same sampled testing set, the sampled models either outperform or perform as well as the mean models for all methods. This result answers the question raised at the close of Section 5: The benefit of reduced random error in the mean training set does not outweigh the costs of reducing the size of the training set.

Finally, for a localization model to function at the scale of a city street, its error must be below a certain threshold. Ideally, the error should be small enough to tell whether a target device has moved from one business to another. Based on this reasoning, the best machine learning classifier (svm-sampled-sampled), the best machine learning RSSI-coordinate regressor (svm-sampled-sampled), and the best traditional fingerprinting (v-sampled-sampled) are the best candidates, as they all exhibit above 90% probability of having four meters or less localization error. However, future research must examine their spatial robustness (e.g., does the model work outside the testbed?) and temporal robustness (e.g., does the model work on data collected a week or a month later?).

9 Conclusion

In this paper, we have systematically evaluated the localization performance of several RSSI-based passive localization methods using data collected in a semi-controlled, obstruction-free outdoor environment. The methods include two adaptations of traditional active localization methods (multilateration and fingerprinting) and three machine learning methods (machine learning-boosted multilateration, machine learning classification of reference points, and machine learning regression of coordinates). The results show that the machine learning methods perform better than the traditional methods. Further, the SVM sampled model in the classification-based approach, the SVM sampled model in the RSSI-coordinate regression-based approach, and the traditional fingerprinting approach seem to have sufficiently good performance to be considered for use in real-world mobility monitoring. However, several limitations in formulating and evaluating these methods must be addressed and improved upon in future research.

First, all the models in this paper are trained and tested on data collected in a five-hour window on the same day. In practice, we have documented and verified that RSSI measured under identical emitter, sensor, and visible environmental conditions exhibits temporal variability [28]. Thus, a localization model must possess temporal robustness such that the time of data collection is irrelevant to its performance. Further, the exceptional performance of the mean-mean curves in almost all models appears to be an artifact due to the information overlap between the mean testing and training sets. It will be interesting to see whether the mean-mean curves remain exceptional when the mean testing set comes from data collected during a different session. Future research will pursue this evaluation.

In addition, the data used in this paper were collected in a near-ideal environment, free of obstructions and adverse weather events. We have yet to include samples with missing data in the analysis (e.g., three instead of four RSSI measurements in a sample). However, signal noise and incomplete data are inevitable in practice. Thus, future research will investigate how model performance deteriorates when data are collected with missing features (e.g., not all sensors capture a probe request), under unfavorable weather conditions (e.g., rain), in the presence of obstructions (e.g., walls, people), or with interference from other devices.

Relatedly, all the models trained in this paper are based on data collected by exactly four sensors. This presents a scalability issue, in which the addition or removal of a few sensors could require model re-training. Since scaling sensor deployments up and down is not uncommon, future research will focus on generating models resistant or adaptable to small changes in sensor availability.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This research is supported by the City of West Palm Beach, the Knight Foundation, and the Community Foundations of Palm Beach and Martin Counties. The authors would like to thank Chris Roog, Executive Director of the Community Redevelopment Agency in the City of West Palm Beach, for his continued assistance and support in realizing MobIntel; and thank the Florida Atlantic University *I-SENSE* team for providing hardware, software, and logistics support. This work was partially supported through the NSF ERC for Smart StreetScapes (CS3) (EEC-2133516).

References

[1] L. Sun, S. Chen, Z. Zheng, L. Xu, "Mobile Device Passive Localization Based on IEEE 802.11 Probe Request Frames," Mobile Information Systems, **2017**, 1–10, 2017, doi:10.1155/2017/7821585.

[2] W. Xue, W. Qiu, X. Hua, K. Yu, "Improved Wi-Fi RSSI Measurement for Indoor Localization," IEEE Sensors Journal, **17**(7), 2224–2230, 2017, doi:10.1109/JSEN.2017.2660522, conference Name: IEEE Sensors Journal.

[3] K. K. Mamidi, K. P. V., "ALTAR: Area-based Localization Techniques using AoA and RSS measures for Wireless Sensor Networks," in 2019 International Conference on contemporary Computing and Informatics (IC3I), 160–168, 2019, doi:10.1109/IC3I46837.2019.9055582.

[4] X. Li, Z. D. Deng, L. T. Rauchenstein, T. J. Carlson, "Contributed Review: Source-localization algorithms and applications using time of arrival and time difference of arrival measurements," Review of Scientific Instruments, **87**(4), 041502, 2016, doi:10.1063/1.4947001, publisher: American Institute of Physics.

[5] G. Guo, R. Chen, F. Ye, X. Peng, Z. Liu, Y. Pan, "Indoor Smartphone Localization: A Hybrid WiFi RTT-RSS Ranging Approach," IEEE Access, **7**, 176767–176781, 2019, doi:10.1109/ACCESS.2019.2957753, conference Name: IEEE Access.

[6] N.-S. Duong, T.-M. Dinh, "Indoor Localization with lightweight RSS Fingerprint using BLE iBeacon on iOS platform," in 2019 19th International Symposium on Communications and Information Technologies (ISCIT), 91–95, 2019, doi:10.1109/ISCIT.2019.8905160, iSSN: 2643-6175.

[7] C.-H. Lin, L.-H. Chen, H.-K. Wu, M.-H. Jin, G.-H. Chen, J. L. Garcia Gomez, C.-F. Chou, "An Indoor Positioning Algorithm Based on Fingerprint and Mobility Prediction in RSS Fluctuation-Prone WLANs," IEEE Transactions on Systems, Man, and Cybernetics: Systems, 1–11, 2019, doi:10.1109/TSMC.2019.2917955, conference Name: IEEE Transactions on Systems, Man, and Cybernetics: Systems.

[8] J. Wang, J. Luo, S. J. Pan, A. Sun, "Learning-Based Outdoor Localization Exploiting Crowd-Labeled WiFi Hotspots," IEEE Transactions on Mobile Computing, **18**(4), 896–909, 2019, doi:10.1109/TMC.2018.2849416, conference Name: IEEE Transactions on Mobile Computing.

[9] L. L. Oliveira, L. A. Oliveira, G. W. A. Silva, R. D. A. Timoteo, D. C. Cunha, "An RSS-based regression model for user equipment location in cellular networks using machine learning," Wireless Networks, **25**(8), 4839–4848, 2019, doi:10.1007/s11276-018-1774-4.

[10] Z. He, Y. Li, L. Pei, R. Chen, N. El-Sheimy, "Calibrating Multi-Channel RSS Observations for Localization Using Gaussian Process," IEEE Wireless Communications Letters, **8**(4), 1116–1119, 2019, doi:10.1109/LWC.2019.2908397, conference Name: IEEE Wireless Communications Letters.

[11] S. Yiu, M. Dashti, H. Claussen, F. Perez-Cruz, "Wireless RSSI fingerprinting localization," Signal Processing, **131**, 235–244, 2017, doi:10.1016/j.sigpro.2016.07.005.

[12] R. Sari, H. Zayyani, "RSS Localization Using Unknown Statistical Path Loss Exponent Model," IEEE Communications Letters, **22**(9), 1830–1833, 2018, doi:10.1109/LCOMM.2018.2849963, conference Name: IEEE Communications Letters.

[13] S. Sadowski, P. Spachos, "RSSI-Based Indoor Localization With the Internet of Things," IEEE Access, **6**, 30149–30161, 2018, doi:10.1109/ACCESS.2018.2843325, conference Name: IEEE Access.

[14] S. Mazokha, F. Bao, J. Zhai, J. Hallstrom, "MobIntel: Sensing and Analytics Infrastructure for Urban Mobility Intelligence," in 2020 IEEE International Conference on Smart Computing (SMARTCOMP) (SMARTCOMP 2020), Bologna, Italy, Italy, 2020.

[15] F. Bao, S. Mazokha, J. O. Hallstrom, "MobIntel: Passive Outdoor Localization via RSSI and Machine Learning," in 2021 17th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), 247–252, 2021, doi:10.1109/WiMob52687.2021.9606338.

[16] D. Alshamaa, F. Mourad-Chehade, P. Honeine, "Mobility-based Tracking Using WiFi RSS in Indoor Wireless Sensor Networks," in 2018 9th IFIP International Conference on New Technologies, Mobility and Security (NTMS), 1–5, 2018, doi:10.1109/NTMS.2018.8328704, iSSN: 2157-4960.

[17] H. Du, C. Zhang, Q. Ye, W. Xu, P. L. Kibenge, K. Yao, "A hybrid outdoor localization scheme with high-position accuracy and low-power consumption," EURASIP Journal on Wireless Communications and Networking, **2018**(1), 4, 2018, doi:10.1186/s13638-017-1010-4.

[18] M. Youssef, M. Mah, A. Agrawala, "Challenges: device-free passive localization for wireless environments," in Proceedings of the 13th annual ACM international conference on Mobile computing and networking, MobiCom '07, 222–229, Association for Computing Machinery, New York, NY, USA, 2007, doi:10.1145/1287853.1287880.

[19] J. Wilson, N. Patwari, "Radio Tomographic Imaging with Wireless Networks," IEEE Transactions on Mobile Computing, **9**(5), 621–632, 2010, doi:10.1109/TMC.2009.174, conference Name: IEEE Transactions on Mobile Computing.

