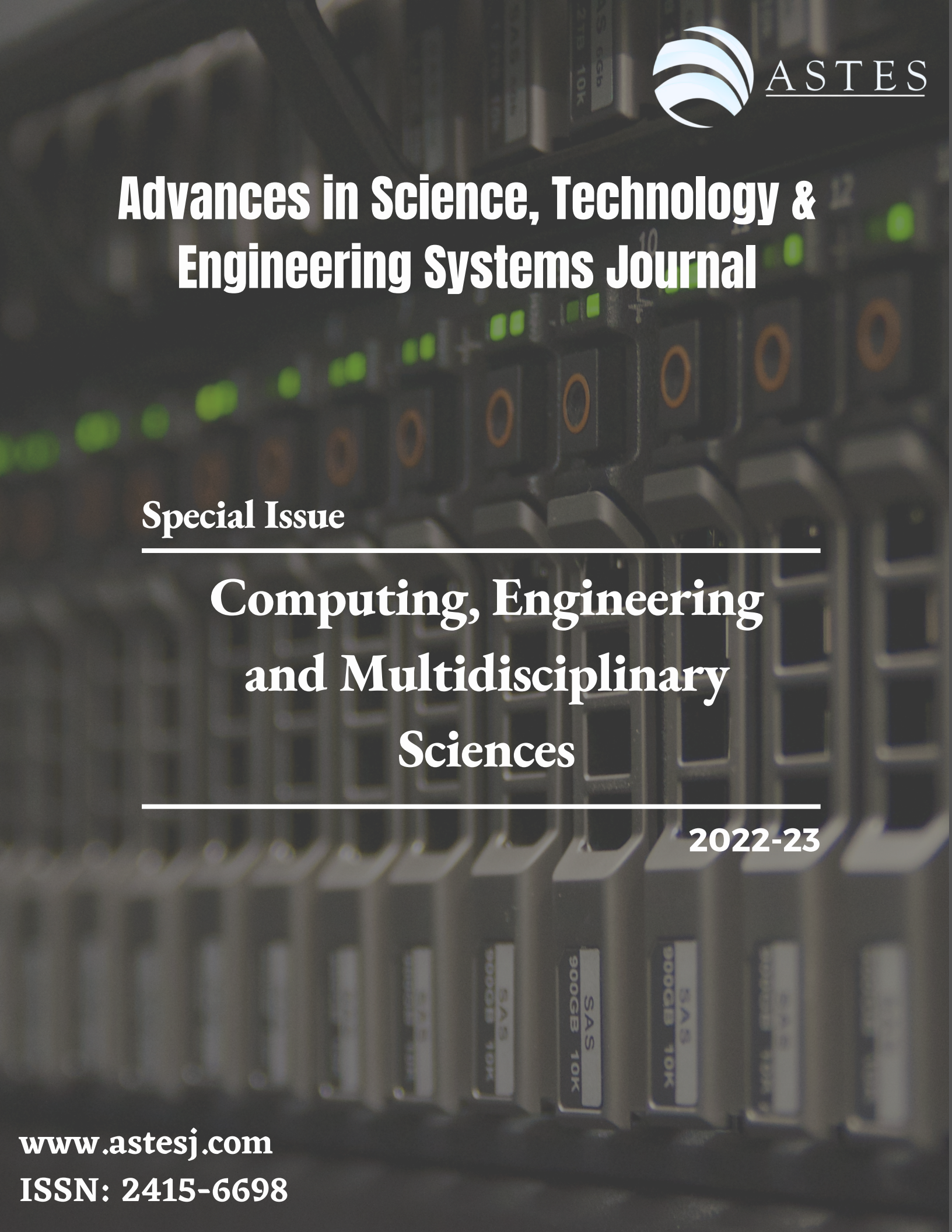


ASTES

Advances in Science, Technology & Engineering Systems Journal

Special Issue

Computing, Engineering and Multidisciplinary Sciences

2022-23

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 Computing, Engineering and Multidisciplinary Sciences (2022–23) in the *Advances in Science, Technology and Engineering Systems Journal (ASTES Journal)* reflects the deepening integration of computational technologies with engineering practices and multidisciplinary research frameworks. As the complexity of modern scientific and societal challenges continues to grow, the convergence of diverse fields has become essential for generating innovative, scalable, and impactful solutions. This issue presents a collection of contributions that exemplify how computing and engineering, when combined with insights from multiple disciplines, can address emerging demands across technological and industrial landscapes.

A prominent theme throughout this issue is the pervasive role of advanced computing techniques in transforming engineering systems. Contributions explore the application of artificial intelligence, machine learning, data analytics, and intelligent control systems in optimizing performance, enhancing reliability, and enabling predictive decision-making. These approaches are applied across a wide spectrum of domains, including smart manufacturing, communication networks, energy systems, and automation. The research highlights how computational intelligence is not only augmenting engineering capabilities but also fostering new paradigms of system design and operation.

The multidisciplinary nature of this issue is further reflected in its engagement with sustainability, resilience, and innovation. Several papers focus on environmentally conscious engineering solutions, renewable energy integration, and efficient resource management, emphasizing the importance of sustainable development in contemporary research. Others address challenges in infrastructure, healthcare technologies, and digital ecosystems, demonstrating how cross-disciplinary collaboration can generate holistic solutions that are both technically robust and socially relevant.

Methodological diversity is a defining strength of this collection. The studies employ a combination of theoretical frameworks, simulation-based analyses, experimental investigations, and real-world case studies. This integrative approach ensures that the research maintains academic rigor while also offering practical applicability. Many contributions emphasize scalability and adaptability, providing models and systems that can be extended across different sectors and operational contexts.

The 2022–23 timeframe offers an important backdrop characterized by rapid digital acceleration and increasing emphasis on resilient and intelligent systems. As industries and societies continue to adapt to evolving global conditions, the role of computing and multidisciplinary engineering has become more critical than ever. The works presented in this issue reflect this transition, highlighting innovations that support connectivity, automation, sustainability, and informed decision-making.

The editorial team expresses its sincere gratitude to the authors for their high-quality contributions and to the reviewers for their careful and constructive evaluations. Their dedication has ensured the integrity and scholarly value of this special issue, reinforcing the journal's commitment to excellence in multidisciplinary research.

This special issue demonstrates the transformative potential of integrating computing, engineering, and multidisciplinary sciences. By fostering collaboration across diverse domains

and presenting forward-looking research, it contributes meaningfully to the advancement of knowledge and the development of innovative solutions for complex global challenges..

Guest Editor

Prof. María Jesús Espinosa Trujillo

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Special Issue **October 2022**

CONTENTS

Markov Regime Switching Analysis for COVID-19 Outbreak Situations and their Dynamic Linkages of German Market
by Kangrong Tan and Shozo Tokinaga

Characterization and Investigating the Effect of Gate-Insulator Thickness on Co-Axial Cylindrical Carbon Nanotube Field Effect Transistor
by Suchismita Sen, Argha Sarkar and Pinaki Chakraborty

Design of an Open Source Anthropomorphic Robotic Hand for Telepresence Robot
by Jittaboon Trichada, Traithep Wimonrut, Narongsak Tirasuntarakul and Eakkachai Pengwang

Localization of Impulsive Sound Source in ShallowWaters using a Selective Modal Analysis Algorithm
by Faraz Talebpour, Saeed Mozaffari, Mehrdad Saif and Shahpour Alirezaee

Inferring Student Needs Based on Facial Expression in Video Images
by Yu Yan, Eric Wallace Cooper and Richard Lee

Temperature-Compensated Overcharge Protection Measurement Technology
by Jin Uk Yeon, Ji Whan Noh and Innyeal Oh

Measurement System for Evaluation of Radar Algorithms using Replication of Vital Sign Micro Movement and Dynamic Clutter
by Christoph Domnik, Daniel Erni and Christoph Degen

A Multiplatform Application for Automatic Recognition of Personality Traits in Learning Environments
by Víctor Manuel Bátiz Beltrán, Ramón Zatarain Cabada, María Lucía Barrón Estrada, Héctor Manuel Cárdenas López and Hugo Jair Escalante

Analysis and Trend Estimation of Rainfall and Seasonality Index for Marathwada Region
by Himanshu Bana and Rahul Dev Garg

Fuzzy MPPT for PV System Based on Custom Defuzzification
by Abdelmadjid Allaoui, Mohamed Nacer Tandjoui and Chellali Benachaiba

Indoor Positioning: Comparing Different Techniques and Dealing with a user Authentication use Case
by Joaquín Pérez Balbela and Aruna Prem Bianzino

A Circuit Designer's Perspective to MOSFET Behaviour: Common Questions and Practical Insights

by Ralf Sommer, Carsten Thomas Gatermann and Felix Vierling

Hybrid Intrusion Detection Using the AEN Graph Model
by Paulo Gustavo Quinan, Issa Traoré, Isaac Woungang, Ujwal Reddy Gondhi and Chenyang Nie

Nonlinear Model Predictive Control of Rover Robotics System
by Serdar Kalaycioglu and Anton de Ruiter

Navigation Aid Device for Visually Impaired using Depth Camera
by Hendra Kusuma, Muhammad Attamimi and Julius Sintara

Forecasting the Weather behind Pa Sak Jolasid Dam using Quantum Machine Learning
by Chaiyaporn Khemapatapan and Thammanoon Thepsena

Proportional Derivative and Proportional Integral Derivative Controllers for Frequency Support of a Wind Turbine Generator in a Diesel Generation Mix
by Abdul Ahad Jhumka, Robert Tat Fung Ah King, Chandana Ramasawmy and Abdel Khoodaruth

Omni-directional Multi-view Image Measurement System in the Co-sphere Framework
by Yung-Hsiang Chen and Jin H. Huang

Distribution Management Problem: Heuristic Solution for Vehicle Routing Problem with Time Windows (VRPTW) in the Moroccan Petroleum Sector
by Younes Fakhradine El Bahi, Latifa Ezzine, Zineb Aman, Imane Moussaoui, Miloud Rahmoune and Haj El Moussami

Analysis of Linear and Non-Linear Short-Term Pulse Rate Variability to Evaluate Emotional Changes during the Trier Social Stress Test
by Alvin Sahroni, Isnatin Miladiyah, Nur Widiasmara and Hendra Setiawan

Detecting the Movement of the Pilot's Body During Flight Operations
by Yung-Hsiang Chen, Chen-Chi Fan and Jin H. Huang

Day-Ahead Power Loss Minimization Based on Solar Irradiation Forecasting of Extreme Learning Machine
by Adelhard Beni Rehiara, Sabar Setiawidayat and Frederik Haryanto Sumbung

On the Polytopic Modelling & Robust H∞ Control of Nonlinear Systems Subject to Cyber-attack: Application to Attitude Stabilization of Quadrotor
by Bezzaoucha-RebaÏ Souad

Human-Centered Design, Development, and Evaluation of an Interface for a Microgrid Controller
by Mohammed Mahfuz Hossain, Thomas Ortmeyer and Everett Hall

Development and Analysis of Models for Detection of Olive Trees
by Ivana Marin, Sven Gotovac and Vladan Papić

HistoChain: Improving Consortium Blockchain Scalability using Historical Blockchains
by Marcos Felipe and Haiping Xu

The First Application of the Multistage One-Shot Decision-Making Approach to Reevaluate a Technology Project Decision Problem
by Mohammed Al-Shanfari

Hybrid Machine Learning Model Performance in IT Project Cost and Duration Prediction
by Der-Jiun Pang

Hybrid Discriminant Neural Networks for Performance Job Prediction
by Temsamani Khallouk Yassine, Achchab Said, Laouami Lamia and Faridi Mohammed

Metaheuristic Optimization Algorithm Performance Comparison for Optimal Allocation of Static Synchronous Compensator
by Abdulrasaq Jimoh, Samson Oladayo Ayanlade, Emmanuel Idowu Ogunwole, Dolapo Eniola Owolabi, Abdulsamad Bolakale Jimoh and Fatina Mosunmola Aremu

Social Financial Technologies for the Development of Enterprises and the Russian Economy
by Evgeniy Kostyrin and Evgeniy Sokolov

Assessment of Scattered-Bend Loss in Polymer Optical Fiber (POF) Displacement Sensor
by Latifah Sarah Supian, Danial Haikal Mohd Razali, Chew Sue Ping, Nurul Sheeda Suhaimi, Sharifah Aishah Syed Ali, Nani Fadzlina Naim and Harry Ramza

Detecting CTC Attack in IoMT Communications using Deep Learning Approach
by Mario Cuomo, Federica Massimi and Francesco Benedetto

Active Simulation of Grounded Parallel-Type Immittance Functions Employing VDBAs and All Grounded Passive Components
by Pratya Mongkolwai, Pitchayanin Moonmuang, Worapong Tangsrirat and Taweepol Suesut

Tunable Resistorless Phase Shifter Realization with a Single VDGA
by Orapin Channumsin, Jirapun Pimpol, Tattaya Pukkalanun and Worapong Tangsrirat

A Model for Teaching Mathematics to Gifted Students Based on an Effective Combination of Various Approaches for their Preparation
by Zhanna Dedovets, Mikhail Rodionov and Anna Novichkova

Design and Comparative Analysis of Hybrid Energy Systems for Grid-Connected and Standalone Applications in Tunisia: Case Study of Audiovisual Chain
by Saidi Mohamed, Habib Cherif, Othman Hasnaoui and Jamel Belhadj

Photoluminescence Properties of Eu(III) Complexes with Two Different Phosphine Oxide Structures and Their Potential uses in Micro-LEDs, Security, and Sensing Devices: A Review
by Hiroki Iwanaga

Design and Implementation of an Automated Medicinal-Pill Dispenser with Wireless and Cellular Connectivity
by Chanuka Bandara, Yehan Kodithuwakku, Ashan Sandanayake, R. A. R. Wijesinghe and Velmanickam Logeeshan

Smart Healthcare Kit for Domestic Purposes
by Yehan Kodithuwakku, Chanuka Bandara, Ashan Sandanayake, R.A.R Wijesinghe and Velmanickam Logeeshan

A Review of the Role of Information Technology in Brazilian Higher Educational Institutions during Covid-19 Pandemic
by Luís Cláudio Dallier Saldanha

Hybrid Neural Network Method for Predicting the SOH and RUL of Lithium-Ion Batteries
by Brahim Zraibi, Mohamed Mansouri, Salah Eddine Loukili and Said Ben Alla

Investigation of Swimming Behavior and Performance of the Soft Milli-Robots Embedded with Different Aspects of Magnetic Moments
by Xiuzhen Tang and Laliphat Manamanchaiyaporn

Three-phase Continuously Variable Series Reactor – Realistic Modeling and Analysis
by Mohammadali Hayerikhiyavi and Aleksandar Dimitrovski

How a Design-Based Research Approach Supported the Development and Rapid Adaptation Needed to Provide Enriching Rural STEM Camps During COVID and Beyond
by Rebecca Zulli Lowe, Adrienne Smith, Christie Prout, Guenter Maresch, Christopher Bacot and Lura Murfee

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 202-211 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Three-phase Continuously Variable Series Reactor – Realistic Modeling and Analysis

Mohammadali Hayerikhiyavi*, Aleksandar Dimitrovski

Department of Electrical and Computer Engineering, University of Central Florida, Orlando, 32816, USA

ARTICLE INFO

Article history:
Received: 28 February, 2023
Accepted: 11 May, 2023
Online: 25 June, 2023

Keywords:
Continuously Variable Series Reactor (CVSR)
Magnetic Amplifier
Gyrator-Capacitor model
Hysteresis
DC bias control

ABSTRACT

Continuously Variable Series Reactor (CVSR) is a device that can vary the reactance in an ac circuit by controlling the magnetization of a ferromagnetic core, shared by ac and dc windings. The bias dc current can change the equivalent ac reactance(inductance) in order to, for example, control load flow, damp oscillations, or fault current limitation. Gyrator-Capacitor (G-C) approach in modeling electromagnetic devices provides a strong and practical way in simulating an integrated system composed of magnetic and electric/electronic circuits. The G-C model provides key advantages in analysis of electromagnetic devices, including CVSR. Understanding the performance and the operational characteristics of the CVSR is essential for its proper utilization in the power grid. This paper presents a detailed G-C approach that includes the ferromagnetic core nonlinearities, namely, hysteresis and saturation. The approach has been applied in modeling the electromagnetic coupling between the ac and dc circuits of a three-phase CVSR. Analysis of the effect of different control dc circuit types on the equivalent ac inductance is presented, during operating conditions at different ferromagnetic states. In addition, induced voltages across the windings and the power exchange with the control circuit are presented.