[20] G.-l. Wang, X.-b. Hong, X.-m. Guo, Y. Fang, "A Data-Driven Three-Dimensional RSS Model for Device-Free Localization and Tracking," in 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 1705–1713, 2019, doi:10.1109/HPCC/SmartCity/DSS.2019.00233.

[21] A. Guillen-Perez, M.-D. Cano, "Counting and locating people in outdoor environments: a comparative experimental study using WiFi-based passive methods," ITM Web of Conferences, **24**, 01010, 2019, doi:10.1051/itmconf/20192401010, publisher: EDP Sciences.

[22] T. Kulshrestha, D. Saxena, R. Niyogi, J. Cao, "Real-Time Crowd Monitoring Using Seamless Indoor-Outdoor Localization," IEEE Transactions on Mobile Computing, **19**(3), 664–679, 2020, doi:10.1109/TMC.2019.2897561, conference Name: IEEE Transactions on Mobile Computing.

[23] J. Du, C. Yuan, M. Yue, T. Ma, "A Novel Localization Algorithm Based on RSSI and Multilateration for Indoor Environments," Electronics, **11**(2), 2022, doi:10.3390/electronics11020289.

[24] F. Shang, W. Su, Q. Wang, H. Gao, Q. Fu, "A Location Estimation Algorithm Based on RSSI Vector Similarity Degree," International Journal of Distributed Sensor Networks, **10**(8), 371350, 2014, doi:10.1155/2014/371350, publisher: SAGE Publications.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, **12**, 2825–2830, 2011.

[26] Wes McKinney, "Data Structures for Statistical Computing in Python," in Stéfan van der Walt, Jarrod Millman, editors, Proceedings of the 9th Python in Science Conference, 56 – 61, 2010, doi:10.25080/Majora-92bf1922-00a.

[27] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, S. . . Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," Nature Methods, **17**, 261–272, 2020, doi:https://doi.org/10.1038/s41592-019-0686-2.

[28] S. J. Pan, V. W. Zheng, Q. Yang, D. H. Hu, "Transfer Learning for WiFi-based Indoor Localization," AAAI Conference on Artificial Intelligence, 6, 2008.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 7, No. 6, 283-294 (2022)
www.astesj.com
Special Issue on Innovation in Computing, Engineering Science & Technology

ASTES Journal
ISSN: 2415-6698

Technical Aspects and Social Science Expertise to Support Safe and Secure Handling of Autonomous Railway Systems

Clemens Gnauer[*,1], Andrea Prochazka[1], Elke Szalai[2], Sebastian Chlup[3], Anton Fraunschiel[4]

[1]*Forschung Burgenland, Research Area for Sustainable Innovation, Eisenstadt, 7000, Austria*

[2]*University of Applied Sciences Burgenland, Social Studies, Eisenstadt, 7000, Austria*

[3]*Austrian Institute of Technology, Digital Safety and Security, Institute, Vienna, 1210 , Austria*

[4]*IQSoft GmbH, Vienna, 1120, Austria*

ARTICLE INFO

Article history:
Received: 30 August, 2022
Accepted: 03 December, 2022
Online: 27 December, 2022

Keywords:
Systems Engineering
Transport & Logistic Management
Social Work

ABSTRACT

In recent years the development of autonomous vehicles has increased tremendously and a variety of methodologies had been applied to make them more safe and secure. This work shows a multilevel approach combining Failure Mode, Effects and Criticality Analysis of an autonomous railway system with sociological and technical aspects to support safe operations and human-machine interactions in the field of autonomous railway systems. This approach includes all relevant technical components, as well as the assessment of measures for a safety process based on the Failure Mode, Effects and Criticality Analysis. We applied the Persona-Roberta model to assess safety aspects at the interface between humans and machines and applied both results to establish training materials. The results provide answers to questions about the avoidance of technical errors, discussions on security and safety aspects and shows organizational development tools for accident prevention. In the future the created knowledge will be used to improve trust in digital solutions and Cyber-Physical Systems.

1 Introduction

This paper is an extension of work originally presented in 2021 at the 5th International Conference on System Reliability and Safety (ICSRS) [1]. In the paper entitled "Knowledge Based Training Derived from Risk Evaluation Concerning Failure Mode, Effects and Criticality Analysis in Autonomous Railway Systems" the following four topics have been discussed:

- Railway vehicles, autonomous systems, safety & security, risk mitigation
- Development of a component meta-model (Figure 1)
- Risk evaluation with Failure Mode Effects and Criticality Analysis (FMECA)
  focused on technical and social aspects and the combination of these aspects to find solutions for minimizing risks
- Finding a solution to the issues of interaction between autonomous systems, Cyber- Physical Systems (CPS) and humans (Persona–Roberta model)

In thier work [2] looked at a meta-model that would assist the development of Cyber-Physical Systems (CPS) which are to be used to manage and control autonomous train systems on auxiliary railway lines in a safe and secure manner. In doing so, they discovered that most meta-models do not meet the particular requirements of the railway domain, particularly when focusing on the integration of the combination between safety and security issues. Our meta-model, which was developed in response to these findings, incorporates a wide range of particular technological and sociocultural factors as well as their interoperability. It serves as the foundation for the study described in this publication . In the extension for this article, the authors would like to discuss aspects of the research process in the interaction between CPS and people working at an operating site regarding safety and security aspects. This also concerns tools for developing trust and trustworthiness in technology and CPS. Therefore, we developed training materials for different target groups, which are working with and within these systems in different positions. For that, it is important to understand, which person has tacit knowledge and can decide actions and which person can only operate along a guideline.

The authors have been conducting research on the topic of au-

[*]Corresponding Author: Clemens Gnauer, Campus 1, A-7000 Eisenstadt, 0043577055471 & clemens.gnauer@forschung-burgenland.at

tonomous rail-bound vehicles on branch lines in Linz, Austria for over a total of 18 months. The overall project duration of the research project in which the presented research had been conducted was three years from August 2019- January 2021. Through direct access to the practical use of the planned autonomous system, safety aspects have been comprehensively discussed in the project. The authors overall objective is to connect social, technical and legal issues in an integrative way to achieve greater safety and security for humans and machines in autonomous systems.

Based on the data and research reports on technology acceptance and social issues in technology social research, a mixed methods approach was chosen, which helps to illuminate the research object with different methods according to the grounded theory [3]. Mixed methods research is often used in the behavioural, health, and social sciences, especially in multidisciplinary settings and complex situational or societal research. This approach is chosen when data is to be collected using different methods. These can be quantitative and qualitative methods, but they are not used side by side or in parallel, but triangulated to produce results. In the present research project, this approach was the right one, as the research question asks for social and technical aspects and interdisciplinary cooperation was used. Therefore, desk research methods were used to evaluate existing literature along categories of technical and social aspects. Based on the results, discussions were held with experts to generate information for the development of the Persona-Roberta model indicators. Furthermore, with this basis, the existing FMECA was extended to include social aspects. The aim of all these steps was to extend the error correction measures on the technical side to include social correction measures. These are the training materials developed from these steps, which in turn were fed from the tacit knowledge of the people responsible for the operation. In this way, a knowledge cycle was created to move from person-controlled rail-based operation to safe autonomous operation.

Doing this we set the following research objectives:

- Implementing social aspects to support the technical development of an autonomous railway system
- Extension of FMECA with design thinking methods
- Development of training concepts and materials to ensure safe and secure operation for humans and machines interacting with rail-bond vehicles

The rest of the paper is structured as follows. In section 2, relevant related scientific literature on the subjects of autonomous railway vehicles, FMECA and training materials and the Persona-Roberta model are presented. The next section 3 presents all objects used in the FMECA, failure mitigation and improvement opportunities in detail, followed by section 4, where training materials derived from the FMECA are presented. In the conclusion, a summary of the research is given and further research is presented.

2 Related Works

2.1 Autonomous Railway Vehicles

Recent technological improvements have led to more focus on the sector of autonomous vehicles. A variety of technologies are contributing to the progress of projects and research in this fields. When considering autonomous driving, roughly three parts, scene recognition, path planning and vehicle control are required [4]. While the first demands localization, object-detection and object-tracking algorithms, the second uses motion and mission planning. The latter makes use of the motion planner and uses path following. All these components use algorithms to aid and enable autonomous driving for vehicles. Regarding the railway sector, a variety of projects have targeted different aspects of autonomous driving for trains and railway vehicles, such as self-driving trains, autonomous train systems and others. Technologies such as the Internet of Things (IoT), AI, Dedicated Short-Range Communication (DSRC) and Positive Train Control (PTC) all enhance the autonomy of railway vehicles. Various work in the field of autonomous railway vehicles have been done in academia already. Speed scheduling for autonomous railway vehicle control systems has been researched using a variety of algorithms such as Neural Network [5], Neuro-Fuzzy systems [6], fuzzy interference system [7]. They propose the need for a secure and safe operation of trains enabled by an AI supported system.

Autonomous railway vehicles can be classified based on four Grades of Automation (GoA) [8]. GoA 1 describes controlled manual operations where the driver manually drives the train, opens passenger doors, etc. and is supported by a safety system, the Automatic Train Protection (ATP). GoA 2 includes automatic braking and change of rail tracks, etc. but is still dependent upon a driver. GoA 3 represents trains that operate autonomously but need a driver to attend the train, be onboard and take control in case of an emergency. GoA 4, however, are fully autonomous trains without a driver present. A lot of progress has been done in recent years regarding fully autonomous train systems. Rio Tinto, for examples, uses fully automated trains for its iron-ore freights in Pilbara, Australia. The system, called AutoHaul, operates 50 automated and unmanned trains on a 1500 km railroad, reducing time and costs [9].

In a recent work by [10], autonomous maintenance technologies for localisation and navigation have been presented. Respective technologies, such as on-board sensors, like Inertia Measurement Unit (IMU), tachometers, satellite-based position systems and cameras are used to identify the exact location, speed and other factors regarding the trains. In a further step, these technologies can be used for autonomous movement and driving of a train, as expressed by [11], for monitoring passengers and goods as well as trains and other systems. These advancements and technologies, however, bring up further issues regarding safety of humans, passengers in particular, and other objects involved that are interacting in one way or the other with an automated/driverless railway vehicle or corresponding control system. These factors need to be addressed appropriately.

2.2 Failure Mode Effects and Criticality Analysis

The Failure Mode and Effects Analysis (FMEA) is a formalized method that enables the identification of possible failures of components of an overall system and their further classification based on root causes, failure modes and estimation of risk [12]. Therefore, errors are identified for specific objects of a system and technical effects for this error are described as well as possible reasons for the respective errors. In this article, we present results from a special form, the Failure Mode Effects and Criticality Analysis

(FMECA)that additionally includes a Criticality Analysis. This is represented in the risk priority number (RPN) assigned to each risk for each failure. The RPN is calculated by multiplying severity, occurrence and detection rates for each failure [13], [14].

When looking at scientific research in the field, FMECA has been applied in various areas of railroad systems. For example, in [15] author conducted an FMECA assessment for heating, ventilation and air conditioning (HVAC) systems of transportation railway systems. They identified the main criticalities and in a further step applied a new analytical method to identify a threshold risk value. In their study, [16] used an FMECA on the passenger door systems of railroad rolling stock. Additionally, they looked at inexpensive and reliable criteria in that context and developed a mileage-based preventive maintenance program to lessen failure recurrence.

To find an optimum FMECA process and methodology for railroad systems, [17] have done a structured evaluation of various FMEA approaches. As a result, they contrasted the approaches and traits of the FMECA standards MIL-1629a, SAE-J1739, and IEC-60812 for use in the automotive, electronics, and military industries, respectively. They discovered that each was deficient in certain components necessary for a comprehensive approach to FMECA for train systems, and, as a result they combined elements of failure modes (SAE-J1739), maintenance analysis (MIL-1629a), and localization of impacts (IEC-60812). These examples demonstrate that research in this area is ongoing, however it appears that there are no studies that examine the social implications of risk reduction and error correction in FMECA. This gave us even more motivation to pursue improved FMECA research in the railway industry and incorporating social measures for error correction.

2.3 Training materials and the Persona-Roberta Model

The chosen models of error analysis (FMECA) and the developed Persona–Roberta model make it clear that it is relevant and important to seriously work on these levels in order to create meaningful error correction measures, but also training and work documents for the executing employees.

It is known from models of technology assessment that an awareness of the possible consequences of a technical implementation, but also of a human action, must be comprehensively created. These models are complemented by technology acceptance approaches that help to convey assessment processes or develop teaching tools for the people concerned. Acceptance is the subject of numerous scientific studies. One of the best-known models for explaining acceptance is the Technology Acceptance Model (TAM) [18]. This model describes the acceptance of new technology in IT. Technology interacts strongly with the product, the application of knowledge and with the development of a product. Both terms structure technology developments and the access of people to these developments. This approach is the basis for the further development of the Persona-Roberta model chosen in the research project, because as the error analysis and the development of the error correction measures (FMECA) show, in addition to technical knowledge, it also - or above all - requires trust in the new technology.

Knowledge about acceptance requires knowledge about the attitudes of the counterpart and in order to make these visible, the aforementioned model of Persona(s) was chosen. Interactions between people and technology (human-machine) require an exact understanding of precisely these interactions. For this purpose, the Persona(s) model can be extended to the Robertas, which makes interfaces and interactions between these two systems visible. In the present use case - in the case of self-propelled rail-bound vehicles in connection with industrial production, this existing knowledge of the employees must be recognised, used and further developed in order to ensure safe - in the sense of error- and accident-free - operations. In the research project, established models were further developed and combined by the various disciplines represented in order to find customised solutions. The combination of FMECA theory, TAM-models and the Persona–Roberta model allowed the derivation of proposals for the development of communication tools and training materials for the staff concerned.

”In the future, there will be a co-existence between humans and machines. ( ) The relationship (...) will therefore be crucial for the experience” [19]. This pattern is also about trust in the machine (sensors, camera, IoT, ...) , about knowing how it works and about the impact of the interaction with his Persona as an expert for autonomous railways. The development of the Persona(s) supported the research process and made it possible to draw on the generated ideas of possible expectations and fears. This is important, because one result of fear could be errors or wrong reactions in the event of a malfunction, which must be avoided. For this reason, the Persona model was expanded and shows in detail the possible interactions between human and machines. The development was based on the Persona-Robona model in [20] and makes it possible to describe and understand relationships between the systems that are responsible for a smooth process. At the same time, required knowledge modules, expertise and communication requirements become visible, which can subsequently be addressed through training measures [20].

3 Results of Failure Mode Effects and Criticality Analysis

In this section, the FMECA objects identified in the research project [1] are presented and their respective RPNs and technical, as well as socio-technical error corrective measures are displayed.

The research concentrated on auxiliary railway lines, while especially focusing on aspects of security and safety related to this form of railway transportation. In this matter, the authors present the autonomous railway vehicle as an exemplary extract from the proposed autonomous railway system. Figure 1 shows the components of the respective component of an autonomous railway vehicle. For each component of the meta-model the conducted FMECA is presented, including technological as well as socio-technical error correction measures.

3.1 Camera

When it comes to retrieving information about surroundings, cameras represent one of the most common technologies. They represent a component that is comparatively cheap but has a significant impact on the perception of the surroundings. In this sense we are not talking about a simple camera but a component capable of

processing image data in order to detect and identify potential obstacles such moving humans or static objects. In combination with other components a camera can contribute to the decision making within an automated vehicle and therefore, has a vital impact on the safety[21].

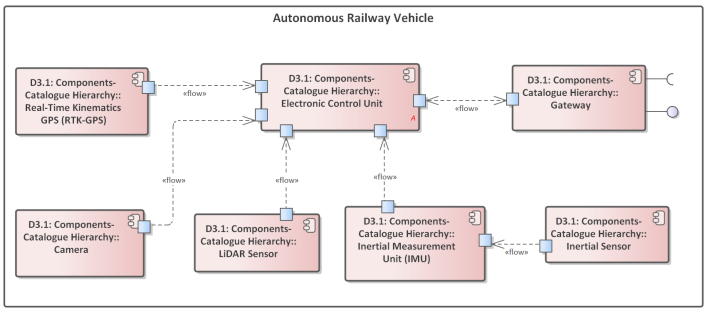


Figure 1: Subcomponents of the autonomous railway vehicle as a component of the meta-model

current (respectively planned state) object Camera			
possible errors	defect on the camera and/or the lense	unclean lense	error in the power connection
possible technical effects	obstacles are not recognized (by Persona or Roberta depending on what needs to be done)	pictures are not displayed correctly (Persona - doesn't recognize that the camera is the reason for the failure)	no action possibilities in case of missing pictures
possible error reasons	object falls down / accident	deposit on the lens	power cable damaged
estimation of the potential danger as well as possible measures for error prevention			
occurrence (O)	6	7	3
severity (S)	10	8	10
detection (D)	4	1	6
risk priority number (RPN)	**240**	**56**	**180**
technical error correction measure	protect camera or lens	cleaning the lens, self-cleaning lens, second camera	install power cable alternatively, redundant power supply
socio-technical error correction measure	interaction between Roberta and Persona	at the Persona level: schedule regular cleaning / "cleaning plan"	check the cable sheathing for visible damage