1. Introduction

This article augments the underlying work introduced in the 2022 IEEE Kansas Power and Energy Conference (KPEC) [1]. The major contributions and improvements to that work included here are the following:

1) Comprehensive analysis of the performance of the CVSR in terms of the induced voltages across the dc winding, to assess the counter impact of the power system on the control circuit.

2) Modeling of hysteresis in order to capture the effect of the phenomenon on the induced voltages in the windings and make the model as realistic as possible.

3) Power exchange between the controlled and control circuits.

Contemporary power grids operate under increased stress and strain due to the growing demand for electric energy, along with the growing penetration of variable renewable sources. The primary concern system operators have in running the power system and satisfying the demand is to deals with the contingencies in generation and transmission, system oscillations and other events that may result in instabilities and result with blackouts [2]. This is mainly due to the absence of a comprehensive load flow control. Traditionally, load flow control has been implemented using phase-shifting transformers, shunt capacitors and reactors, generator controls, switching system elements on and off and, in the past few decades, various types of power electronics-based controllers [3]. However, these devices either provide only limited control or they come at very high cost. In addition, the meshed topologies of the power systems, make it quite complicated for some of these control means to be effectively employed. More recently, continuously variable series reactor (CVSR) technology was proposed as an alternative option [3-5].

CVSR is a series-connected reactor in the ac power circuit that can continuously vary its reactance within the design boundaries. It is characterized with high reliability and low maintenance, installation, and operating costs [4,6]. Continuous control is achieved by varying a bias control current. Whereas in FACTS controllers, power flows through the high-rated power electronic components, the CVSR control circuit is isolated from the power system and can use low-rated power electronic converters.

Line decongestion and overload relief are easily accomplished by the CVSR through inserting additional impedance in the power circuit in series with its ac winding. Furthermore, CVSR can also

*Corresponding Author: Mohammadali Hayerikhiyavi ; Email: Mohammad.ali.hayeri@knights.ucf.edu

https://dx.doi.org/10.25046/aj080323

be used in dynamic applications such as oscillations damping and fault currents limiting by varying the impedance accordingly. Due to the versatility of applications, it is essential to study the mutual impact of the CVSR and the power grid under different operating conditions [4,5].

Gyrator-Capacitor (G-C) modeling approach is a suitable and effective method for modeling magnetic circuits in detailed analyses of power magnetic devices. It directly links electrical and magnetic circuits for comprehensive studies of complex hybrid devices which are part of the power system. In this approach, the analogy between the magnetomotive force (MMF) and the electromotive force is preserved, however the electrical current is analogous to the rate of change of the flux. As a result, the permeance (inverse reluctance) becomes analogous to capacitance.

In a basic representation of the CVSR, as is usual, the magnetic circuit is generally modeled with no core losses, hysteresis in particular. However, some research has been done to include this effect in the analysis [7]. the Jiles-Atherton method can have convergence problems which can be detrimental in transient studies. Furthermore, in some cases, it may result with high percentage error. In [8], alternative models have been examined, with their accuracy and ease of computation compared. The study indicates that Rayleigh model offers sufficient accuracy for cores with high coercivity. On the other hand, Potter model employs simpler mathematical expressions but may lead to significant errors in certain cases. The Frölich and Preisach models provide results that are consistent with experimental findings, although they may not be suitable for dynamic analysis due to their high computation burden. It should be stated that another characteristic of these methods is parameter sensitivity which may also lead to significant errors.

The G-C model represents hysteresis by adding a resistor. The value of this resistor depends on the core geometry, and specific losses [9].

In most studies of electromagnetic devices with controlled bias flux, for simplicity, the source in the control circuit of the device is considered as an ideal current source. This, of course, is not realistic. Three other bias control sources are considered here that have quite different internal impedances. One is an ideal voltage source with zero internal impedance, as opposed to the infinite impedance of the ideal current source. The other two are typical power electronics-based sources: H-bridge converter and buck converter. Their internal impedance values are realistic, in between the two extremes of the ideal sources.

Accurate modeling and simulation of power system elements is fundamental to enhance our understanding of the behavior of the components and their interaction with the power system. This is true in general, but it is also essential in the analysis of the CVSR – a device in which the power circuit is magnetically coupled with a control circuit of much lower ratings. The usual approximations that are justified at the power system level can lead to significant errors and oversights of detrimental conditions on the sensitive controls. Hence, to investigate the impact that the power system has on the CVSR itself, it is important to take all the details into account. The goal of this paper is to provide comparison between idealistic and realistic models of the CVSR and highlight the differences in its performance as a result of the improved accuracy.

The rest of the paper is structured as follows: the concept of three-phase CVSR is discussed in Section 2. The G-C model general concept, expanded to include nonlinearities from core saturation and hysteresis, and applied to the three-phase CVSR is introduced in Section 3. Section 4 introduces the different types of bias dc sources. Section 5 provides a case study with results from simulations with different models and presents the impact on the CVSR in terms of equivalent ac inductance, induced voltages, and the power exchange with the control bias winding. The conclusions are drawn in Section 6.

2. Three-phase CVSR

A three-phase CVSR with a five-legged magnetic core, as shown in Fig. 1, includes three-phase ac winding wound on the inner legs that is part of a three-phase ac power circuit that delivers power to the load. Typically, the three inner legs have air gaps to achieve the desired nominal reactance in unsaturated conditions, and to prevent core saturation even at small ac currents. In this particular case, five dc coils are wound on each leg, connected in series and fed by a dc source. The outer legs are gapless in order to provide unimpeded return path for the bias flux. To balance the bias MMFs, the number of turns in the outer dc windings is 1.5 times the number of turns on the inner ones [1]. The controlled bias dc current creates a flux passing through the entire core, hence controlling the ac inductance of the three-phase CVSR. The induced voltages on the coils are proportional to $d\Phi_{leg}/dt$ and the voltage across the whole dc winding is their algebraic sum: $V_{bias} = \sum_{i=1}^{5} V_{dc,i}$.

The core can be heterogeneous whose specifications are taken from [1].

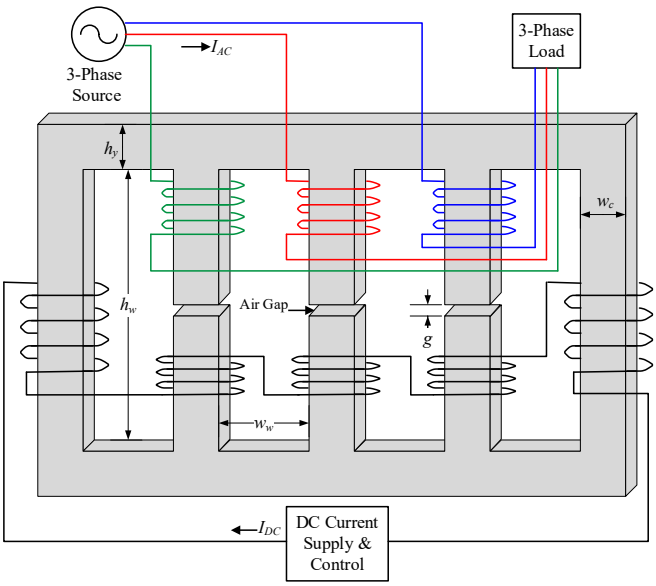


Figure 1: Three-phase CVSR [1]

3. Gyrator-Capacitor Model

3.1. General Concept

Magnetic circuits are typically represented using the electric circuit analogy [10]. The corresponding circuits are constructed using resistors. Magnetic circuits store energy and they are not

suitably modeled by resistors which only dissipate energy. The G-C modeling concept illustrated in Fig. 2 maintains the power equivalence and is energy/power invariant [11,12].

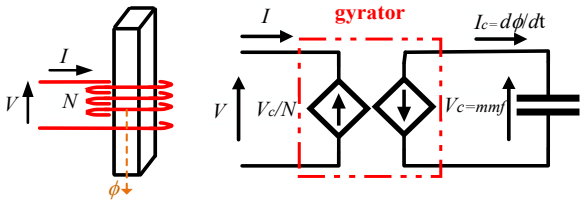


Figure 2: Magnetic circuit and its equivalent gyrator-capacitor model

These expressions lead to representation of magnetic permeances, reciprocal values of magnetic reluctances, (magnetic conductances) with capacitances. Nonlinear magnetic paths with nonlinear permeances are represented with nonlinear capacitors. Each coil is represented by a gyrator that relates voltage and current with the number of turns *N*.

The reluctances can be expressed approximately with (1):

$$\mathcal{R}(\Phi) = (\mu_r(\Phi)\mu_0 l)/A \quad (1)$$

where: $\mathcal{R}$ is the magnetic reluctance which depends in general on the magnetic flux Φ; $\mu_0 = 4\pi\times10^{-7}$ μ_r is the relative magnetic permeability of the material; A is the cross-sectional area; l is the mean length of the core path.

3.2. Hysteresis Modeling

Hysteresis is modeled on one of the most common electromagnetic devices– a two-winding transformer shown in Fig. 3. The hysteresis is modeled by adding a resistor in series with the core capacitance, which represents the magnetic circuit shared by the two windings. The modeling of hysteresis is explicit, meaning that it is reflected in the core B-H characteristic [9]. Besides modeling the ensuing hysteretic losses, more importantly, it captures the other effects of the hysteretic nonlinearity on the performance of the CVSR, as shown later in the case study.

In Fig. 3, L_1, R_1, L_2, and R_2 represent the primary and secondary winding impedances on the electrical side. The core saturation is modeled by a nonlinear capacitor on the magnetic side. This concept can be extended to the three-phase CVSR.

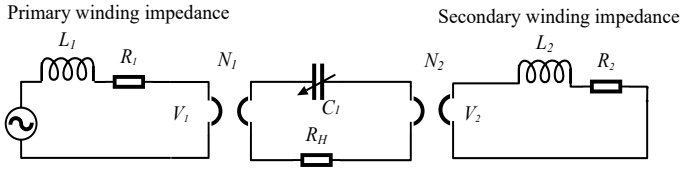


Figure 3: G-C model of a two-winding transformer with hysteretic losses

3.3. Three-phase CVSR

An improved G-C model of the physical configuration of the CVSR from Fig.1, with nonlinear core in MatLAB /Simulink® is represented in Fig. 4.

The nonlinear magnetic paths are modeled with variable capacitors, and linear permeances Cg model the air gaps in the inner legs. They can also include the fringing effect by an effective increase in the cross-sectional area. As described, the coils of the ac and dc windings are modeled with gyrators. The hysteresis in the core is modeled with corresponding resistors ($R_1 - R_{13}$) [10].

The capacitances that represent leakage permeances can be divided into two primary categories: leg leakage permeances ($C_{14} - C_{18}$) and yoke leakage permeances ($C_{19} - C_{26}$).[13]

The dc electric control circuit is connected to the dc source via five gyrators that represent each dc windings.

4. Bias Dc Sources

As stated before, four different methods to control the current in the dc winding are assumed. They are applied to explore how different types of the dc electric control circuits impact the operation of the three-phase CVSR in terms of equivalent ac inductance, induced voltage across dc winding, ac terminal

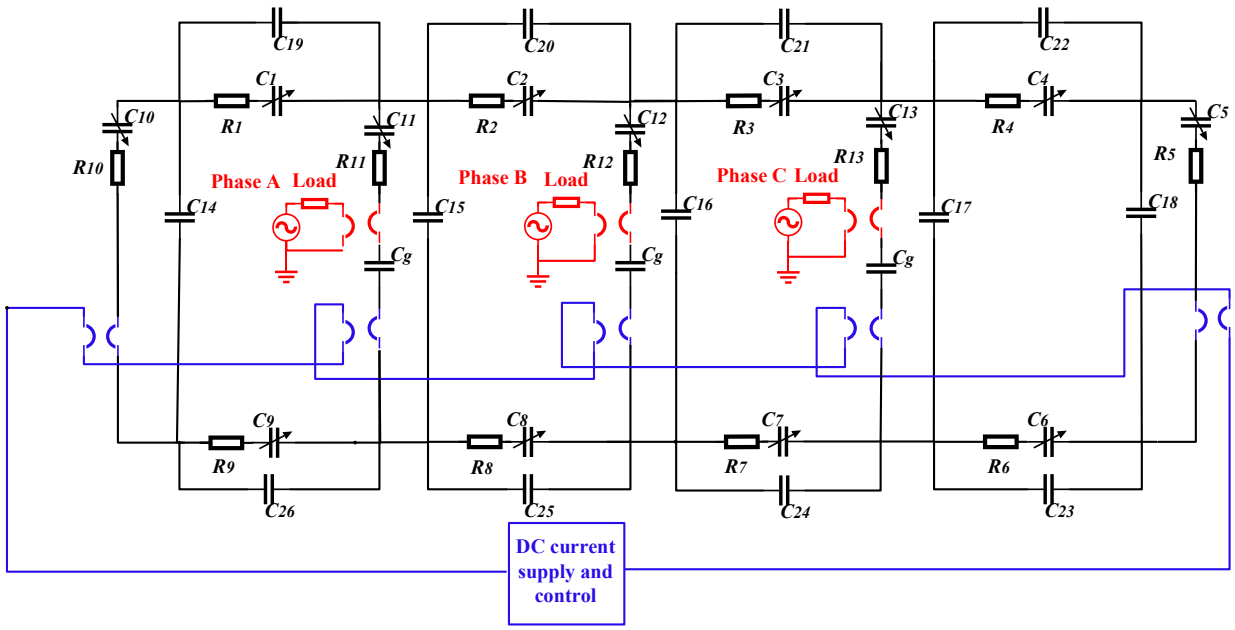


Figure 4: Gyrator-capacitor model of three-phase CVSR

voltage, and power interchange with the dc control circuit during different operating conditions.

All of the dc circuit controllers (DCCs) are simulated to supply the dc winding with five coils at a set of desired current values. The ideal current source has an infinite internal impedance, while the ideal voltage source has no impedance. H-bridge and buck converter are more realistic sources, with an internal impedance value between zero and infinity [1].

Fig. 5 illustrates an H-bridge converter composed of four IGBTs and a front-end rectifier. The PI controller coefficients are selected to ensure standard overshoot, settling time, and ripple in a steady state. A simple buck converter (step-down converter) is illustrated in Fig. 6. It operates in two modes: a) charging and b) discharging. During the charging mode, the IGBT turns on, and the dc winding current ramps up. During the discharging mode, the IGBT turns off, and the dc winding current ramps down [1].

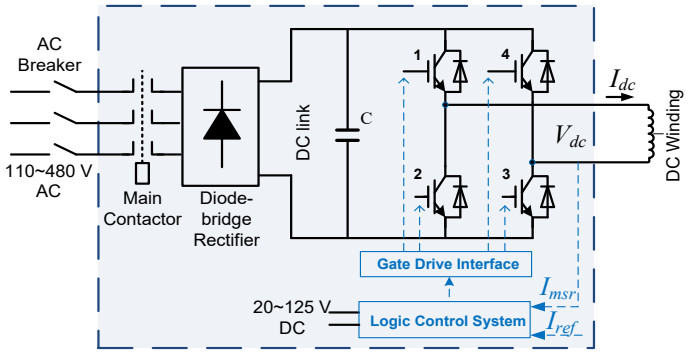


Figure 5: H-bridge converter

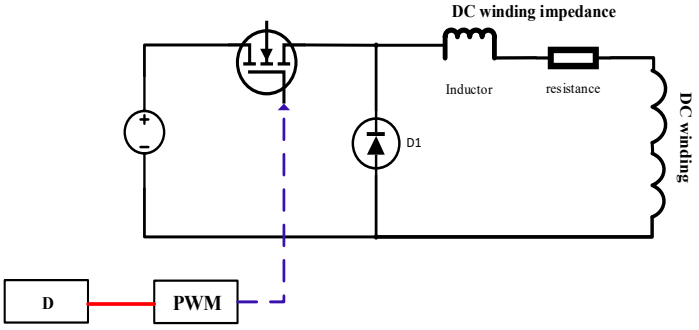


Figure. 6. Buck converter

5. Case Study

To explore the three characteristic operating states of the CVSR with unsaturated, partially saturated, and fully saturated core, simulations have been conducted for three different dc bias currents: 0 A, 15 A, and 45 A. The simulations results include waveforms for the terminal voltages and currents for each dc source type, at each characteristic bias value. Based on these waveforms, the effective impedances are computed.