Figure 2: Camera object from FMECA

For a graphical depiction of the result of the FMECA for the object camera see Figure 2. When we connect the camera component to the FMECA, we identified the following possible errors: a defect on the camera itself, an unclean lens or an error in the power connection. The possible technical effects of the error defect on the camera itself are, firstly, that obstacles are not recognized, secondly, pictures are not displayed correctly, or lastly, that there are no action possibilities in case of missing pictures. Possible reasons for the above-mentioned errors are accidents caused by fallen objects, deposits on the lens or damages on the power cables. Regarding the estimation of the potential danger as well as possible measures for error prevention, we evaluate the error possibilities with the metrics occurrence (O), severity (S) and detection (D). For each of them we developed a metric corresponding to the potential occurrences of errors, as well as to the risk of severity and detection in several rounds based on the respective expert knowledge of the entire project team. The RPN determined was assigned to three categories with the help of an ABC analysis. The first category is reserved for RPNs between 0 and 299, where the risk is assessed as low and will be accepted. The second category is for RPNs between 300 and 749, where the risk is assessed as medium and, thus, taking measures to omit the failure mode is recommended. The third category is for RPN from 750 to the maximum possible RPN of 1000, the risk is assessed as high and, thus, assessed that an action is necessary, since on this level people could be in danger [1]. Doing this we receive the RPN for each scenario. In the shown case of the camera, we receive the RPN 240 for ”obstacles are not recognized” caused by a ”defect on the camera itself”, the RPN 56 for ”pictures are not displayed correctly” caused by an ”unclean lens” and the RPN 180 for ”no action possibilities in case of missing pictures” caused by an ”error in the power connection”. For the high-risk areas identified in this way, troubleshooting measures will be developed in a separate next step. For all possible errors listed in the FMECA, measures have also been taken to reduce the cause of errors within the system. These measures are divided into technical and sociological measures [1]. For the technical corrective measure, we determine tasks to protect the camera and/or lens to eliminate the possibility of a defect of the camera or lens. To avoid the possible errors induced by unclean lenses we can clean the lens and/or use self-cleaning lenses and/or a built-in redundancy by using a second camera. To avoid an error in the power connection we focus on the opportunities of alternative ways and areas for the installation, as well as a safer installation of the power chords and/or implementing a redundant power supply. As socio-technical corrective measures we determine the following actions to increase the safety level of the object camera. Continuous interaction between Persona(s) and Roberta(s) to detect possible defects on the camera and/or lens as soon as possible.

3.2 *Real Time Kinematics Global Positioning System*

Global Navigation Satellite Systems (GNSS) offer various ways of positioning solutions. One of which is standalone Global Positioning System (GPS). It utilizes signals retrieved from a radio link to calculate the current position [22]. However, standalone GPS provides an accuracy of only 2-5 metres [23] and therefore does not represent the up-to-date best solution when it comes to automating vehicles. Consequently, other solutions to determine the position more accurately were invented. With Differential GPS (D-GPS) a more precise position can be calculated. By adding a reference station whose position is known, an additional stationary point can be added to the calculation [24]. Thus, the moving object receives information from the satellites as well as the reference station which enables a precision of 0.3-0.8 meters [23]. The currently most advanced version of GNSS based position detection is Real Time Kinematics (RTK). They are based on the same principles as D-GPS but involve sophisticated computations and formulas during the course of processing the data. With RTK accuracies of 1-5 centimetres can be achieved [23].

For a graphical depiction of the result of the FMECA for the object Real Time Kinematics Global Positioning System see Figure 3. When we connect the Real Time Kinematics RTK-GPS component to the FMECA, we identified the following possible errors: a loss of the power connection to the RTK-GPS and a loss of the possibility

to connect to the reference station.

current (respectively planned state) object Real Time Kinematics RTK-GPS		
possible errors	lost power connection to the RTK-GPS	connection to reference station not possible
possible technical effects	unknown position of the autonomous vehicle	incorrectly displayed position of the autonomous vehicle
possible error reasons	damaged power cable mechanical damage by a person	RTK-GPS damaged, reference station damaged
estimation of the potential danger as well as possible measures for error prevention		
occurrence (O)	3	4
severity (S)	8	6
detection (D)	4	6
risk priority number (RPN)	**96**	**144**
technical error correction measure	install power cable alternatively	regular inspection intervals with checklists for the stations
socio-technical error correction measure	check the cable sheathing for visible damage	regularly check compliance with the inspection plan

Figure 3: RTK object from FMECA

The possible technical effects of the error loss of the power connection are, firstly, that the position of the autonomous vehicle is unknown and, secondly, incorrectly displayed positions of the autonomous vehicle. Possible reasons for the above-mentioned errors are damages on the power cables and/or damages on the reference station. Regarding the estimation of the potential danger as well as possible measures for error prevention we evaluate the error possibilities with the metrics occurrence (O), severity (S) and detection (D). For each of them we developed a metric corresponding to the potential occurrences of errors as well as to the risk of severity and detection. Doing this we received the RPN 96 for damages on the power cables and the RPN 144 for damages on the reference station.

3.3 Light Detection and Ranging Sensor

In the past Light Detection and Ranging (LiDAR) was utilized in areas such as measuring ground topography, vegetation canopies or predicting forest stand structures [25]. Nowadays the application field of LiDAR has adapted towards automation and autonomous vehicles. With LiDAR one can measure the distance to targets which might be potential obstacles. This is done by directing a laser onto surfaces within the environment and measure the time-of-flight, i.e. the time needed for the reflected light to return to the light source [26]. Moreover, the change in wavelength is considered. This way, the environment can be scanned and a digital representation of the surroundings is generated. Especially, when thinking about autonomous systems such as robots, vehicles or railway vehicles it becomes apparent that static information and manual measurements are not sufficient, as an up-to-date representation of the current surrounding of the automated entity is required. Therefore, LiDAR takes care of remote sensing in terms of obstacle detection by scanning the environment for autonomous vehicles and other objects and plays an essential role in automatically navigating through a given path.

current (respectively planned state) object Lidar Sensor							
possible errors	obstacle is overlooked	broken sensor	lost power connection to the Lidar Sensor	distance measurement does not work	EMERGENCY STOP sensors trigger at the wrong time, although there is no danger	EMERGENCY STOP sensors are not triggered in the event of danger	positioning at the parking position, blast furnace or steel mill does not work at the wrong time
possible technical effects	crash with obstacle	values are not forwarded or forwarded incorrectly	no response from Lidar Sensor	crash with obstacle	availability is limited - autonomous operation is difficult	Danger to persons and damage to property possible	end of journey unclear
possible error reasons	speed measurement faulty	broken sensor	damaged power cable mechanical damage by a person	unclean sensor	sensors parameterised too sensitively or environmental influences	Sensors incorrectly parameterized or environmental influences	positioning sensors failed
estimation of the potential danger as well as possible measures for error prevention							
occurrence (O)	2	6	4	7	8	9	4
severity (S)	10	7	4	9	4	10	3
detection (D)	8	7	3	4	3	9	6
risk priority number (RPN)	**160**	**294**	**48**	**252**	**96**	810	**72**
technical error correction measure	compare speed measurement using GPS data	plausibility check, comparison with other sensors	install power cable alternatively, redundant power supply	install sensor alternatively. Install self-cleaning for LIDAR (e.g. ultrasound)	use certified systems or carry out certification, redundancy	use certified systems, redundancy	for important positions install 2 sensors
socio-technical error correction measure	persona must not intervene	set limit values and training measure on this table / knowledge	check the cable sheathing for visible damage	clean sensors at regular intervals	train the operating people	train the operating people	integrate regular test intervals

Figure 4: LiDAR object from FMECA

For a graphical depiction of the result of the FMECA for the object Light Detection and Ranging Sensor see Figure 4. When we connect the Lidar Sensor to the FMECA, we identify the following possible errors: obstacle is overlooked, broken sensor, lost power connection to the LiDAR sensor, distance measurement does not work, EMERGENCY STOP sensors trigger at the wrong time, although there is no danger, EMERGENCY STOP sensors are not triggered in the event of danger and positioning at the parking position, blast furnace or steel mill does not work at the wrong time. The possible technical effects of these errors are: crash with obstacle, Sensor values are not forwarded or forwarded incorrectly, no response from LiDAR sensor, crash with obstacle, availability is limited and ,thus, autonomous operation is difficult, danger to persons and damage to property possible and end of journey unclear. Regarding the estimation of the potential danger as well as possible measures for error prevention we evaluate the error possibilities with the metrics occurrence (O), severity (S) and detection (D). For each of them we developed a metric corresponding to the potential occurrences of errors as well as to the risk of severity and detection. Doing this we received the RPNs from left to right RPN 160, RPN 294, RPN 48, RPN 252, RPN 96, RPN 810 and RPN 72. For the technical corrective measures, we determine tasks to protect the LiDAR Sensor. To avoid the ”possible error obstacle is overlooked” we compare speed measurement using GPS data, plausibility checks and comparison with other sensors to avoid broken sensors and a safer installation of the power connections to avoid errors in the power connection to the LiDAR Sensor. To avoid ”the possible error distance measurement does not work” we install sensors alternatively and/or install self-cleaning capabilities for the LiDAR unit (e.g. ultrasound). To avoid ”EMERGENCY STOP sensors trigger at the wrong time, although there is no danger” as well as ”EMERGENCY STOP sensors are not triggered in the event of danger”, we use certified systems and/or carry out certification, redundancy and avoid that ”positioning at the parking position, blast furnace or steel mill does not work at the wrong time” we install two separate LiDAR sensors for important positions. As socio-technical corrective measures we determine the following actions to increase the safety level of the system Lidar Sensor. In case of ”obstacle is overlooked” the Persona(s) must not intervene, in case of ”broken sensor” limiting values should be set and training measure on this

table / knowledge, in case of ”lost power connection to the LiDAR Sensor” one has to check the cable sheathing for visible damage. In case of ”distance measurement does not work” the sensors must be cleaned at regular intervals. In case of ”EMERGENCY STOP sensors trigger at the wrong time, although there is no danger” and ”EMERGENCY STOP sensors are not triggered in the event of danger” a measure is to train the operating people and in case of ”positioning at the parking position, blast furnace or steel mill does not work at the wrong time” regular test intervals should be integrated.