Fig. 7 shows the equivalent circuit on the controlled ac power system side with source, load, and CVSR's inductance in series.

The ac equivalent inductance is obtained from the current through the ac winding and the terminal voltage (Fig. 7). Based on the G-C model definition, the terminal voltage across the L_{CVSR} (ac variable equivalent inductance of the CVSR) is equal to the gyrator induced voltage on the primary side. It is obtained from the basic Ohm's Law (2):

$$Z_{ac} = V_{ac}/I_{ac} \tag{2}$$

where: V_{ac} is the terminal voltage, and I_{ac} is the load current.

For simplicity, it has been supposed that the power system is balanced. Consequently, all of the simulation results shown here are for one phase.

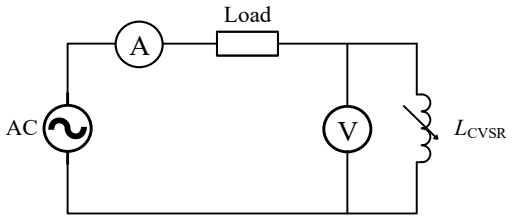


Figure 7: Ac electric circuit

5.1. CVSR without hysteresis

Fig. 8 illustrates the B-H characteristic for the inner legs of the CVSR at 0 A [14].

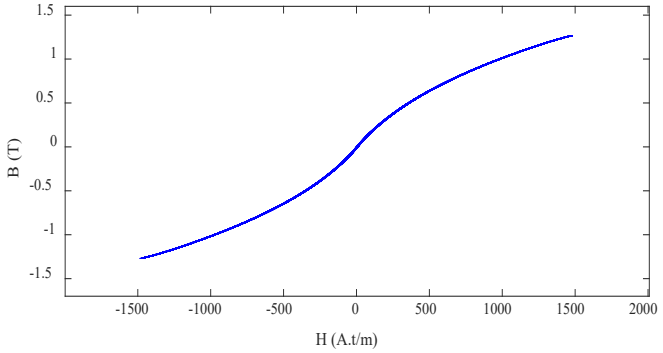


Figure 8: B-H characteristic of the inner legs without hysteresis

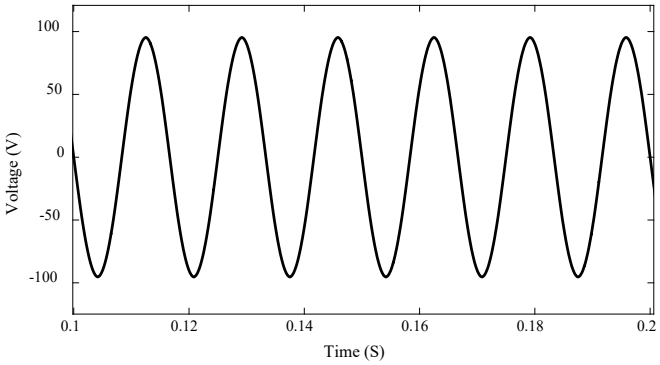


Figure 9: Terminal voltage (I_{dc} = 0 A)

5.1.1. 0A dc bias

The terminal voltage and current through the ac winding (load current) in one phase, from the G-C model at 0 A, are shown in Figs. 9 and 10, respectively. At no bias, the voltage and current waveforms are identical for all source types. In all operational circumstances and for every control source, the current passing through the AC winding is approximately equivalent to the nominal load current value of 20.9 A. This is because, in

comparison to the device's equivalent inductance, the load impedance holds more significance. Thus, the analysis will now shift its focus solely on the terminal voltage.

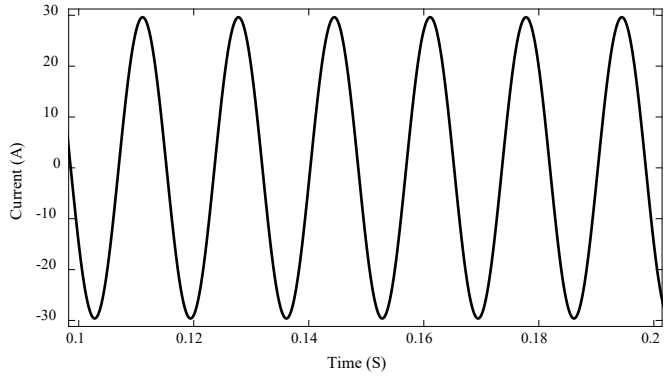


Figure 10: Load current (I_{dc} = 0 A)

The induced voltage across the dc winding in this case is shown in Fig. 11.

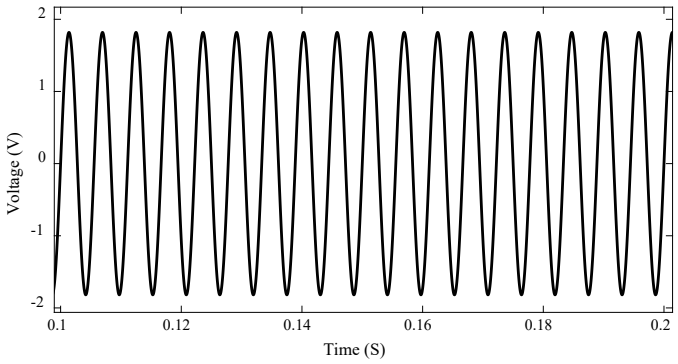
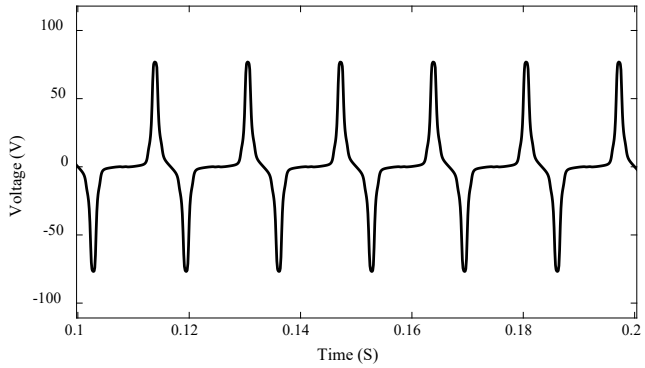


Figure 11: Voltage across dc winding (I_{dc} = 0 A)

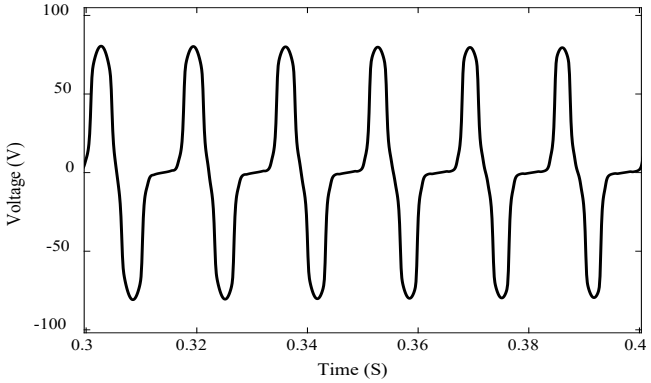
5.1.2. 15A dc bias

Figs. 12 (a)-(d) depict the terminal voltage on one phase of the CVSR for different dc controlled circuits, when the dc bias is raised to 15 A, the inner legs go into partial saturation, while the others are unsaturated. The flat regions in the voltage waveforms are caused by the saturation. The thick parts in the last two figures are the ripple effect of the PWM frequency (2.5 kHz). The controls move back and forth through the nonlinear B-H characteristic during the transition between the unsaturated and saturated region. This effect does not occur with an ideal current source. It can be seen that the induced voltage with the H-bridge converter is smoother with smaller fluctuations than that with the buck converter due to the closed- loop control. The figures show steady-state conditions, which depend on the settling time of the DCCs, hence the different time periods.

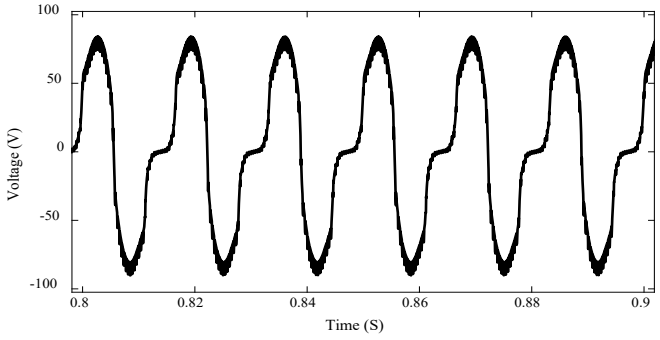
Figs. 13 (a)-(b) show the induced voltages across the dc winding for an ideal current source and an H-bridge converter. The voltage for a buck converter is roughly identical to the latter in terms of shape. The induced voltage is distorted at times when parts of the core enter saturation and it has triple the fundamental frequency of the system. Therefore, the ripple effect on the voltage waveforms is even more exaggerated with realistic DCCs. This effect aside, they tend to have the same shape and peak values.



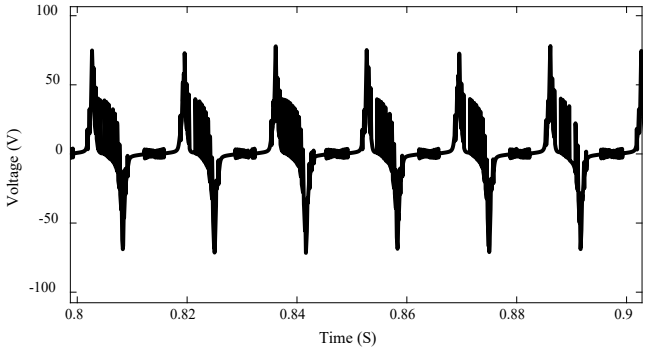
(a) Ideal current source



(b) Ideal Voltage source

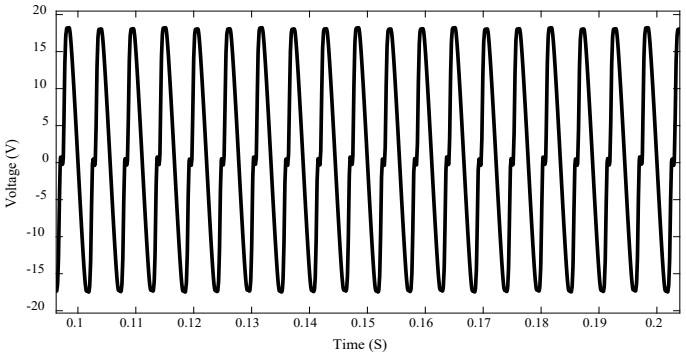


(c) H-bridge converter

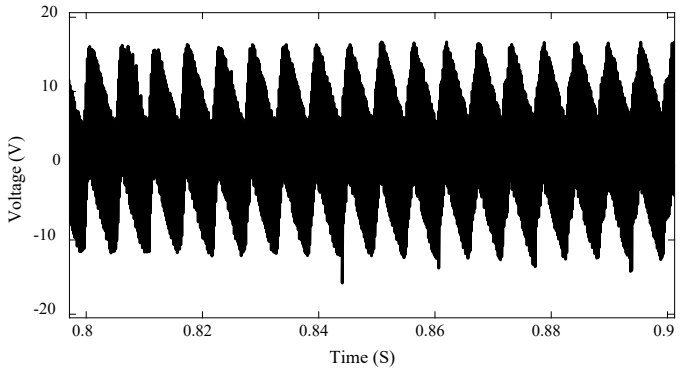


(d) Buck converter

Figure 12: Terminal voltages for different dc bias sources (I_{dc} = 15 A)



(a) Ideal current source



(b) H-bridge converter

Figure 13: Induced voltage across dc winding for different dc sources (I_{dc} = 15 A)

The power exchange with the dc winding is calculated using a "rolling window" approach. Starting at the beginning of the simulation, a window with *m* consecutive samples in time is chosen. The window then moves with each next sample. The power is calculated for each rolling window subset. The window size *m* depends on the fundamental period *T* and the sampling frequency of the data (step size). A longer window size produces smoother results.

The power transferred to the dc winding for an ideal current source is shown in Figure 14. After the initial transient, there is no real power exchange, apart from the small ripple. However, there is a noticeable apparent power transfer present.

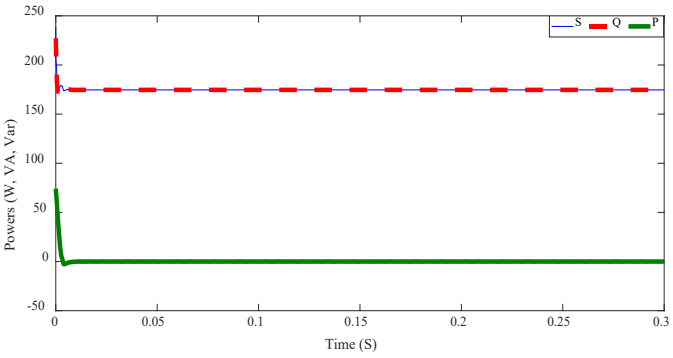


Figure 14: Power transfer to dc winding (I_{dc} = 15 A)

5.1.3. *45A dc bias*

At a high dc offset of 45 A, the CVSR core goes into complete saturation. Hence, the fluxes through the legs decrease. Also, the terminal voltage and the voltage across the dc winding are very low. Fig. 15 shows the terminal voltage on one phase for the ideal dc current source.

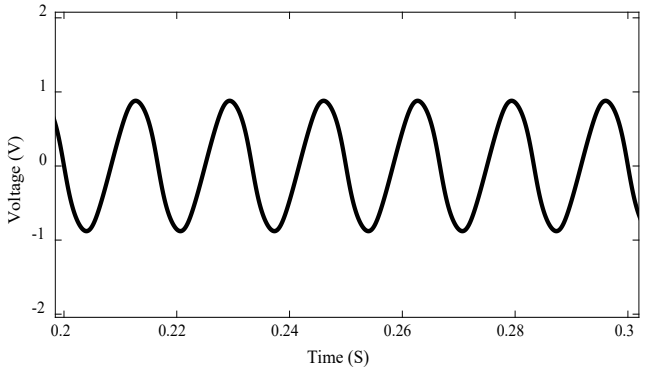


Figure 15: Terminal voltage (I_{dc} = 45 A)

Figs. 16(a)-(c) on the next page show the voltage across the dc winding for an ideal current source, an H-bridge converter, and a buck converter, respectively.

The power transferred to the dc winding for an H-bridge converter is shown in Fig. 17. After the transient, there is no real power and a smaller than before apparent power exchange.

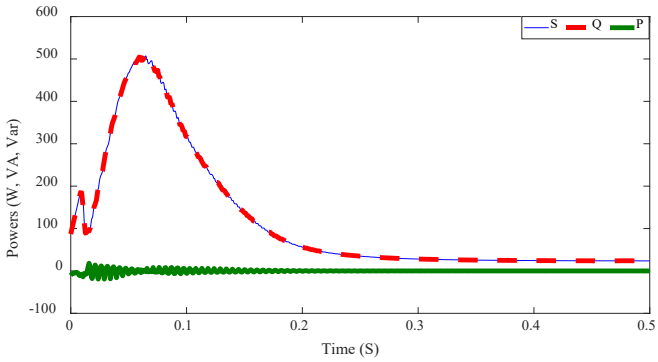
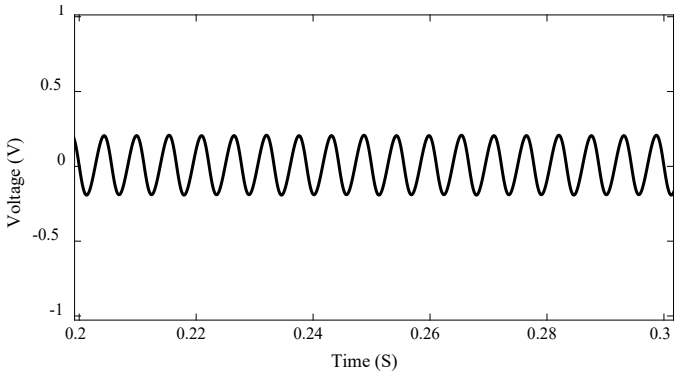


Figure 17: Power transfer to dc winding (I_{dc} = 45 A)

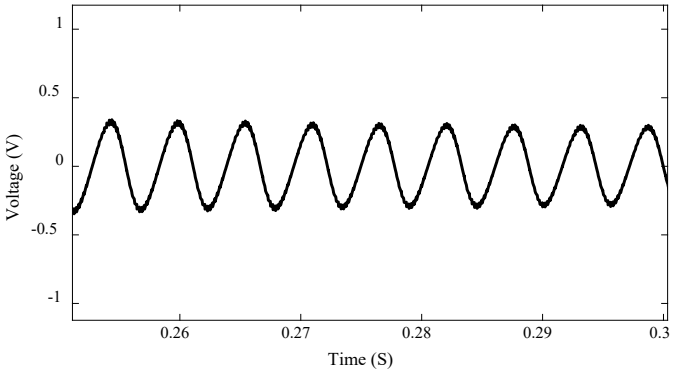
5.2. *CVSR with hysteresis*

The B-H characteristic for the inner legs at 0A is presented in Fig. 18. The hysteresis is small, but visible. The assumed material for the core is soft ferromagnetic and has small hysteretic characteristic. This is consistent with the practice how power electromagnetic devices are built to reduce the core losses. Still, it is of interest to see how this effect will influence the performance of the CVSR.

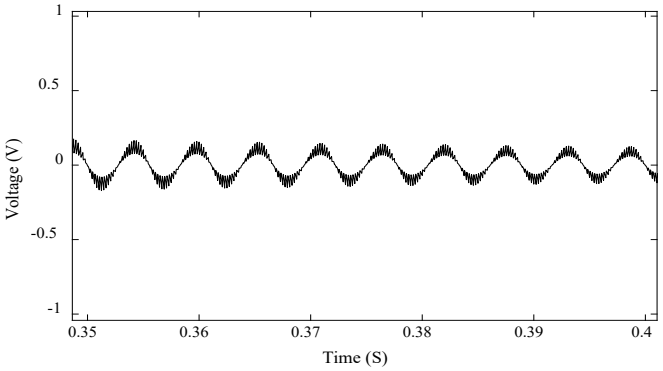
Again, like in the case without considering hysteresis, three different dc bias currents are considered: 0A, 15A, and 45A. The same analyses as before are performed and the results are compared with the previous case.



(a) Ideal current source



(b) H-bridge converter



(c) Buck converter

Figure 16: Voltage across dc winding for different dc sources (I_{dc} = 45 A)

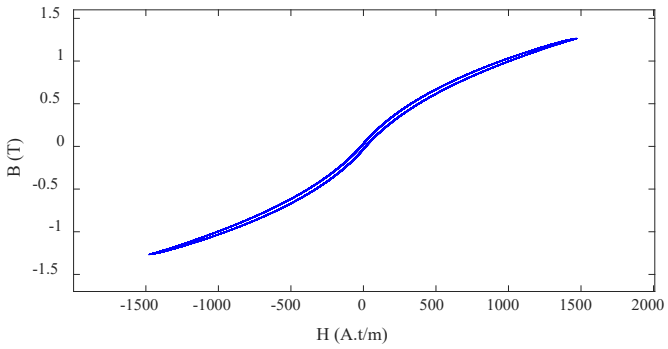


Figure 18: B-H characteristic of the inner legs with hysteresis

5.2.1. 0A dc bias

Fig. 19 shows the terminal voltage on one phase of the CVSR for an ideal current source. Because of the hysteresis, the voltage is not purely sinusoidal. The effect is not obvious and, for comparison, a pure sinusoid is plotted with a dashed red line. The same also for the voltage across the dc winding in Fig. 20.

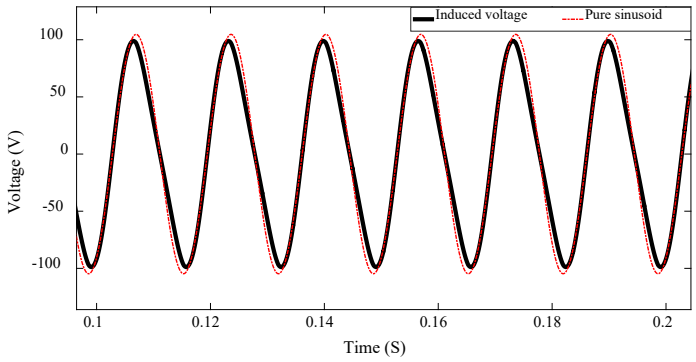


Figure 19: Terminal voltage (I_{dc} =0 A)

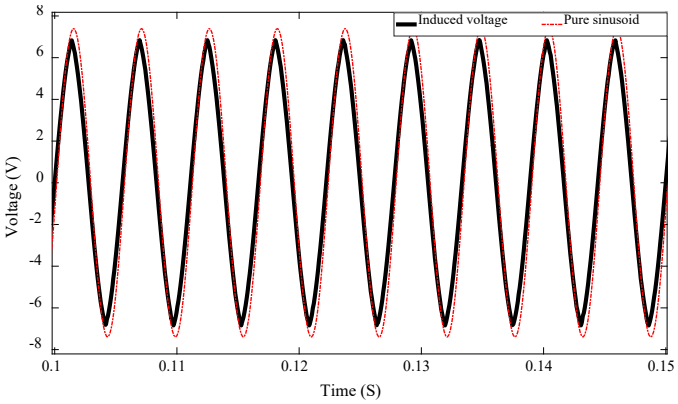


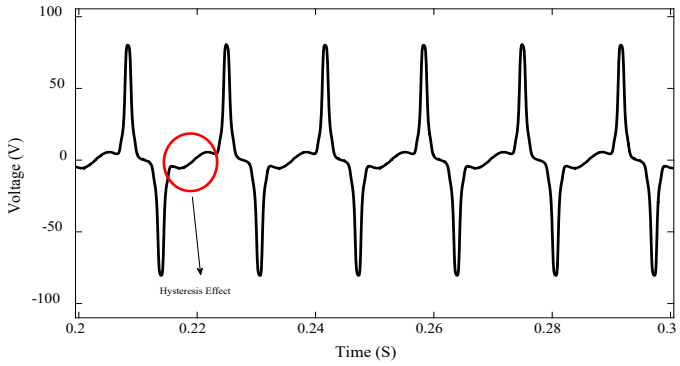
Figure 20: Voltage across dc winding (I_{dc} = 0A)

5.2.2. 15A dc bias

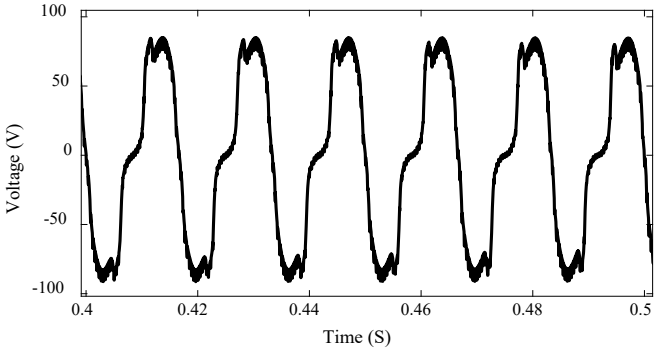
In Figs. 21 (a)-(b), the terminal voltage across one phase for an ideal current source and an H-bridge converter are shown, respectively, at dc bias equal to 15 A. At this bias current, the inner legs go into partial saturation, while the others are unsaturated. The hysteresis effect is encircled in red.

Figs. 22 (a)-(b) show the voltage across the dc winding for the same dc source types. Again, the effect from the hysteresis on the voltage waveform is encircled in red. It can also be seen that, although the hysteresis is quite small, the peaks of the induced voltages are significantly higher than those in the previous case for the same scenario and vary. This shows the value of the improved modeling in analysis of a device like CVSR.

The power transferred to the dc circuit for an ideal current source, shown in Fig. 23, is also significantly higher than in the previous case due to the higher induced voltages from including the hysteresis effect.

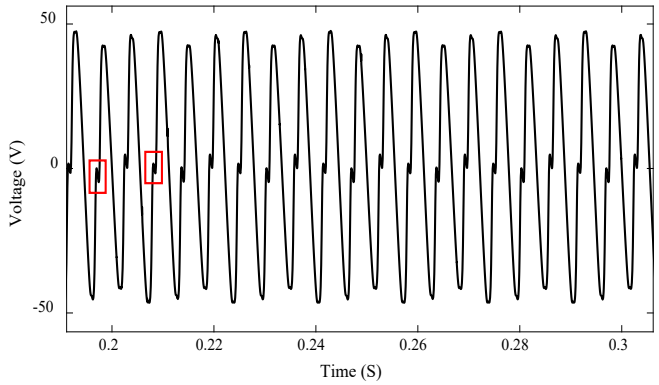


(a) Ideal current source

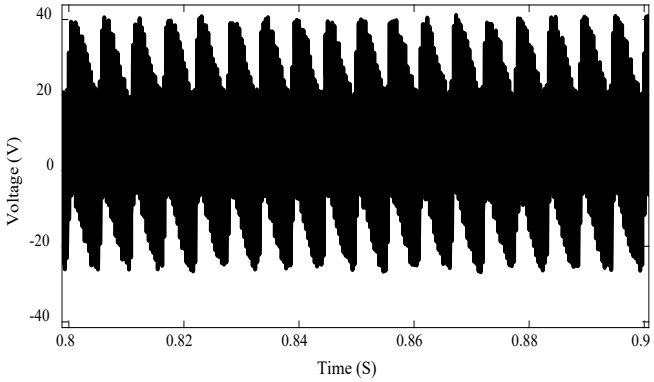


(b) H-bridge converter

Figure 21: Terminal voltage for different dc electric control circuits (I_{dc} = 15 A)



(a) Ideal current source



(b) H-bridge converter

Figure 22: Induced voltage across dc winding for different dc sources (I_{dc} = 15 A)

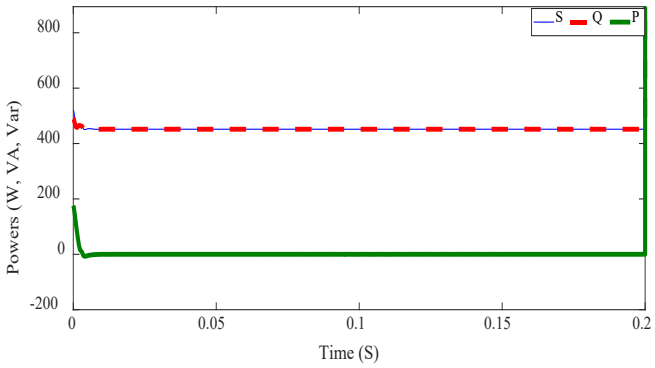
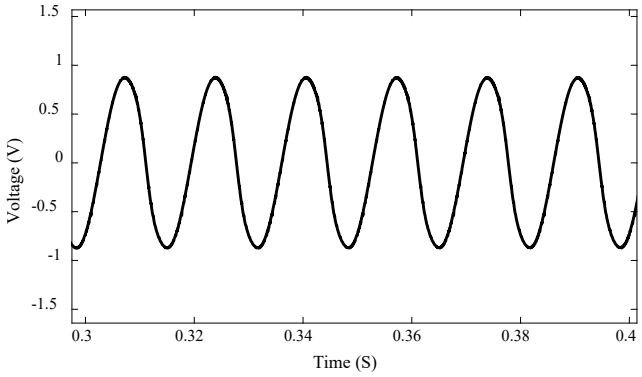


Figure 23: Power transferred to dc winding (I_{dc}= 15A)

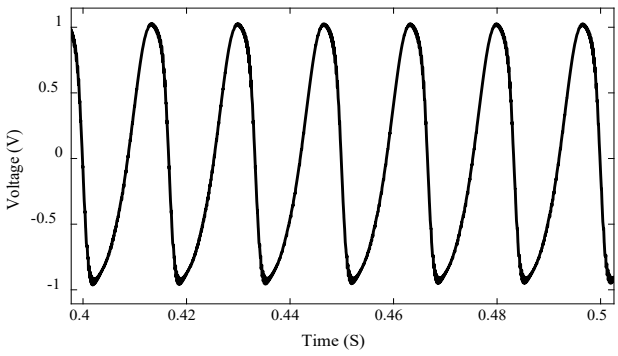
5.2.3. *45A dc bias*

At a high dc offset of 45 A, the CVSR core again goes into complete saturation. The fluxes through the legs decrease due to core fully saturation. Hence, the terminal voltage and the voltage across the dc winding are very low. Figs. 24 (a)-(b) show the terminal voltage on one phase of the CVSR for an ideal current source and an H-bridge converter, respectively.

Figs. 25 (a)-(b) show the induced voltage across the dc winding for the same dc source types. In Fig. 25, the distorted peaks encircled in red, due to the hysteresis effect, are still visible.

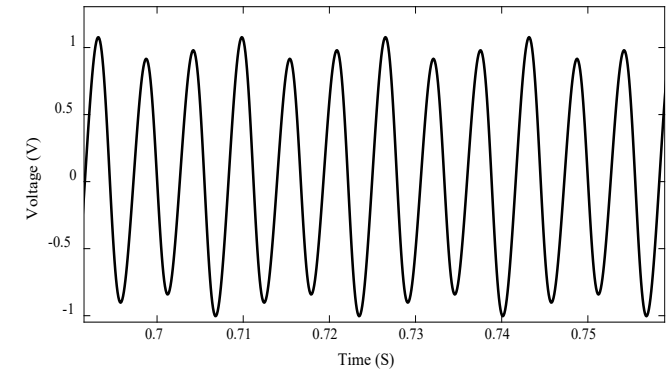


(a) Ideal current source

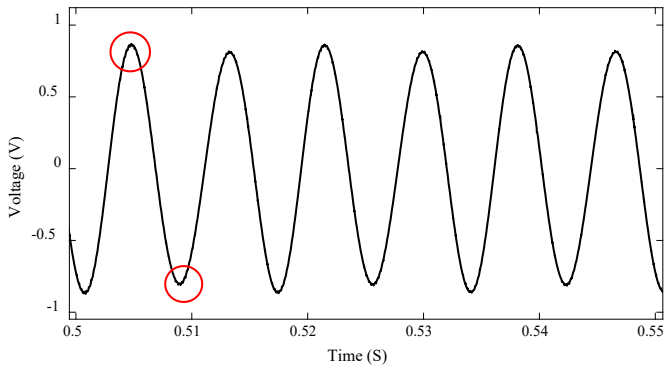


(b) H-bridge converter

Figure 24: Terminal voltage (I_{dc} = 45 A)



(a) Ideal current source



(b) H-bridge converter

Figure 25: Induced voltage across dc winding for different dc sources (I_{dc} = 45 A)

The power transferred to the dc winding for a Buck converter in this case is shown in Fig. 26.