3.4 Inertial Measurement Unit (IMU)

Inertial Measurement Units (IMU) are utilized in a variety of today’s technical systems. Most commonly, IMUs are applied in the automotive and aerospace sector. But even smartphones rely on the technology to measure velocity, acceleration, rotation and gravitational force [27]. An IMU relies mainly on two types of sensors. An accelerometer capable of measuring the inertial acceleration and a gyroscope aimed at detecting the angular rotation. Alongside a given starting location e.g. retrieved via D-GPS or Real Time Kinematics GPS an IMU can give information about a vehicle’s position and even orientation. For this, linear and angular acceleration measurements are used to improve the precision of the location determination. Some newer IMUs even integrate a magnetometer to further boost the gyroscope measurements [28] , [29]. Moreover, IMUs in contrast to cameras, LiDARs or other sensing components need no exposure to the outside and can be mounted at any part of the vehicle and are not affected by outside conditions. [29]. Consequently, they can aid in a precise localization of an autonomous railway vehicle, especially in areas where other types of sensors reach their limits e.g. in tunnels where the GPS signal is lost.

current (respectively planned state) object Inertial Measurement Unit IMU				
possible errors	broken sensor	input is incorrectly processed/At the Persona level: identification of submitter/At the Roberta level: check state of interaction	output is incorrectly processed/At the Persona level: identification of submitter/At the Roberta level: check state of interaction	error in the power connection
possible technical effects	values are not forwarded or forwarded incorrectly	precision is not given	wrong information in the output	no action possibilities in the IMU
possible error reasons	sensors incorrectly calibrated At the Persona level: identify who calibrated the sensor	wrong input	wrong output	damaged power cable machanical damage by a person
estimation of the potential danger as well as possible measures for error prevention				
occurrence (O)	6	2	2	3
severity (S)	4	4	4	4
detection (D)	3	8	8	4
risk priority number (RPN)	**72**	**64**	**64**	**48**
technical error correction measure	calibration monitoring	testdrive before first use	testdrive before first use	install power cable alternatively, redundant power supply
socio-technical error correction measure	define table of values for calibration and develop training measure on this table/knowledge	define error table - when is it wrong?, Train the Personas	set up error correction measures using Roberta; At the personal level: knowledge of recognizing and adapting faulty programs	check the cable sheathing for visible damage

Figure 5: Inertial Measurement Unit object from FMECA

For a graphical depiction of the result of the FMECA for the object Inertial Measurement Unit see Figure 5. When we connect the Inertial Measurement Unit IMU to the FMECA, we identified the following errors: ”broken sensor”, ”incorrectly proceeded input”, ”incorrectly proceeded output” and ”errors in the power connection”. The possible technical effects of the error ”broken sensor” are that values are not forwarded or forwarded incorrectly. The technical effect of the error ”incorrectly proceeded input” is that the needed precision is not given. The technical effect of the error ”incorrectly proceeded output” is that wrong information is used for further actions and the technical effect for ”error in the power connection” is that there is no action possible for the IMU. Possible reasons for the above-mentioned errors are: incorrectly calibrated sensors, wrong inputs, wrong outputs and damages on the power cable. Regarding the estimation of the potential danger as well as possible measures for error prevention we evaluate the error possibilities with the metrics occurrence (O), severity (S) and detection (D). For each of them we developed a metric corresponding to the potential occurrences of errors as well as to the risk of severity and detection. Doing this we received the RPNs from left to right in Figure 5 RPN 72, RPN 64, RPN 64 and RPN 48. As technical corrective measures we determine tasks to protect the IMU to eliminate the possibility of failure. In this case we determine the calibration of the monitors to avoid broken sensors, test drives before first usage, to avoid incorrect inputs as well as incorrect outputs and alternative ways for the installation, as well as a safer installation of the power connections. As socio-technical corrective measures we determine the following actions to increase the safety level of the system IMU. To avoid errors in the calibration we define ”table of values for calibration and develop training measure on this table/knowledge”. To avoid incorrect inputs, we identified the measure ”define error table and train the Persona(s)”. To avoid incorrect outputs, we set up error correction measures using Robertas. At the Persona level we train the knowledge of recognizing and adapting faulty programs and to avoid errors in the power connection, we implement continuous cable sheathing checking.

3.5 Inertial Sensor

For a graphical depiction of the result of the FMECA for the object Inertial Sensor see Figure 6. When we connect the Inertial Sensor to the FMECA, we identified the following possible errors: ”no data output”, ”a broken sensor”, ”incorrect values”, ”errors in the power connection”.

current (respectively planned state) object Inertial Sensor				
possible errors	no data output	broken sensor	incorrect values	lost power connection to the Inertial Sensor
possible technical effects	data won't be forwarded	values are not forwarded or forwarded incorrectly	incorrect values are passed on	no response from Internal Sensor
possible error reasons	damaged power cable machanical damage by a person	broken sensor	sensors incorrectly calibrated	damaged power cable machanical damage by a person
estimation of the potential danger as well as possible measures for error prevention				
occurrence (O)	3	6	4	4
severity (S)	4	4	6	4
detection (D)	4	3	5	3
risk priority number (RPN)	**48**	**72**	**120**	**48**
technical error correction measure	install power cable alternatively, redundant power supply	plausibility check, comparison with other sensors	implement regular intervals for calibration	install power cable alternatively, redundant power supply
socio-technical error correction measure	check the cable sheathing for visible damage	set limit values and training measure on this table / knowledge	train personas to an acceptable range of values	knowing where which power cables are, which measures must be taken to start emergency supply - training measure

Figure 6: Inertial Sensor object from FMECA

The possible technical effects of the error ”no data output” is that data won’t be forwarded. The technical effect of the error ”broken sensor” is that values are not forwarded or forwarded incorrectly. The technical effect of the error ”incorrect values” is that incorrect values are passed on and the technical effect of an error in the power

connection is that there is no response from the Inertial Sensor. Possible reasons for the above-mentioned errors are damages in the cables, broken sensors, incorrectly calibrated sensors, and damages on the power cable. Regarding the estimation of the potential danger as well as possible measures for error prevention we evaluate the error possibilities with the metrics occurrence (O), severity (S) and detection (D). For each of them we developed a metric corresponding to the potential occurrences of errors as well as to the risk of severity and detection. Doing this we received the RPNs from left to right in Figure 6 RPN 48, RPN 72, RPN 120 and RPN 48. As the technical corrective measure, we determine tasks to protect the Inertial Sensor to eliminate the possibility of failure. In this case we determine an alternative installation of the cables to avoid the error of "no data output", plausibility checks and comparison with other sensors to avoid "broken sensor". To avoid "incorrect values", we implement regular intervals for calibration and to avoid errors in the power connection we implement a safer installation of the power connections. As socio-technical corrective measures we determine the following actions to increase the safety level of the system Inertial Sensor. To avoid errors in the power connection we implement continuous cable sheathing checking. To avoid errors caused by a broken sensor we set limits of values and develop training measures on this table/knowledge. To avoid incorrect values, we define an acceptable range of values and train the Persona(s). To avoid errors in the power connection we implement trainings to identify and know the exact locations of power cables and which measures must be taken to start emergency power supply.

3.6 Gateway

A Gateway or network gateway is used to establish communication between networks. It represents a hardware component capable of addressing a host potentially unknown to the user and enables the exchange of data with various other components. In our use case, we require a gateway for the automated railway vehicle to communicate with the infrastructure, provide information as well as receive information in terms of controls and reactions. Gateways find multiple applications in networking and in the area of IoT [30], [31].

current (respectively planned state) object Gateway			
possible errors	data will not be transmitted	disruption in the data transmission	lost power connection to the Gateway
possible technical effects	movement command is not executed	data records are not accepted	no response from Gateway
possible error reasons	data line interrupted	data timestamps faulty	damaged power cable mechanical damage by a person
estimation of the potential danger as well as possible measures for error prevention			
occurrence (O)	5	5	3
severity (S)	4	2	4
detection (D)	7	9	3
risk priority number (RPN)	**140**	**90**	**36**
technical error correction measure	query to the ECU if data flow is available	timestamp sanity-check	Install power cable alternatively, redundant power supply
socio-technical error correction measure	at the Persona level: train how to recognize that data line is broken and requiers response	sensor/signal if timestamp is incorrect/training to pay attention to this	check the cable sheathing for visible damage

Figure 7: Gateway object from FMECA

For a graphical depiction of the result of the FMECA for the object Gateway see Figure 7. When we connect the Gateway to the FMECA, we identified the following possible errors: "data will not be transmitted", "distribution in the data transmission", "errors in the power connection". The possible technical effects of the error "data will not be transmitted" are that movement commands are not executed. The technical effects of a distribution in the data transmission are that data records are not accepted and the technical effect of an error in the power connection is that that there is no response from the Gateway. Possible reasons for the above-mentioned errors are interruptions in the data line, faulty data timestamps and damages on the power cables. Regarding the estimation of the potential danger as well as possible measures for error prevention we evaluate the error possibilities with the metrics occurrence (O), severity (S) and detection (D). For each of them we developed a metric corresponding to the potential occurrences of errors as well as to the risk of severity and detection. Doing this we received the RPNs from left to right in Figure 7 of RPN 140, RPN 90 and RPN 36. As the technical corrective measure, we determine tasks to protect the Gateway to eliminate the possibility of failure. To avoid the possible error that data will not be transmitted, we implement a query to the ECU if data flow is available. To avoid a distribution in the data transmission, we implement timestamp sanity checks and to avoid errors in the power connection we implement continuous cable sheathing checking. As socio-technical corrective measures we determine the following actions to increase the safety level of the system Gateway. To avoid errors that data will not be transmitted we train the employees on how to recognize that data lines are broken and require a response. To avoid a distribution in the data transmission we implement alerts that show when timestamps are incorrect and train the employees to pay attention to these alerts. To avoid errors in the power connection, we implement continuous cable sheathing checking.