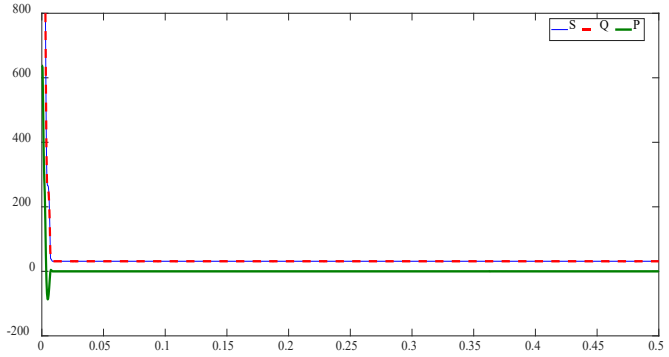


Figure 26: Power transferred to dc winding (I_{dc} = 45 A)

Tables I and II summarize the terminal voltage values and the resulting effective CVSR impedance values derived from the results of the analysis. The differences in the impedance reveal that both the dc source type and the hysteresis effect significantly impact the effective AC reactance of the CVSR.

6. Conclusion

The paper presents an improved realistic model of a three-phase CVSR, based on the gyrator-capacitor modeling approach. To improve the accuracy for an electromagnetic device that operates in the whole range of its B-H characteristic like the CVSR, hysteresis and core saturation nonlinearities are considered. Capacitors model permeances (magnetic conductances), and nonlinear capacitors can model core saturation. Additionally, a resistor connected in series with the core capacitor represent core hysteresis. Simulations of the improved G-C model of three-phase CVSR under different conditions and at different values of the bias dc current have been performed. Different dc control source types have also been considered. Results from the comprehensive analysis show significant impacts on the performance of the CVSR from a more realistic model.

Table 1: AC impedance and terminal voltage for different DC sources and currents – no hysteresis

Bias current / Bias source	0A (No saturation)		15A (Partial saturation)		45A (Full saturation)	
	Impedance	Voltage	Impedance	Voltage	Impedance	Voltage
Ideal current	3.185 Ω	66.64 V	1.31Ω	27.3 V	0.060 Ω	0.88 V
Ideal voltage	3.196 Ω	66.81 V	1.82 Ω	38.1 V	0.119 Ω	1.85 V
Buck converter	3.192 Ω	66.73 V	1.56 Ω	32.4 V	0.080 Ω	1.21 V
H-bridge	3.192 Ω	66.74 V	1.69 Ω	35.4 V	0.077 Ω	1.15 V

Table 2: AC Impedance and Terminal Voltage for Different DC Sources and Currents – With Hysteresis

Bias current / Bias source	0A (No saturation)		15A(Partial saturation)		45A (Full saturation)	
	Impedance	Voltage	Impedance	Voltage	Impedance	Voltage
Ideal current	3.57 Ω	74.64 V	1.97 Ω	41.3 V	0.070 Ω	1.07 V
Ideal voltage	3.73 Ω	77.82 V	2.51 Ω	52.3 V	0.145 Ω	2.07 V
Buck converter	3.71 Ω	77.73 V	2.24 Ω	46.4 V	0.092 Ω	1.33 V
H-bridge	3.71 Ω	77.74 V	2.23 Ω	46.8 V	0.088 Ω	1.24 V

References

[1] M. Hayerikhiyavi and A. Dimitrovski, "Impact of Different Types of DC Bias Sources on the Effective Impedance of a CVSR," in 2022 IEEE Kansas Power and Energy Conference (KPEC), 1-6, 2022, doi: 10.1109/KPEC54747.2022.9814785.

[2] A. Dimitrovski, Z. Li, B. Ozpineci, "Applications of saturable-core reactors(SCR) in power systems," in 2014 IEEE PES T&D Conference and Exposition, 1–5, 2014, doi:10.1109/TDC.2014.6863404.

[3] A. Dimitrovski, Z. Li, B. Ozpineci, "Magnetic Amplifier-Based Power-FlowController," IEEE Transactions on Power Delivery, **30**(4), 1708–1714, 2015, doi:10.1109/TPWRD.2015.2400137.

[4] M. Hayerikhiyavi and A. Dimitrovski, "Gyrator-Capacitor Modeling of A Continuously Variable Series Reactor in Different Operating Modes," in 2021 IEEE Kansas Power and Energy Conference (KPEC), 1-5, 2021, doi: 10.1109/KPEC51835.2021.9446236.

[5] M. Hayerikhiyavi and A. Dimitrovski, "Comprehensive Analysis of Continuously Variable Series Reactor Using G-C Framework," in 2021 IEEE Power & Energy Society General Meeting (PESGM), 1-5, 2021, doi: 10.1109/PESGM46819.2021.9637971.

[6] M. Young, A. Dimitrovski, Z. Li and Y. Liu, "Gyrator-Capacitor Approach to Modeling a Continuously Variable Series Reactor," in IEEE Transactions on Power Delivery, **31**(3), 1223-1232, 2016, doi: 10.1109/TPWRD.2015.2510642.

[7] Valadkhan S, Morris K, Khajepour A. "Review and Comparison of Hysteresis Models for Magnetostrictive Materials". Journal of Intelligent Material Systems and Structures. 2009;**20**(2):131-142.

[8] M. F. Jaafar and M. A. Jabri, "Study and modeling of ferromagnetic hysteresis," in 2013 International Conference on Electrical Engineering and Software Applications, 1-6, 2013, doi: 10.1109/ICEESA.2013.6578426.

[9] M. Hayerikhiyavi and A. Dimitrovski, "Improved Gyrator-Capacitor Modeling of Magnetic Circuits with Inclusion of Magnetic Hysteresis," in 2022 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), 1-5, 2022, doi: 10.1109/TD43745.2022.9816976.

[10] S. D. Sudhoff, B. T. Kuhn, K. A. Corzine, B. T. Branecky, "Magnetic Equivalent Circuit Modeling of Induction Motors," IEEE Transactions on Energy Conversion, **22**(2), 259–270, 2007, doi:10.1109/TEC.2006.875471.

[11] G. M. Shane, S. D. Sudhoff, "Refinements in Anhysteretic Characterization andPermeability Modeling," IEEE Transactions on Magnetics, **46**(11), 3834–3843,2010, doi:10.1109/TMAG.2010.2064781.

[12] D. C. Hamill, "Gyrator-capacitor modeling: a better way of understanding magnetic components," Proceedings of 1994 IEEE Applied Power Electronics Conference and Exposition - ASPEC'94, Orlando, FL, USA, 1994, pp. 326-332 vol.1, doi: 10.1109/APEC.1994.316381.

[13] S. Pokharel and A. Dimitrovski, "Modeling of An Enhanced Three-phase Continuously Variable Reactor," in 2020 IEEE Power & Energy Society General Meeting (PESGM), 1-5, 2020, doi: 10.1109/PESGM41954.2020.9282074.

[14] M. Hayerikhiyavi and A. Dimitrovski, "Voltage Balancing Using Continuously Variable Series Reactor," in 2023 IEEE Texas Power and Energy Conference (TPEC), 1-5, 2023, doi: 10.1109/TPEC56611.2023.10078529.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 3, 11-18 (2023)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Markov Regime Switching Analysis for COVID-19 Outbreak Situations and their Dynamic Linkages of German Market

Kangrong Tan[*,1], Shozo Tokinaga[2]

[1]*Faculty of Economics, Kurume University, Fukuoka, 839-8502, Japan*

[2]*Department of Economics, Kyushu University, Fukuoka, 819-0395, Japan*

ARTICLE INFO

Article history:
Received: 22 February, 2023
Accepted: 25 April, 2023
Online: 15 May, 2023

Keywords:
COVID-19 outbreak
German stock index(DAX)
Markov Regime Switching Analysis (MRSA)
GARCH(Generalised Autoregressive heteroscedasticity)
Stock returns
Growth rates of the disease
Volatility

ABSTRACT

This paper deals with the analysis of the dynamic linkage, co-movement between COVID-19 outbreak situations and German stock market. Firstly, Markov Regime Switching Analysis(MRSA) is proposed and employed to investigate the situations in the pandemic, as to catch the dynamics of how the daily number of the newly-infected changes, and also to assess the impact of the pandemic situations on German Market. Secondly, we compute the log growth rates of the weekly new cases and the log-returns of weekly DAX index, then fit the GARCH models to both of them to acquire their volatilities. We then employ the MRSA model once more to expose the dynamic linkages and co-movement between these two volatilities series. Through our empirical analyses, we find that GARCH models can capture the dynamics of stock returns and the growth rates. On the other hand, the MRSA models work well to identify the dynamics between different regimes with different states in dealing with the volatilities obtained from the estimated GARCH models. Our proposed econometric methods are highly practical, it indicates the possibility of replicating the results obtained in this study to assess the impact of other epidemics and negative factors on economic activities. Knowing what may happen during a pandemic, more effective measures and actions can be taken to protect people while dealing with another pandemic in the future.

1 Introduction

This paper is an extension of work originally presented in CICS2021 [1]. After the presentation at the conference CICS2021, we have done more researches on this issue and more interesting empirical results are obtained and updated in this extension version.

In the past decades, many approaches have been developed to tackle a time-varying time series. The main idea is how to separate the whole data set(interval) into several subsets(subintervals) with different statistical characteristics.

One of the methodologies is to detect change point or structural change in data [2]–[9]. One is to track the time-varying series using a Particle Filter [10], [11]. But, in this study, we propose to apply the Markov Regime Switching Analysis (MRSA) models to the time-varying data. The reason is that a time-varying series canbe efficiently and appropriately split into several subintervals with different statistical characteristics(or different state-specific regimes) by using the MRSA models. Namely, the dynamics of these subintervals can be well captured by MRSA [12]–[21].

And in many cases the regime changes are corresponding to the change points. Regimes indicate the situations of the infected cases, or the specific-state subintervals(market states).

Another methodology we employ in this study is GARCH(Generalised Autoregressive heteroscedasticity) model [22], [23]. Since GARCH model can help us get different and consecutive subintervals based on the conditional heteroscedasticity and has been utilized ranging from finance to psychology [24]–[27]. We fit the stock returns of DAX Index and the growth rates of COVID-19 to GARCH models to get their volatilities. And then apply MRSA to the volatilities, to investigate and assess the impact of the pandemic on the German stock market [1], [12]–[21].

Focusing upon the changes of DAX Index is that Germany is the industry leader country in the Europe Union(EU), it is important to know what impact had on the German stock market during the COVID-19 pandemic. Knowing what may happen during a pandemic, we can predict and do more better while dealing with another pandemic in the future.

The remainder of this paper is organized as follows. Section 2 shows an overview of our proposed Markov Regime Switching

[*]Corresponding Author: Kangrong Tan, Faculty of Economics, Kurume University, Japan. Tel: 0942-43-4411 & Email: camox.wein.london@gmail.com

Analysis(MRSA). Section 3 explains the data sets we use in this study. Section 4 summarizes the outbreak situations in Germany. Section 5 presents our empirical results in situation analysis of the disease outbreak, and its impact on the market, by using the MRSA models. Section 6 displays the MRSA results by using the volatilities obtained by the GARCH model fittings. Section 7 gives concluding remarks.

2 An Overview of Markov Regime Switching Analysis

We here just give an overview of Markov Regime Switching Analysis before we present the results of our empirical analyses below [12]–[21].

Usually MRSA is utilized to depict the dynamics of a time series by using regime states. Regime states describe the different segments of a time series in different states. The common example of regime states is of two different states, denoted as s_t, either $s_t = 0$ or $s_t = 1$, presenting two different regimes.

A common MRSA model is to do regression assuming different regime states existing in the data. The advantage of this method is that one can expose the varieties around the means. Such as,

$$y_t = \mu_{s_t} + \epsilon_t \tag{1}$$

where $s_t = 0$ or 1 means that data y_t have two different mean levels in different segments, where ϵ_t follows a Gaussian process with mean zero and variance σ^2. State zero (s_t=0) corresponds to Regime 1, and verse vice.

Another usage of MRSA is to build an Markov regime switching autoregressive Model. A generalised presentation of the model can be described as follows, assuming the AR model with order p, namely,

$$y_t = c_{s_t} + \alpha_{s_t} x_t + \beta_{1 s_t} y_{t-1} + \cdots + \beta_{p s_t} y_{t-p} + \epsilon_{s_t} \tag{2}$$

where x_t is called as exogenous variable. The transition probabilities are usually defined as a first order Markov chain. Thus, a transition probability matrix P with n states can be represented as follows.

$$P = (p_{ji}) = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix} \tag{3}$$

where $i, j = 1, 2, ...n$, with $\sum_{j=1}^{n} p_{ji} = 1$, and p_{ji} is the probability of transitioning from regime i to regime j.

While, a simplified case of (2) is that a Markov regime switching autoregressive model without an exogenous variable.

The following model shows the time-varying data y_t follows two different AR(1) equations with different parameters, for example.

$$y_t = \begin{cases} \beta_0 + \beta y_{t-1} + \epsilon_t, & (s_t = 0) \\ (\beta_0 + \beta_1) + \beta y_{t-1} + \epsilon_t, & (s_t = 1) \end{cases} \tag{4}$$

What (1)-(4) mean is that the time-varying observations fluctuate when state or regime shifts from one to another.

The estimation of parameters of regime-switching models is implemented by maximizing the likelihood function.

3 Data Description

In this study, we use the following two data sets.

1) DAX(daily), which is known as the GER40 as well, comprised of 40 German companies traded on the Frankfurt Exchange. We obtained the data from Yahoo(https://de.finance.yahoo.com). The period of the data set is ranged from Jan, 2020 to Jul, 2021.

2) The newly infected cases of COVID-19(daily). We obtained the data from WHO's Website, in the same period.

These two data sets are open to the public on above-mentioned websites, and can be downloaded freely, at any time.

4 Situations of COVID-19 Outbreak

Germany is the industry leader in the Europe Union(EU), and it is important to investigate how the German stock market was impacted from the pandemic. We thus focus upon the situations in Germany.

In Germany, the ups and downs of the newly confirmed number(daily) are shown in Figure 1, duration from January 22, 2020 to July 30, 2021.

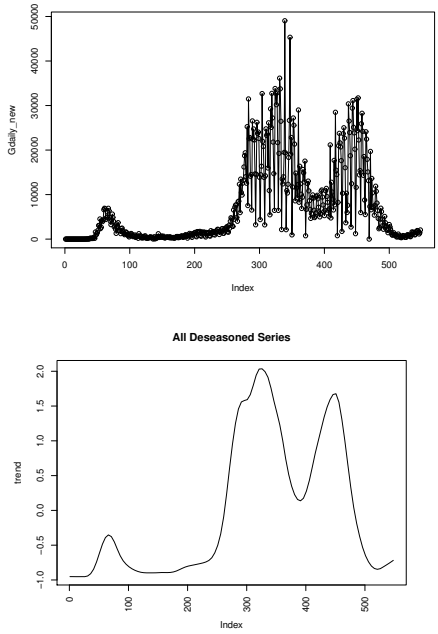


Figure 1: Daily newly confirmed number(left) and its trend(right)

It shows the newly confirmed number(daily) fluctuated with the lapse of time. The left small peak appeared when fewer cases were confirmed. However, gradually the newly confirmed number rose from zero level up to a big peak, and then decayed to its half level around. Sooner, the third peak appeared when the newly confirmed number surged once more. One may be interested in identifying change points, we had successfully identified the change points in this time series based upon a Bayesian approach[8], [9].

With the spread of infection of COVID-19 in Europe, lockdown policy had been employed in many countries, since the newly confirmed cases were growing up and up. The government enforced lockdown in Nov, 2020, and it had not been lifted until Jun, 2021, about 8 months long.

Besides, mask-wearing, social distancing, staying home and other public health measures had been also implemented, these measures had shown their effects on reducing newly infected cases.

5 Empirical Research

We then carry out some empirical analyses by using the above-mentioned two data sets.

5.1 Markov Switching Regression Analysis

Firstly, the Markov regime switching regression model is applied to the data. Table 1 shows the estimated results.

Table 1: Estimated results

Intercept	Regime 1	Regime 2
Estimate	12889	844.336
Std. Error	557	43
t value	23	19
$Pr(>\|t\|)$	2.2e-16***	2.2e-16***
AIC	10178.62	
Likelihood	-5087.31	

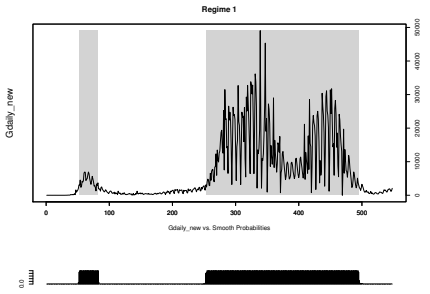


Figure 2: Regime 1 is represented in gray

Figure 2 shows the estimated areas of Regime 1 in gray. Seen from the figure, the peaks of daily infected number are identified as Regime 1 in gray, and verse vice.

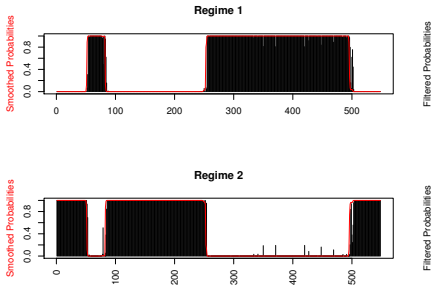


Figure 3: Plots of estimated probabilities of Regime 1 and 2

Figure 3 displays the estimated transition probabilities of Regime 1 and 2.

The corresponding transition probabilities are estimated and shown in Table 2. While, the corresponding confidence intervals(Level= 0.95) for the Intercepts estimated above are shown in Table 3.

Table 2: Estimated transition probabilities

	Regime 1	Regime 2
Regime 1	0.9926	0.0074
Regime 2	0.0074	0.9926

Table 3: Confidence intervals for the estimated intercepts

Intercept	Regime 1	Regime 2
Estimation	12889.56	844.34
Lower	11796.06	760.02
Upper	13983.05	928.65

Thus, the ability and the practicality of MRSA to precisely identify the different subintervals with their-own specific states have been confirmed distinctly.

5.2 Daily Growth Rate of the Disease

In this section, we discuss how the daily growth rates of the disease evolved. The daily growth rate is defined as follows.

$$dailygrowthrate_t = log(dailynewcases)_t - log(dailynewcases)_{t-1} \quad (5)$$

The plot of the calculated daily growth rates is shown in Figure 4.

It is shown that the daily infected cases were increasing(the first surge of the growth rates) in the early stages, and those infected people mostly got proper treatment in hospitals and discharged from the hospitals later.

We also can confirm the second surge of the growth rates around the Xmas holidays. And the growth rates got lower around the 500th day. We then apply the Markov regime switching regression model to the daily growth rates.

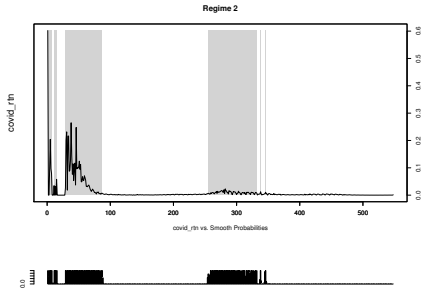


Figure 4: Regime 2 in gray indicates the areas of increasing growth rates

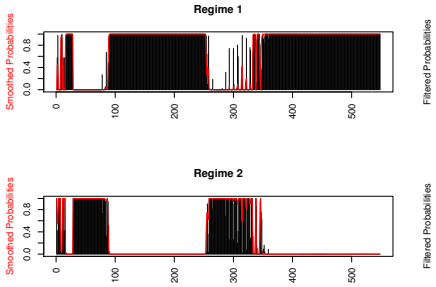


Figure 5: The Smoothed transition probabilities

Figure 4 and 5 display the two switching regimes and their corresponding smoothed transition probabilities, respectively.

The estimated results are summarized in the following Table 4 and 5. Table 4 shows the estimated intercept and its standard error, t value, and Pr(> |t|), and Table 5 gives the estimated transition probabilities, respectively.

Table 4: Estimated intercept and the corresponding statistics

Intercept	Regime 1	Regime 2
Estimate	0.0016	0.0395
Std. Error	0.0001	0.0058
t value	16	6.8103
Pr(> $\|t\|$)	2.2e-16 ***	9.7e-12 ***
AIC	-4408.92	
Likelihood	2206.46	

Table 5: Estimated transition probabilities

Transition probabilities	Regime 1	Regime 2
Regime 1	0.9807	0.0444
Regime 2	0.0193	0.9556

5.3 *Applications of Markov Switching Autoregression Model*

Here we apply the above-mentioned Markov switching autoregressive model of order p to our two data sets. In the following applications, we transform these two daily data sets into weekly ones, namely, DAX_{weekly}, and $Newcases_{weekly}$.

Application 1: Autoregressive model of order 1

That is to say, we fit the variables as follows.

$$DAX_{weekly} \sim Newcases_{weekly} + DAX_{weekly}(t-1). \quad (6)$$

As a result of this model setting, the following facts are revealed and summarized in Table 6 and 7.

Looking at these tables, it is clear that most of the estimated parameters are statistically significant, except the intercept in Regime 1, and the coefficient of $Newcases_{weekly}$ in Regime 2.

The multiple R-squared is estimated as 0.8053, and 0.986 in Regime 1 and 2, respectively.

Meanwhile, the state transition probabilities are summarized in Table 8.

The confidence intervals(CI)(Level= 0.95) for the estimated parameters, namely, intercept, the coefficients of $Newcases_{weekly}$ and $DAX_{weekly}(t-1)$ are displayed in Table 9 and 10.

Table 6: Estimated results of Regime 1

Regime 1	Intercept	$Newcases_{weekly}$	$DAX_{weekly}(t-1)$
Estimate	1325.08	0.3516	0.8482
Std. Error	1376.41	0.1887	0.1116
t value	0.963	1.8633	7.6004
$Pr(>\|t\|)$	0.3357	0.06242 .	2.95e-14***
AIC	1101.393		
Likelihood	-544.6966		

Table 7: Estimated results of Regime 2

Regime 2	Intercept	$Newcases_{weekly}$	$DAX_{weekly}(t-1)$
Estimate	521.332	-0.007	0.9700
Std. Error	221.256	0.016	0.0167
t value	2.356	-0.409	58.0838
$Pr(>\|t\|)$	0.01846*	0.683	2e-16***

Table 8: Estimated state transition probabilities

	Regime 1	Regime 2
Regime 1	0.6348	0.0852
Regime 2	0.3652	0.9148

Table 9: Confidence intervals for the parameters in Regime 1

	Intercept	$Newcases_{weekly}$	$DAX_{weekly}(t-1)$
Estimation	1325.0807	0.3516	0.8482
Lower	-1367.0203	-0.0182	0.6298
Upper	4017.1817	0.7214	1.0665

Table 10: Confidence intervals for the parameters in Regime 2

	Intercept	$Newcases_{weekly}$	$DAX_{weekly}(t-1)$
Estimation	521.3317	-0.0067	0.97
Lower	87.5582	-0.0386	0.937
Upper	955.1051	0.0252	1.0027

These two shifting regimes and their corresponding transition probabilities are displayed in Figure 6 and 7, respectively. Seen from the Figure 6, Regime1(in gray) captures the steep descent of DAX.

As can be seen from the results above, it is clear that, the Markov regime switching autoregressive model of order one captures the sudden ups and downs of the observations. It means that, in the early stages, volatile movements in the market are observed due to the growth of daily new cases, but, after the fortieth day, the change of the stock index slowed down, as if the market had acclimatized to the disease.

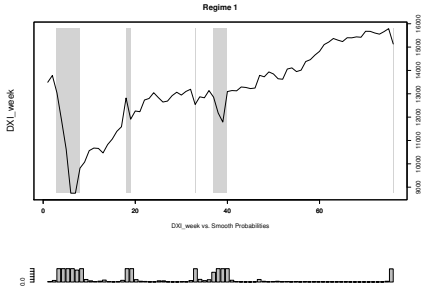


Figure 6: Regime 1(in gray) captures the steep decent of DAX

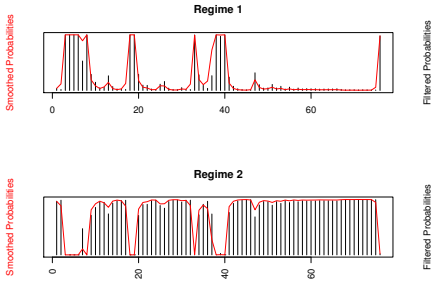


Figure 7: Smoothed transition probabilities

Application 2: Autoregressive model of order 2

We here fit the variables of our data sets by using Markov regime switching autoregressive model of order 2. Namely,

$DAX_{weekly} \sim Newcases_{weekly} + DAX_{weekly}(t-1) + DAX_{weekly}(t-2)$.

The numerical results are summarized in Table 11 and Table 12. Seen from these tables, it is clear that most of the estimated statistics are better than the results of AR(1) model obtained above, even the multiple R-Squared is better in each Regime.

Table 11: Corresponding statistics of Regime 1

	Intercept	$Newcases_{weekly}$	$DAX_{weekly}^{(t-1)}$	$DAX_{weekly}^{(t-2)}$
Estimate	2642.633	0.367	1.337	-0.584
Std. Error	1365.812	0.170	0.227	0.262
t value	1.935	2.152	5.892	-2.231
$Pr(>\|t\|)$	0.0530	0.0314*	3.8e-9***	0.0257*
R-squared	0.857			
AIC	1083.065			
Likelihood	-533.5325			

Table 12: Corresponding statistics of Regime 2

	Intercept	$Newcases_{weekly}$	$DAX_{weekly}^{(t-1)}$	$DAX_{weekly}^{(t-2)}$
Estimate	586.227	0.0006	0.867	0.098
Std. Error	217.604	0.016	0.078	0.075
t value	2.694	0.038	11.134	1.302
$Pr(>\|t\|)$	0.0071**	0.9699	2e-16***	0.193
R-squared	0.9871			

The estimated state transition probabilities are listed in Table 13.

Table 13: Estimated state transition probabilities

	Regime 1	Regime 2
Regime 1	0.7071	0.0904
Regime 2	0.2929	0.9096

Furthermore, looking at Table 11 , we see that the weekly new cases do have impact on the market, meanwhile DAX_{weekly}(t-1), DAX_{weekly}(t-2) show statistically significant in this AR(2) model setting. It also can be confirmed that Figure 8 captures the stock market plunges precisely. In other words, it reveals that Markov Switching Autoregressive Analysis can identify change points precisely as well as other tools, the change point identification based upon a Bayesian approach, for example.

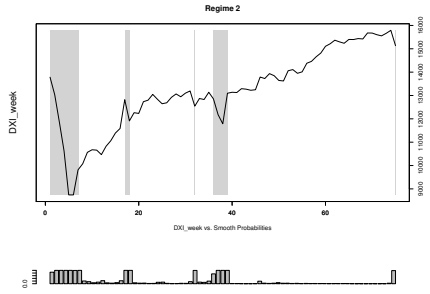


Figure 8: Regime 2(in gray) captures the market plunges

6 Fitting the Datasets Using GARCH Models

In this section, we first fit our two datasets using GARCH models, since GARCH model is considered to be a useful mean for capturing

the dynamics in data.

The dynamics or the momentum can be described by the volatilities series obtained from GARCH model. Second, we apply MRSA to the volatilities to see whether there exists some sort of co-movement between the weekly growth rates of the disease and the weekly returns of stock index.

Details of GARCH model are omitted here, one may get more information from other references, such as, [22], [23]. Hereafter, we just display our empirical results.

Our model fittings reveal that GARCH(1,1) and GARCH(1,0) are the best choices for the weekly returns of DAX and the weekly growth rates of the disease.

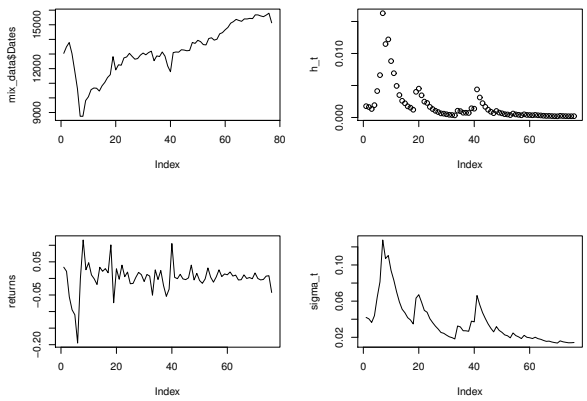


Figure 9: Plots of the results of GARCH(1,1) for returns

Table 14: Summarized results for GARCH(1,1)

	omega	alpha1	beta1	skew
Estimation	4.4675e-05	0.30499	0.70171	0.84704
Std err	3.785e-05	0.1497	0.1032	0.1021
t Value	1.180	2.037	6.798	8.299
Pr(> |t|)	0.2379	0.0416*	1.1e-11***	2e-16***
AIC		-4.0190		
BIC		-3.8964		
HQIC		-3.9700		
Likelihood		-156.72		

The plots of weekly stock returns and the weekly number of the newly infected are shown in Figure 9 (leftup) and Figure 10 (leftup) below. It seems that the DAX Index didn't react sensitively, in corresponding to the changes of the weekly new cases. However, we get completely different results if we fit the weekly stock returns and growth rates using GARCH models, and then apply the Markov regime switching autoregressive model to volatilities obtained from the GARCH model fittings.

6.1 Fitting Results of DAX Index

By setting different p, q values in GARCH(p,q), we find that GARCH(1,1) fits the weekly DAX returns most appropriately, where $returns_{weekly} = log(r_t) - logr_{t-1}$, where r_t is the price at week t. It indicates the weekly returns follow a skewed normal distribution based upon the numerical results.

The fitting results are shown in Table 14. Looking at the table, we see it is a good fit.

The original time series of weekly DAX Index, its returns, h_t, and σ_t values obtained from the GARCH(1,1) model, are shown in Figure 9, respectively.

It can be confirmed in both Table 14 and Figure 9, that the dynamics of the weekly stock returns is well-captured by GARCH(1,1) model.

6.2 Fitting Results of Weekly Confirmed Number

Similarly to the DAX index above, we calculate the corresponding weekly growth rates of the disease, namely, $growthrate_{weekly} = logn_t - logn_{t-1}$, where n_t is the number at week t. We then fit the weekly data set using GARCH(p,q) model. By setting different p, q values in the model, we find that GARCH(1,0) fits the weekly growth rates most appropriately.

The fitting results are shown in Table 15. Looking at the table, we see a skewed normal distribution is detected, because of the data's statistical properties. It indicates that the dynamics of the statistical properties of the data is well captured by a skewed normal distribution. And what is more important is that all the corresponding statistics are statistically significant, a good fit as well.

Table 15: Summarized results for GARCH(1,0)

	omega	alpha1	skewness
Estimate	0.2337	0.2936	1.0158
Std. Error	0.0465	0.1534	0.1322
t value	5.025	1.914	7.685
Pr(> |t|)	5.03e-07***	0.0557.	1.53e-14***

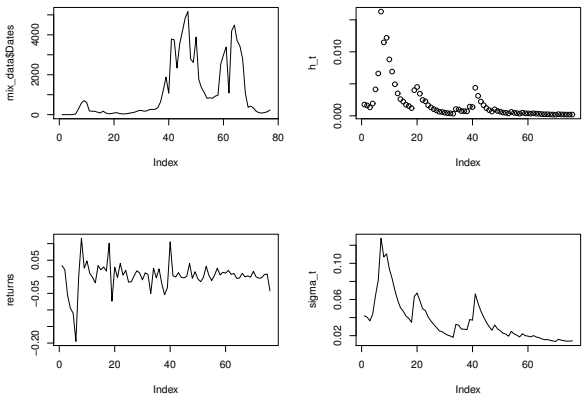


Figure 10: Plots of the results of GARCH(1,0) for growth rates

Similarly, the original time series of weekly number of new cases, weekly growth rates, h_t, and σ_t values obtained from the GARCH(1,0) model, are shown in Figure 10, respectively.

It can be confirmed in both Table 15 and Figure 10, that the dynamics of the weekly growth rates is well captured by GARCH(1,0) model.