3.7 Electronic Control Unit

For a graphical depiction of the result of the FMECA for the object camera see Figure 8. When we connect the Electronic Control Unit (ECU) to the FMECA, we identified as possible errors a mechanical damage, a big in the control software, a software update error, and errors in the power connection.

current (respectively planned state) object Electronic Control Unit ECU				
possible errors	mechanical damage	bug in the control software	software update error	lost power connection to the ECU
possible technical effects	broken ECU	crashed ECU	control function of the ECU is not fulfilled	no response from ECU
possible error reasons	solder joints loosened, damaged by thermal effects (e.g. flying sparks, liquid steel)	incorrect input validation (e.g. readings from sensors outside the expected range)	software-update	damaged power cable mechanical damage by a person
estimation of the potential danger as well as possible measures for error prevention				
occurrence (O)	3	4	4	3
severity (S)	9	7	9	8
detection (D)	6	6	7	5
risk priority number (RPN)	**162**	**168**	**252**	**120**
technical error correction measure	chenage position, use covers	defensive programming, intensive testing (fuzzing, random, etc.)	functional check after software updates with the staggered-update-system	install power cable alternatively, redundant power supply
socio-technical error correction measure	at the persona level: Detect mechanical damage and damage in the solderings, initiate repair	training for the software developers, integration of acceptance procedures	at the persona level: event-related controls	check the cable sheathing for visible damage

Figure 8: Electronic Control Unit object from FMECA

The possible technical effects of the error mechanical damage are broken ECU. The technical effect of the error bug in the control software is a crashed ECU. The technical effect of a software update error is that the ECU cannot fulfil their control function and the technical effect of an error in the power connection is that there is no

response from the ECU. Possible reasons for the above-mentioned errors are loosened solder joints, incorrect input validations, wrong software updates and damages on the power cable. Regarding the estimation of the potential danger as well as possible measures for error prevention we evaluate the error possibilities with the metrics occurrence (O), severity (S) and detection (D). For each of them we developed a metric corresponding to the potential occurrences of errors as well as to the risk of severity and detection. Doing this we received the RPNs from left to right RPN 162, RPN 168, RPN 252 and RPN 120. For the technical corrective measure, we determine tasks to protect the ECU to eliminate the possibility of failure. In this case we determine a safer installation for the ECU and/or the usage of covers, defensive programming and intensive testing, a functional check after software updates with the staggered-update-system and a safer installation of the power connections. As socio-technical corrective measures we determine the following actions to increase the safety level of the system ECU. To avoid errors in the calibration we define table of values for calibration and develop training measures on this table/knowledge. To avoid mechanical damages, we define scheduled controlling and train the Persona(s). To avoid a crashed ECU, we implement trainings for the software developers and integrate of acceptance procedures. To avoid software update errors, we implement event-related controls at the Persona level and to avoid errors in the power connection we implement continuous cable sheathing checking.

3.8 *Combining multiple sensor data for safety autonomous driving*

In the previous chapter multiple sensor types were pointed out. Cameras, RTK GPS , LiDAR and IMUs were picked as an example to clarify the need of different types of sensors for localization and obstacle detection. Especially, in autonomous/automated (railway) vehicles sensors represent a vital factor when it comes to perceiving surroundings and position detection. From a safety perspective there must also be a redundancy within these components in case of failure. The components must be either present multiple times or another component can take over in terms of failure. However, this also requires ways to identify failures, on the one hand, and to be able to correct them, on the other hand. Therefore, data from multiple sensors must be combined in order to enable thorough decision making. This is also called sensor fusion [32]. With sensor fusion data from multiple sensors can be analysed and incorporated. In our case information is exchanged with the infrastructure over a gateway. Thus, yielding additional data for the sensor fusion to compute. Moreover, discrepancies between sensors can be identified and corrected. This way, information from contradicting data sources can be identified and data from various sources enables precise decision making.

4 Development of training materials

4.1 *Persona-Roberta model*

The interaction between humans and machines is illustrated by the Persona-Roberta model in Figure 9). It shows, what interactions there are between the systems that have to work together in automated operation and do flawlessly so as not to put anyone in danger. On one side are Personas that act as a model and in the analysis of the possible social factors that can act in the cooperation. These include social categories that characterise a person, an employee, such as his or her age and language, but also his or her professional experience or position in the company. All these factors have an influence on their actions and their understanding of the overall system, but also on their willingness to learn and develop, which is necessary as technical and organisational systems evolve. On the other hand, the Roberta model helps to understand which different technical systems are used, which knowledge is required to operate and understand them and which errors can also occur. This knowledge about the technical processes and systems is necessary to recognise errors and to prevent them before they cause damage. In this field of the model, we have to link it with the FMECA to find solutions for a safe operation at any time. These insights help to connect the measures for error elimination with concrete human knowledge and to take appropriate measures. Critical errors can only be eliminated on a conditional technical basis with suggestions in FMECA. This modelled mapping can be used for the development of error correction measures related to the FMECA as well as for the development of training measures. These should make particular reference to the interfaces and interactions between these systems and design training measures accordingly.

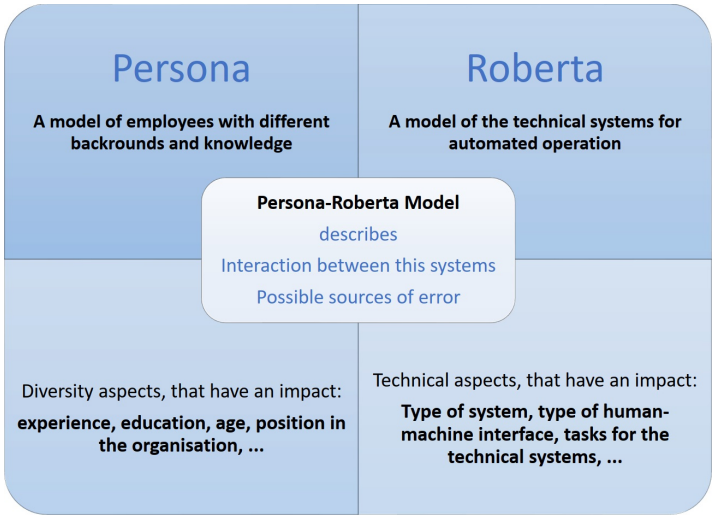


Figure 9: Persona-Roberta model

4.2 *On-the-job training*

Based the theoretical groundwork presented in Section 2 and expert discussions with the industry partners, a training concept was developed that empowers people to gain confidence in the machine and to make the right decisions based on their experience. For the development of the first draft of the training concept, learning theory principles and organisational learning as well as the generated error correction measures from the FMECA were used. Thus, a sound safety concept is a basis of the training concept that helps to comply with safety standards. Furthermore, responsibilities are clearly addressed to persons who bring in existing experience, but

who can also be trained specifically for new requirements.

In order to be able to develop training measures as precisely as possible, it needs:

- Document analysis of work instructions
- Overview of safety guidelines
- Analysis of current training folders
- Interviews or participatory observations

It is also clear that in system-theoretical adult education or systemic-constructivist didactics, no lectures and conversions [33] should happen, but to put self-direction and the learning of the individual person as knowledge carrier in the centre. This is also about the possibility of co-developing training measures [34].

It can also be recommended that training should include a mix of formal and informal teaching steps and that "learning on-the-job" should be established. In any case, any documents produced should be continuously revised in order to incorporate technical changes / further developments, but also to bring new teaching methods and knowledge of the knowledge bearers into training documents. Building on the knowledge management approach to learning organisations, the following initial concept of training or planned documentation was developed. In [35], the author have bridged organisational development and organisational learning with the field of enterprises as For-Profit-Organisations and their highly structured and determined process for learning environments. In [36], the author calls for the necessity of addressing formal and informal power as an aspect of advanced internal trainings, which also seems relevant when addressing possible resistance to new technologies by employees. In [37], the author points out that the organisational identity of an enterprise is also relevant – in our case a self-reference as "innovative", "participative", "learning organisation" or "modern". Research in this area has shown that up to now, the full potential of informal learning processes and tacit knowledge of employees, as well as participative methods have not been tapped by internal trainings.