6.3 *MRSA for Volatilities Obtained from GARCH Models*

In order to investigate the linkage and co-movement between the stock returns and the growth rates of the disease, we carry out MRSA based upon (2) using these two volatilities series($\sigma_t^{return}, \sigma_t^{growthrate}$), which are obtained from the GARCH(1,1) and the GARCH(1,0) models for the weekly returns and growth rates, respectively.

The estimated results are summarized in Table 16 and 17. And the linkage and co-movement are shown in Figure 11 and 12.

Seen from Table 16, the growth rate of the disease, which is regarded as an exogenous variable in the model setting, makes a severe impact on the stock change no matter which regime it stays.

Table 16: Estimated parameters for MRSA

Regime 1	Estimation	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	0.0007	0.0009	0.7778	0.4367
$\sigma_t^{growthrate}$	0.0031	0.0016	1.9375	0.0527 .
σ_{t-1}^{return}	0.8264	0.0059	140.068	2e-16***
R-squared	0.9972			
Regime 2	Estimation	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-0.0151	0.0070	-2.1571	0.031*
$\sigma_t^{growthrate}$	0.0424	0.0107	3.9626	7.4e-05***
σ_{t-1}^{return}	0.9644	0.1090	8.8477	2.2e-16***
R-squared	0.8972			

The estimated state transition probabilities are shown in Table 17.

Table 17: Transition probabilities

	Regime 1	Regime 2
Regime 1	0.8075	0.5387
Regime 2	0.1925	0.4613

The smoothed state transition probabilities are shown in Figure 11, Regime 1(up), and Regime 2(down).

And Regime 1 is displayed in Figure 12. It can be seen that the decayed σ_ts of DAX Index are captured in gray(Regime 1). These decayed parts (intervals) are corresponding to the change points as well.

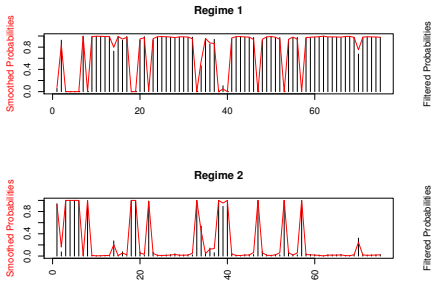


Figure 11: Smoothed state transition probabilities of regimes

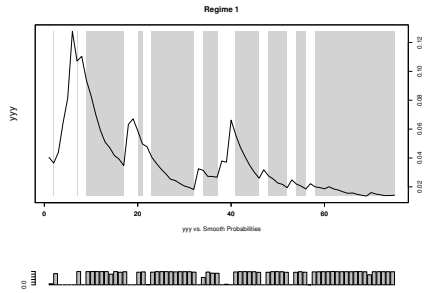


Figure 12: Regime 1 shows the decayed parts of DAX(YYY=σ_t)

Through our numerical analyses, it is confirmed that our proposed the Markov Regime Switching Analysis models are effective in investigating the dynamics, linkage and co-movement between the stock returns and the growth rates of the disease, namely, the statistical characteristics of the data are captured more precisely by using the volatilities obtained from GARCH models.

7 Concluding Remarks

In this paper, we have proposed and employed econometric apparatus and tools, the Markov Regime Switching Analysis method and other methods to investigate the situations of the outbreak of COVID-19 and its impact on German stock market.

It has been confirmed that the dynamics, linkage and co-movement between stock returns and growth rates of the disease are well captured by our proposed MRSA models through our numerical analyses.

Moreover, it has been confirmed that it is effective in precisely capturing the linkages between the stock returns and the growth rates of the disease by using volatilities obtained from GRACH models.

Our proposed methods are highly practical, it indicates the possibility of replicating the results obtained in this study to assess the impact of other epidemics and negative factors on economic activities, and provide researchers and policy-makers with a clue of how a pandemic evolved and what impact on the economies. It may lead to better measures and actions while dealing with another pandemic in the future.

Acknowledgment

The authors are grateful to the reviewers' valuable comments that improved the manuscript. This study was partly supported by the Japan Society for the Promotion of Science(JSPS) KAKENHI Grant Number(c)18K04626. The authors would like to thank the organization.

References

[1] K.R. Tan, S. Tokinaga, "Markov Regime Switching Analysis for the Pandemic and the Dynamics of German Market," Proceedings of International Conference on Computational Science and Computational Intelligence(CSCI), IEEE Xplore, 2021.

[2] D.W.K. Andrews, "Tests for parameter instability and structural change with unknown change points," Econometrica, **61**, 821-856, 1993.

[3] J. Chen, A.K. Gupta, "Change point analysis of a Gaussian model," Statistical Papers, **40**, 323-333, 1999.

[4] F. Desobry, M. Davy, C. Doncarli, "An online kernel change detection algorithm," IEEE Trans. Signal Process, **53**(8), 2961-2974, 2005.

[5] G. Koop, S.M. Potter, "Estimation and forecasting in models with multiple breaks," Review of Economic Studies, **74**(3), 763-789, 2007.

[6] R. Malladi, G.P. Kalamangalam, B. Aazhang, "Online Bayesian Change Point Detection Algorithms for Segmentation of Epileptic Activity," Asilomar Conference on Signals, Systems and Computers, 1833-1837, 2013.

[7] J. Knoblauch, T. Damoulas, "Spatio-temporal Bayesian On-line Changepoint Detection with Model Selection," Proceedings of the 35th International Conference on Machine Learning(ICML-18), 2718-2727, 2018.

[8] K.R. Tan, "Detecting structural changes in stochastic differential equation system based upon a Bayesian approach," Journal of Economic and Social Research, **58**(1-2), 51-67, 2018.

[9] K.R. Tan, "Identifying the pandemic change points of COVID-19 outbreak: case studies in Germany, Italy and Austria," Journal of Economic and Social Research, **61**(2-3), 19-33, 2021.

[10] S. Tokinaga, S. Matsuno, "Estimation of Transition of Credit Rating by Using Particle Filters Based on State Equations Approximated by the Genetic Programming," IEICE Proceeding Series, **45**, 397-400, 2011.

[11] F. Caron, A. Doucet, R. Gottardo, "On-line changepoint detection and parameter estimation with application to genomic data," Stat Comput, **22**, 579-595, 2012.

[12] J.D. Hamilton, "Rational-expectations econometric analysis of changes in regimes: an investigation of the term structure of interest rates," Journal of Economic Dynamics and Control, **12**, 385-423, 1998.

[13] J.D. Hamilton, "A new approach to the economic analysis of nonstationary time series and the business cycle," Econometrica, **57**, 357-384, 1989.

[14] J.D. Hamilton, Time series analysis, Princeton University Press, 1994.

[15] S. Goutte, B. Zou, "Foreign exchange rates under Markov regime switching model," Center for Research in Economic Analysis Discussion Paper, **16**, 1-29, 2011.

[16] A. Ramponi, "VaR-optimal risk management in regime-switching jump-diffusion models," Journal of Mathematical Finance, **3**(1), 103-109, 2013. DOI: 10.4236/jmf.2013.31009

[17] S. Choi, M. Marcozzi, "A regime switching model for the term structure of credit risk spreads," Journal of Mathematical Finance, **5**, 49-57, 2015. DOI: 10.4236/jmf.2015.51005

[18] H. Boubaker, N. Sghaier, "Markov-switching time-varying copula modeling of dependence structure between oil and GCC stock markets," Open Journal of Statistics, **6**, 565-589, 2016. DOI: 10.4236/ojs.2016.64048

[19] S. Gyamerah, P. Ngare, "Regime-switching model on hourly electricity spot price dynamics," Journal of Mathematical Finance, **8**, 102-110, 2018. DOI: 10.4236/jmf.2018.81008

[20] G. Stefano, O. Aydin, S. Abhijit, J. Mazin, "Pricing of time-varying illiquidity within the Eurozone: evidence using a Markov switching liquidity-adjusted capital asset pricing model," International Review of Financial Analysis, **64**, 145-158, 2019. DOI: 10.1016/j.irfa.2019.05.002

[21] Z. Qu, Z. Fan, "Likelihood ratio-based tests for Markov regime switching," Review of Economic Studies, **88**(2), 937-968, 2021. DOI: 10.1093/restud/rdaa035

[22] R. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation," Econometrica, **50**, 987-1007, 1982.

[23] T. Bollerslev, "Generalized autoregressive conditional heteroskedasticity," Econometrics, **31**, 307-327, 1986.

[24] P. Yu, T. Liu, Q. Ding, "Volatility analysis of web news and public attitude by GARCH model," Psychology, **3**, 610-612, 2012. DOI: 10.4236/psych.2012.38092

[25] C. Dritsaki, "An empirical evaluation in GARCH volatility modeling: evidence from the Stockholm stock exchange," Journal of Mathematical Finance, **7**, 366-390, 2017. DOI: 10.4236/jmf.2017.72020

[26] W. Zhang, P. Yang, "Research on dynamic relationship between exchange rate and stock price based on GARCH-in-mean model," Journal of Service Science and Management, **11**, 691-702, 2018. DOI: 10.4236/jssm.2018.116046

[27] A. Ngunyi, S. Mundia, C. Omari, "Modelling volatility dynamics of cryptocurrencies using GARCH models," Journal of Mathematical Finance, **9**, 591-615, 2019.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 1, 12-16 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Characterization and Investigating the Effect of Gate-Insulator Thickness on Co-Axial Cylindrical Carbon Nanotube Field Effect Transistor

Suchismita Sen[1], Argha Sarkar[2,*], Pinaki Chakraborty[1]

[1]*Department of Physics, Raiganj University, Raiganj, 733134, India*

[2]*School of Computer Science and Engineering, REVA University, Bangalore,560064, India*

ARTICLE INFO

Article history:
Received: 29 August, 2022
Accepted: 12 December, 2022
Online: 24 January, 2023

Keywords:
Carbon nanotube
FETToy
Carbon nano tube diameter
Characterization curve

ABSTRACT

Carbon nanotube field effect transistor (CNTFET) has a huge advantage over the Si-MOSFET. In MOSFET switching occurs by altering channel resistivity whereas in CNTFET switching occurs by modulation contact resistance. CNTFET generates three to four times of drive current than MOSFET. Transconductance of CNTFET is four times higher than the MOSFET. The average carrier velocity is also very high almost double in CNTFET than that is in MOSFET. Its power consumption is low. Electron mobility is high. Threshold voltage is also low. It has better control over channel formation. There is no direct tunneling and gate leakage current is also reduced. Herein, the main objective is to investigate the effect of gate-insulator thickness on CNTFET, and to optimize the thickness so that current carrying capacity may reach higher.A detailed simulations have been made and IV characterizationis done to investigate the effect of Gate-Insulator Thickness on Co-Axial Cylindrical Carbon Nanotube Field Effect Transistor.Report shows with the increasing gate-insulator thickness current is decreased significantly. Where as the variation of nano diameter shows that the increasing rate of current is increased when the carbon tube diameter is increased.

1. Introduction

In this age of nanotechnology, the demand of integrated circuits with smaller dimension has increased. On the other side this fast-moving world requires technology with high speed performance and that can consume lower power. To fulfill such requirements, the use of carbon nanotube field effect transistor (CNTFET) over Si- MOSFET has increased widely. Its ability to carry high current makes it more popular [1]. Carbon nanotubes consist of carbon atoms having diameter in nanometer range [2]. The considerable tiny sized carbon and its electronic configuration ensure unique carbon element with versatile structures and alluring properties [2]. Having the title of the strongest material ever measured, graphene is a two-dimensional (one-atom-thickness) allotrope of carbon with a planar honeycomb lattice [3]. It is regarded as the basic building block of carbon nanotubes. The versatile properties of carbon nanotubes (CNT) basically sourced from graphene [2]. Folding one or multiple graphene sheets with a specific chiral angle creates unique CNT.

Based on the number of folded layers' carbon nanotubes can be classified in two types.

- Single – walled carbon nanotube (SWNTs), having diameter 1nm [4]
- Multi – walled carbon nanotube (MWNTs), having diameter 100 nm

In multilayer formation many layers are interlinked. On the other hand, another classification of CNTFETs can be mentioned based on its geometry.

In a back-gate CNTFET generally SWCNT is used. It was first proposed by Tans et.al. [5]. The I(on)/ I(off) ratio of this type of CNTFET is almost 105 [6]. The parasitic contact resistance of such CNTFET is very high (>1Mohm) [7]. On the other hand, the drain current as well as the value of transconductance is very low. Drain current is of the nano range [6]. Such limitations of back-gate CNTFET drive the researchers to develop a next generation CNTFET.

*Corresponding Author: Argha Sarkar, Email: argha15@gmail.com

Wind et al. have come up with the first top gate CNTFET [8]. In this model the gate is formed over the carbon nanotubes. Though the fabrication process of Top gate CNTFET is little complicated but it is preferred over back-gate CNTFET due to its high drain current of the order of micro and for the greater value of transconductace.

Unlike the other two CNTFETs in Wrap around gate CNTFET the whole nanotube is covered by gate. It is also known as Gate-all-around CNTFET. To expose the ends of the nanotube the wrapping is partially etched and then the source, gate and drain contacts are deposited on the nanotubes. As the entire carbon nanotube is covered, it reduces the leakage current and increases the electrical performance.

In suspended CNTFET method, gate is suspended over a trench to reduce the contact with substrate and gate oxide and it improves the device performance. But the main drawback of such type of CNTFET is here air or vacuum is considered as the dielectric medium. Only short CNTs are used as long tubes may short the device by touching the metal contact.

Depending on the type of electrodes used, the CNTFET classification has been made into three categories. (a) Schottky-barrier (SB) CNTFET (b) Partially gated (PG) CNTFET and (c) doped-S/D CNTFET [9-16]. And the differences are clearly shown in Figure 1.

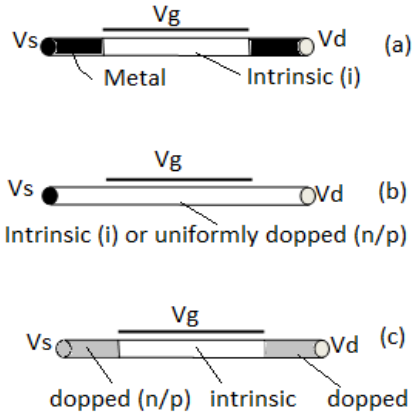


Figure 1: Different types of CNTFET: (a) Schottky-barrier (SB) CNTFET, (b) partially gated (PG) CNTFET (c) doped-S/D CNTFET [9].

Here we have focused on the co-axial cylindrical CNTFET. In such CNTFETs an oxide layer is portrayed arround a cylindrical carbon nano tube. Further a metallic cylindrical layer is deposited in it. This metallic contact can behave as a gate here. At a fixed bias voltage it can induce more channel charge than the other CNTFETs. This is because of its geometry. The capacitive coupling between the gate electrode and the nanotube surface is the maximum for it. Technologies like complementary metal-oxide-semiconductor (CMOS) can be affected by the short channel effects. This improved coupling can prevent this short channel effects. Its geometry of end contact is also important as it can provide the concept of the dimension of Schottky barrier. This Schottky barrier is actually present at the channel near device ends and it can directly influence the current modulation. It also has a huge role in low voltage applications.

2. Simulation of Co-Axial Cylindrical Carbon Nanotube Field Effect Transistor

The In this paper simulations were done for co – axial cylindrical CNTFET using the well-known FETToy tool to see how the characteristic curves depends on different tube parameters like nanotube diameter and gate insulator thickness. Varying the Gate Oxide thickness and nano-tube diameter the drain current can be varied. On the other hand the scaling of the most popular Si-MOSFET almost approaches towards its limiting values. In search of new alternatives this simulation was done to overcome these limitations.