Identify people who have knowledge and those who need to be trained and create appropriate materials. These have the focus of stimulating learning, learning to solve complex problems together and presenting them through simple language and symbols. On-the-job trainings for employees should be designed in a way that they address their specific worries, needs, state of knowledge and enable them to participate within the design process and to bring in their ideas to reduce risk factors – including oral and written didactic methods as well as thinking-out-loud processes to gain access to implicit and tacit knowledge. Elderly workers or users are often seen as digital migrants and a tackle for implementing new technologies. However, we call for applying ethnographical observations of experienced operators to sustain and involve treasure trove of their experience in the development of new technologies. Thus, a modern understanding of innovation which allows modifications during the process is vital. Technical methods like FMECA should be combined with sociological methods like Design Thinking. For the training settings, management tools are also presented to create an understanding of plan-do-check-act [38].

4.3 Tools and settings for developing the trainings and materials

How do people learn in organisations?

One important point is to use existing tacit knowledge of experienced employees and to take it seriously. There are different tools from knowledge management to capture this knowledge. Helpful are: Expert discussions, how-to listens, but also concepts such as micro-articles, which help to prepare knowledge clearly for other people [39]. The selection of interviewees is based on the identified Persona(s) and the concrete contact persons derived from them as well as the work areas identified in the FMECA. Narrative interviews are conducted with these people to help capture their experiences and knowledge of the processes and activities. In this way, it becomes clear which activities and actions are currently being carried out and what changes will occur as a result of automated operation. After this acidity the development of basic materials and trainings settings can start. This should be linked with the results of the FMECA to include possible dangerous situations and solutions to solve it.

In general it is important to define and communicate for update intervals, responsible persons and admins, including an annual training plan. This plan must be communicated by the management and the technical experts of the industry partner. All these steps must be linked to the time requirements for updates and many more. In the training documents, documents are continuously developed and simple training settings are set.

The following framework conditions are planned in the research project: The focus is on knowledge in the organisation and a transfer of knowledge. Documents and a handbook will be developed for the training courses with knowledge carriers, which can be continuously developed and adapted to the current technical requirements. The focus of the developed documents will be on using simple language and clear symbols in order to be able to communicate the complex process of self-driving operation to all persons equally. Furthermore, it will be addressed that not all persons need to know everything, but are comprehensively and well trained in their fields of work.

The following learning concepts are used for this purpose: Self-directed learning [40], as this concept considers the individual's own learning speed and prior knowledge. Basic documents on paper but also on videos are provided and simple tests to check knowledge. In addition to self-directed learning, this teaching concept also allows individual parts of the documents / videos to be easily expanded or exchanged. "The benefits of self-directed learning can be described in an effective manner in terms of the types of learners it develops. The self-directed learners demonstrate a greater awareness in terms of responsibilities." [40] Beside the self-directed learning concept, a peer-learning concept will be used. In the practice settings, learners practice in tandems. In particular, documentation sheets are filled in and reflected on as examples. Focus on the presentation and explanation of all components and their connections, as well as the possible errors. It is also planned to engage specially trained persons (admins), which are responsible for an overview on the whole content and new requirements. The goal of all measures is to train responsibility through learning theoretical expansion! For this purpose, a review and adaptation of all created documents is required at least every 12 months when changing the sensor or increasing

knowledge in real operation. Furthermore, it must also be a matter of discussing planning with staff and checking for understanding. This can happen with small online tests that the person has to do or by independently handling a possible error correction action. This can be documented by the training management with the help of an internship or observation sheet and any retraining can be proposed. In this way, the entire team can develop further.

For the development of initial materials and didactic concepts for in-company training in relation to new work processes through automated rail-bound driving on the premises, the basic work was used in order to make custom-fit proposals from a socio-scientific-technical point of view. These are, as shown: FMECA - Persona-Roberta model - Building blocks of knowledge management - Foundations of organisational learning.

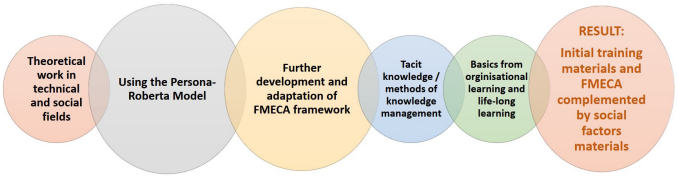


Figure 10: Combination of tools for training material development

As Figure 10 illustrates, there is no simple implementation guide, but there are many tools and methods that can be used to ensure smooth and flawless operation of autonomous rail-bound vehicles in the industry. The tools shown can be used to further develop and use them according to one's own requirements. This is achieved through the cooperation of different disciplines that contribute their technical and social expertise.

These tools and methods were supplemented by the basics of adult education and e-learning, since parts of the planned further education are to function online and with the concept of self-directed learning in the future. All documents are to be developed and reviewed in consultation with the Human Resources, Technical Development and Hazard Management departments. It will also be clearly worked out who is responsible for the final design of all documents and course units, who is responsible for the individual training parts, who adapts and modernises them accordingly if something changes in the sequence of events. It can be assumed that this technology will continue to develop and, therefore, documents, as well as tools and concepts, must be constantly updated. These aspects point to the future and also show the challenges of future research work. It is clear that concepts of adult education and a managing diversity concept [41] should be taken as a basis. People with their knowledge and social characteristics are thus perceived in their entirety as a relevant resource and integrated into knowledge processes. This also closes the circle to knowledge management, which had been used in the research project in particular to raise the explicit knowledge of the employees. How do adults learn? There is no one-size-fits-all answer, but there are frameworks that help to design one's own learning environment in order to be able to communicate new technologies that bring about changes in work processes. The needs of adults, which can motivate them to learn, must be as precisely as possible and relate them to the respective needs (qualification profiles, etc.) [42]

If we combine these starting points with the aspects of organisational learning, managing diversity theory and knowledge management that have already been worked out, they provide sufficient orientation for the selection of didactic methods and the targeted preparation of content and learning materials. Once again, reference should be made to the diversity of learners in an industrial company and to the requirements of clear languages, unambiguous images and small snippets of information that will make it possible to organise the interaction between man and machine with the most well-founded knowledge possible in a safe and error-free manner. A regular evaluation of the success of the training courses with regard to possible incidents complements the overall approach of integrating social and socio-scientific aspects into technical training.

5 Conclusion and further research

The proposed set of tools and models can contribute to more appropriate training materials for humans interacting with machines, in the area of autonomous railway vehicles. Especially, the extended FMECA participates and provides a holistic approach to what a railway operating system should look like and shows which requirements, influences and interfaces must be considered. Additionally, it is an effective technique to spot security as well as safety concerns at a point where planned systems and human actors interact at the very beginning of the design process. However, the descriptions of the subcomponents of the component autonomous railway vehicle in section 3, indicates a strong relationship between the technical and sociological error correction measures. Regarding the estimation of the potential danger as well as possible measures for error prevention we evaluated the error possibilities with the metrics occurrence (O), severity (S) and detection (D). For each of them, a metric was developed that is corresponding to the potential occurrences of errors as well as to the risk of severity and detection in several rounds based on the knowledge of the entire project team. A multitude of tools and models are then implemented to identify and develop adequate training materials. These are based on theoretical work in technical and social fields, the Persona-Roberta model, as a model focusing on interaction between humans and machines, tacit knowledge and methods of knowledge management as well as basics of organisational and life-long learning. We propose that together these tools will help and enhance developing more adequate training materials for the autonomous railway sector. In a further step, these materials should then be generated based on the proceedings that have been described and applied to a genuine context of an autonomous railway system.

The novelty of the research and the article lies in the combination of the different research methods and models, which complement each other and combine technical and social issues. It is well known from knowledge management and organizational development that the introduction of new technical processes that bring innovation to operations and processes requires knowledge about the people who will be affected by them. Therefore, research questions such as "What makes an organizational culture that produces technical and social innovations?" but also "How does the use of the tacit knowledge of the employees succeed?" For that the research team used methods from organizational learning including aspects of

safety and security and of knowledge management. To stick social and technical aspects together the method of "personas" from design thinking processes was extended with the technical aspects – "Roberta". Furthermore, the model of [43] for organizational learning was included.

In the research projects as well as in this article, it could be shown that it is relevant and important to be aware of technical expert knowledge. And besides there must also be knowledge about social factors and social negotiation processes in an organisation, which have an effect on the introduction and operation of new technical systems. It can be stated that interdisciplinary cooperation can deepen the processing of the interaction between machines and humans. The use and further development of known instruments such as the FMECA and models such as the Persona-Roberta model enables a structured derivation of possible training measures and contents. The existing knowledge of (long-term) employees must be used as a basis in order to be able to integrate informal experiential knowledge specifically into the selection of the learning setting and the preparation of the materials. Existing hazard plans in the organisations and technical knowledge complement the development and help to establish safe operations. In summary, it can be said that this contribution shows where research and development must go in the future, in order to further develop autonomous rail-bound driving on company premises, both technically and socially. The presented research results have already been implmented in teachings in a variety of departments at the University of Applied Scienes Burgenland. And the research partners further suggest on implementing the results in upcoming research projects in the field of autonomous railway systems and others with focus on sustainable nationwide mobilty. We further suggest that the implementation of social aspects can aid in developing such a system and support the further mitigation of failures by applying and developing adequate training materials.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment Research leading to these results has received funding from the project BESTE-AB, funded by the Austrian Research Promotion Agency (FFG), coordinated by the Austrian Institute of Technology.

References

[1] C. Gnauer, A. Prochazka, E. Szalai, S. Chlup, S. Luimpöck, A. Fraunschiel, C. Schmittner, "Knowledge Based Training Derived from Risk Evaluation Concerning Failure Mode, Effects and Criticality Analysis in Autonomous Railway Systems," in 2021 5th International Conference on System Reliability and Safety (ICSRS), 47–52, 2021, doi:10.1109/ICSRS53853.2021.9660726.

[2] D. Hofbauer, C. Schmittner, M. Brandstetter, M. Tauber, "Autonomous CPS Mobility Securely Designed," in 2019 IEEE 20th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM), 1–6, 2019, doi:10.1109/WoWMoM.2019.8793050.

[3] D. Walker, F. Myrick, "Grounded theory: An exploration of process and procedure," Qualitative health research, **16**(4), 547–559, 2006, doi:doi:10.1177/1049732305285972.

[4] S. Kato, E. Takeuchi, Y. Ishiguro, Y. Ninomiya, K. Takeda, T. Hamada, "An open approach to autonomous vehicles," IEEE Micro, **35**(6), 60–68, 2015.

[5] N. Anwar, M. S. Khan, K. Ahmed, A. Ahmad, A. Athar, "Speed scheduling of autonomous railway vehicle control system using ANN," International Journal of Scientific & Engineering Research, **2**(6), 1–6, 2011.

[6] A. Ahmad, M. S. Khan, K. Ahmed, N. Anwar, A. Athar, "Speed Scheduling of Autonomous Railway Vehicle Control System using Neuro-Fuzzy System," International Journal of Scientific and Engineering Research, **2**(6), 272–277, 2011.

[7] A. Ahmad, M. S. Khan, K. Ahmed, N. Anwar, U. Farooq, "FIS Based Speed Scheduling System of Autonomous Railway Vehicle," International Journal of Scientific and Engineering Research, **2**(6), 299–304, 2011.

[8] P. Singh, M. A. Dulebenets, J. Pasha, E. D. S. Gonzalez, Y.-Y. Lau, R. Kampmann, "Deployment of autonomous trains in rail transportation: Current trends and existing challenges," IEEE Access, **9**, 91427–91461, 2021.

[9] "Rise of the machines: Rio Tinto breaks new ground with AutoHaul," 2019.

[10] M. Rahimi, H. Liu, I. D. Cardenas, A. Starr, A. Hall, R. Anderson, "A Review on Technologies for Localisation and Navigation in Autonomous Railway Maintenance Systems," Sensors, **22**(11), 2022, doi:10.3390/s22114185.

[11] R. Lagay, G. M. Adell, "The autonomous train: A game changer for the railways industry," in 2018 16th international conference on intelligent transportation systems telecommunications (ITST), 1–5, IEEE, 2018, doi:10.1109/ITST.2018.8566728.

[12] M. Shafiee, F. Dinmohammadi, "An FMEA-based risk assessment approach for wind turbine systems: a comparative study of onshore and offshore," Energies, **7**(2), 619–642, 2014, doi:10.3390/en7020619.

[13] L. S. Lipol, J. Haq, "Risk analysis method: FMEA/FMECA in the organizations," International Journal of Basic & Applied Sciences, **11**(5), 74–82, 2011.

[14] J. Singh, S. Singh, A. Singh, "Distribution transformer failure modes, effects and criticality analysis (FMECA)," Engineering Failure Analysis, **99**, 180–191, 2019, doi:https://doi.org/10.1016/j.engfailanal.2019.02.014.

[15] M. Catelani, L. Ciani, D. Galar, G. Guidi, S. Matucci, G. Patrizi, "FMECA assessment for railway safety-critical systems investigating a new risk threshold method," IEEE Access, **9**, 86243–86253, 2021, doi:10.1109/ACCESS.2021.3088948.

[16] F. Dinmohammadi, B. Alkali, M. Shafiee, C. Bérenguer, A. Labib, "Risk evaluation of railway rolling stock failures using FMECA technique: a case study of passenger door system," Urban Rail Transit, **2**(3), 128–145, 2016, doi:10.1007/s40864-016-0043-z.

[17] J. Kim, H. Jeong, J. Park, "Development of the FMECA process and analysis methodology for railroad systems," International Journal of Automotive Technology, **10**(6), 753–759, 2009.

[18] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS quarterly, 319–340, 1989, doi:10.2307/249008.

[19] M. Lewrick, P. Link, L. Leifer, N. Langensand, Das Design Thinking Playbook: mit traditionellen, aktuellen und zukünftigen Erfolgsfaktoren, Vahlen, 2018.

[20] M. Lewrick, P. Link, L. Leifer, A. Schmidt, Das Design Thinking Toolbook: Die besten Werkzeuge & Methoden, Franz Vahlen, 2019.

[21] D. J. Yeong, G. Velasco-Hernandez, J. Barry, J. Walsh, "Sensor and Sensor Fusion Technology in Autonomous Vehicles: A Review," Sensors, **21**(6), 2021, doi:10.3390/s21062140.

[22] Y. Feng, J. Wang, "GPS RTK Performance Characteristics and Analysis," Journal of Global Positioning Systems, **7**, 2008, doi:10.5081/jgps.7.1.1.

[23] "What is the Difference Between RTK & DGPS?" 2019.

[24] F. Van Diggelen, "GPS and GPS+ GLONASS RTK," in Proceedings of the 10th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1997), 139–144, 1997.

[25] N. Mehendale, S. Neoge, "Review on Lidar Technology," 2020, doi:10.2139/ssrn.3604309.

[26] M. A. Lefsky, W. B. Cohen, G. G. Parker, D. J. Harding, "Lidar Remote Sensing for Ecosystem Studies: Lidar, an emerging remote sensing technology that directly measures the three-dimensional distribution of plant canopies, can accurately estimate vegetation structural attributes and should be of particular interest to forest, landscape, and global ecologists," BioScience, **52**(1), 19–30, 2002, doi:10.1641/0006-3568(2002)052[0019:LRSFES]2.0.CO;2.

[27] "What is IMU?" 2021.

[28] N. Ahmad, R. A. R. Ghazilla, N. M. Khairi, V. Kasi, "Reviews on Various Inertial Measurement Unit (IMU) Sensor Applications," International Journal of Signal Processing Systems, 256–262, 2013, doi:10.12720/ijsps.1.2.256-262.

[29] "Inertial Measurement Units for Localisation and Navigation of Autonomous Vehicles," 2019.

[30] C. P. Kruger, A. M. Abu-Mahfouz, G. P. Hancke, "Rapid prototyping of a wireless sensor network gateway for the internet of things using off-the-shelf components," in 2015 IEEE International Conference on Industrial Technology (ICIT), 1926–1931, IEEE, 2015, doi:10.1109/ICIT.2015.7125378.

[31] R. He, B. Ai, G. Wang, K. Guan, Z. Zhong, A. F. Molisch, C. Briso-Rodriguez, C. P. Oestges, "High-speed railway communications: From GSM-R to LTE-R," Ieee vehIcular technology magazIne, **11**(3), 49–58, 2016, doi:10.1109/MVT.2016.2564446.

[32] S. Ataei, A. Aghakouchak, M. Marefat, S. Mohammadzadeh, "Sensor fusion of a railway bridge load test using neural networks," Expert Systems with Applications, **29**(3), 678–683, 2005, doi:https://doi.org/10.1016/j.eswa.2005.04.038.

[33] R. Arnold, H. Siebert, "Konstruktivistische Erwachsenenbildung. Von der Deutung zur Konstruktion der Wirklichkeit. Baltmannsweiler," 2003.

[34] S. Nolda, Einführung in die Theorie der Erwachsenenbildung, WBG Darmstadt, 2012.

[35] N. Engel, S. Koch, "Betrieb als Gegenstand und Ort organisationspädagogischer Forschung und Praxis," in Handbuch Organisationspädagogik, 793–804, Springer, 2018, doi:10.1007/978-3-658-07512-5_77.

[36] P. Faulstich, Strategien der betrieblichen Weiterbildung: Kompetenz und Organisation, Vahlen, 1998, doi:10.1007/s11618-000-0016-5.

[37] S. Kirchner, "Organisationsidentität und Unsicherheit," in Organisation und Unsicherheit, 69–85, Springer, 2015, doi:10.1007/978-3-531-19237-6.

[38] S. M. Weber, M. Göhlich, A. Schröer, H. Macha, C. Fahrenwald, "Organisation und Partizipation–interdisziplinäre Verhältnisbestimmungen und organisationspädagogische Perspektiven," in Organisation und Partizipation, 9–28, Springer, 2013, doi:10.1007/978-3-658-00450-7_1.

[39] A. Mittelmann, Werkzeugkasten Wissensmanagement, BoD–Books on Demand, 2011.

[40] R. Kapur, "Significance of Self-Directed Learning," 2019.

[41] L. Gardenswartz, A. Rowe, Managing diversity: A complete desk reference and planning guide, McGraw Hill Professional, 1998.

[42] B. Dewe, K. Feistel, Betriebliche Weiterbildung Materialien in didaktischer und bildungsökonomischer Perspektive, Franz Steiner Verlag, 2013.

[43] G. Probst, S. Raub, K. Romhardt, Bausteine des Wissensmanagements, 25–33, Gabler Verlag, Wiesbaden, 2003, doi:10.1007/978-3-322-94790-1_3.