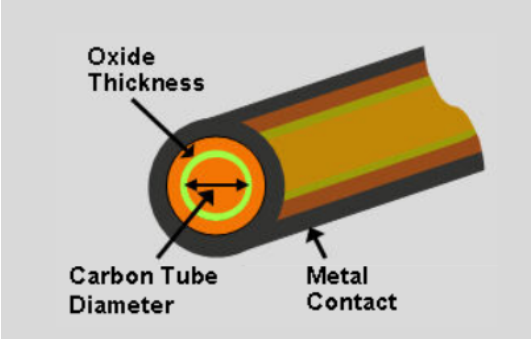


Figure 2: co-axial cylindrical CNTFET

Figure 2 represents a schematic diagram of co-axial cylindrical CNTFET. First for a constant nanotube diameter (1 nm), the simulation was done by varying the gate insulator thickness only. For this simulation the ambient temperature was taken as 300K. Threshold voltage of the used CNTFET was 0.32 V whereas the gate control parameter and drain control parameter were considered as 0.88 and 0.35 respectively. Series resistance was taken zero.

3. Prepare Your Paper before Styling

Before Some data taken during the simulation are given below.

Table 1: (Drain voltage=1 V, Nanotube diameter=1nm)

Gate Voltage (V)	Drain Current (μA)		
	Gate-Insulator thickness 1nm	Gate-Insulator thickness 1.5nm	Gate-Insulator thickness 2nm
0	0.00672	0.0000661	0.0000661
0.0833	0.114	0.00112	0.00112
0.1667	1.9	0.0187	0.0186
0.2500	24.4	0.24	0.232
0.3333	123	1.23	1.13
0.4167	296	3.01	2.69
0.5000	517	5.41	4.75
0.5833	778	8.37	7.26
0.6667	1070	11.9	10.2
0.7500	1400	15.9	13.7
0.8333	1750	20.4	17.5
0.9167	2140	25.1	21.6
1.0000	2550	30	26

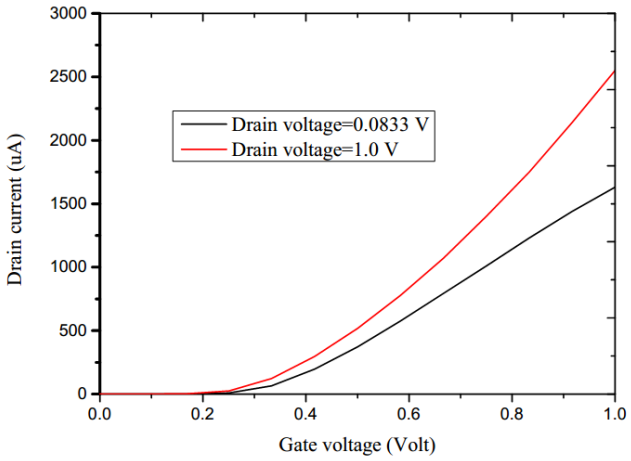


Figure 3: (a) Gate Voltage Vs Drain Current curve for a CNT having diameter 1nm and gate insulator thickness 1nm.

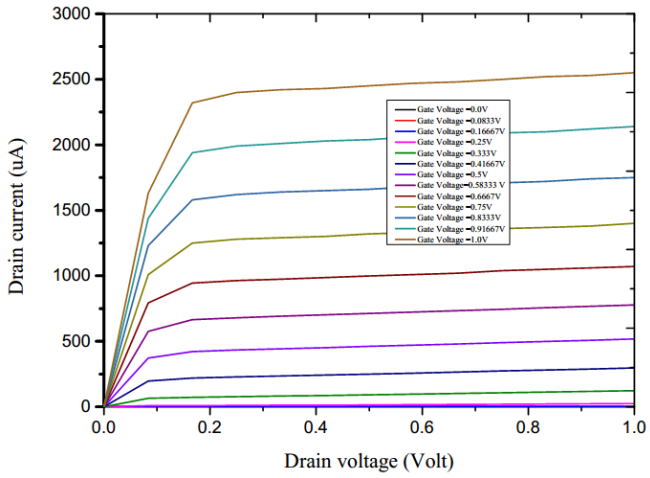


Figure 3: (b) Output Characteristic curve for a CNT having diameter 1nm and gate insulator thickness 1nm .

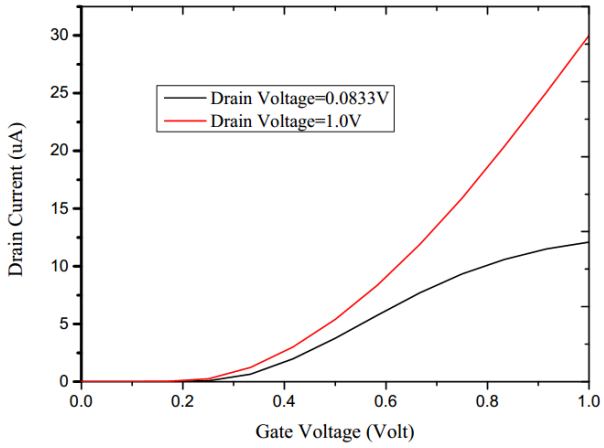


Figure 3: (c) Gate Voltage Vs Drain Current curve for a CNT having diameter 1nm and gate insulator thickness 1.5nm.

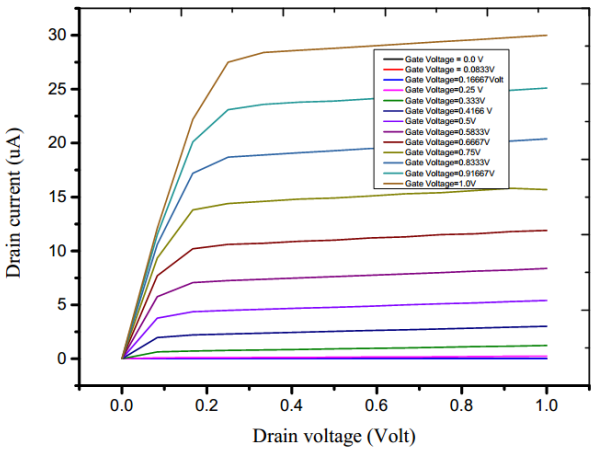


Figure 3: (d) Output Characteristic curve for a CNT having diameter 1nm and gate insulator thickness 1.5nm.

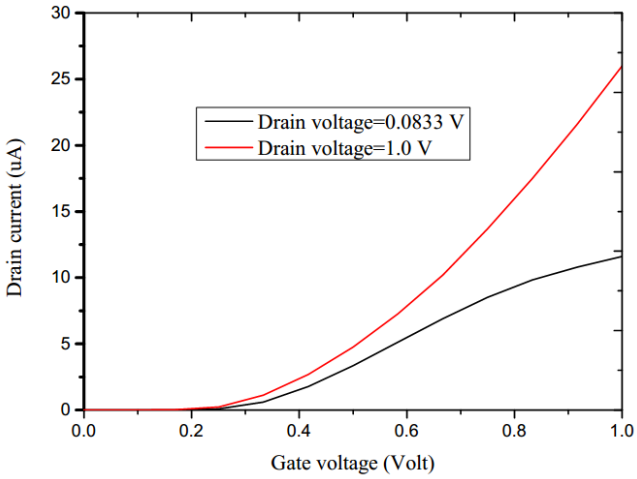


Figure 3: (e) Gate Voltage Vs Drain Current curve for a CNT having diameter 1nm and gate insulator thickness 2nm.

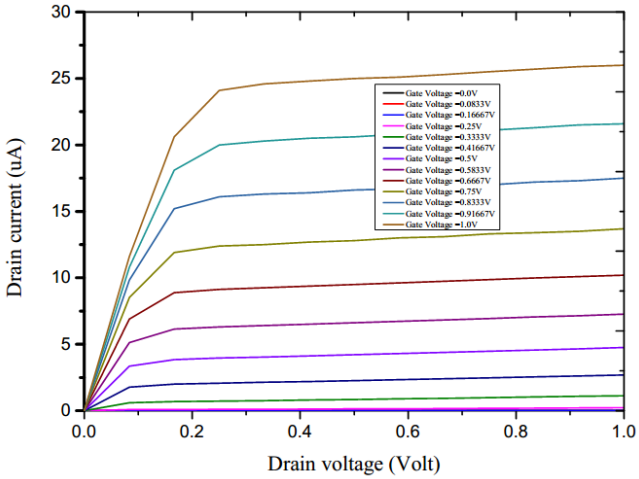


Figure. 3: (f) Output Characteristic curve for a CNT having diameter 1nm and gate insulator thickness 2nm.

From the data of Table-1 it is clear that the drain current vs gate voltage characteristic curve for 1nm gate-insulator thickness is steeper than the other two. Whereas for the same given voltage CNTFET having gate-insulator thickness 2nm gives the lowest drain current. So the drain current is decreased with the increasing gate-insulator thickness. This is due to the fact that with the increase of the gate-insulator thickness the resistance across it is also increased and as a result the drain current is decreased [12]. When the gate dielectric becomes thicker, the electric field within the dielectric becomes smaller for the same gate voltages. Thus the accumulated free carrier near the interface also becomes less. As the carrier density decreases, the drain current decreases as well. Figure 3 shows the input and output characteristic curves for CNTs having different gate insulator thickness but same nanotube diameter i.e. 1nm. Another simulation (shown in figure 4) was done by taking nanotube diameter as 2 nm and varying the gate-insulator thickness from 1nm to 2nm. Other aspects were fixed. I-V characterization is made to investigate the Effect of gate-insulator thickness on co-axial cylindrical CNTFET.

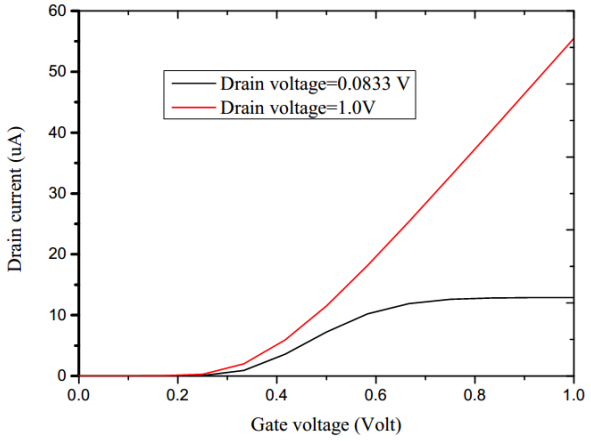


Figure 4: (a) Gate Voltage Vs Drain Current curve for a CNT having diameter 2nm and gate insulator thickness 1nm

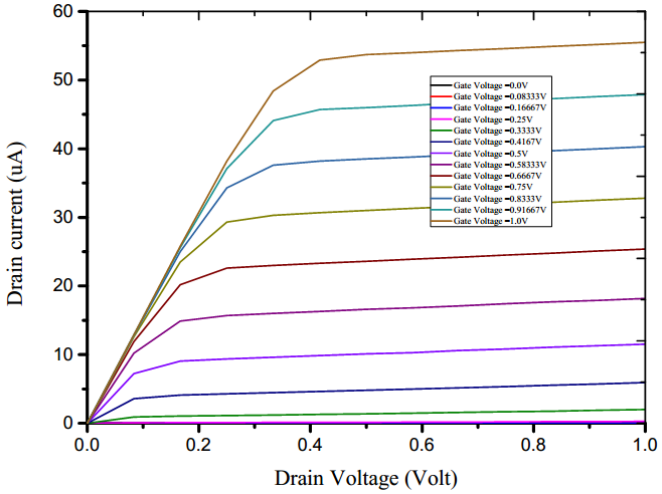


Figure 4: (b) Output Characteristic curve for a CNT having diameter 2nm and gate insulator thickness 1nm

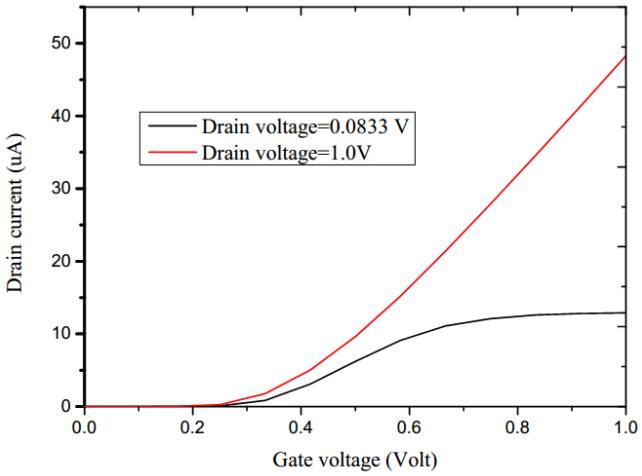


Figure 4: (c) Gate Voltage Vs Drain Current curve for a CNT having diameter 2nm and gate insulator thickness 1.5nm

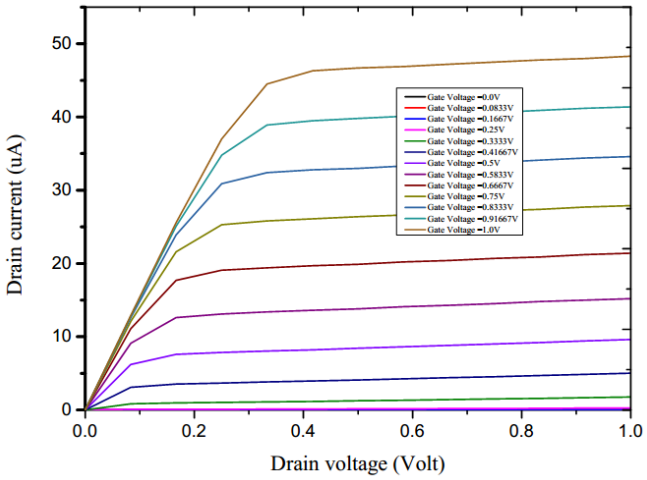


Figure 4: (d) Output Characteristic curve for a CNT having diameter 2nm and gate insulator thickness 1.5nm

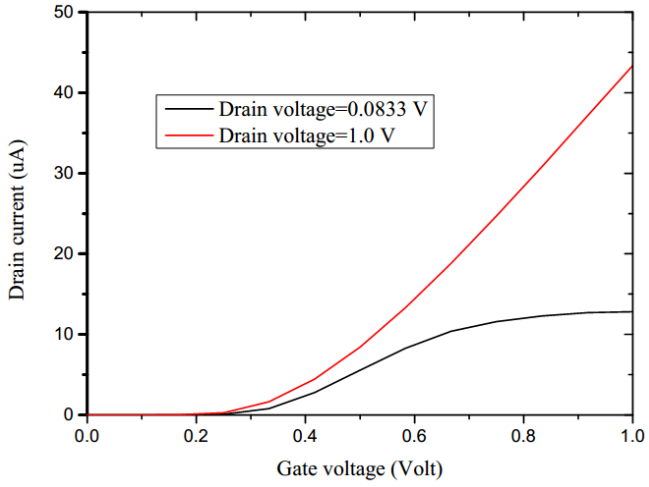


Figure 4: (e) Gate Voltage Vs Drain Current curve for a CNT having diameter 2nm and gate insulator thickness 2nm

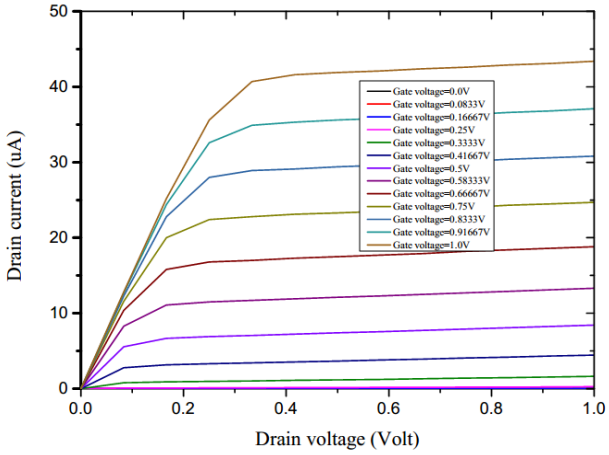


Figure 4: (f) Output Characteristic curve for a CNT having diameter 2nm and gate insulator thickness 2nm

Table 2: (Drain voltage =1 V, Nanotube diameter=2nm) [11]

Gate Voltage (V)	Drain Current (μA)		
	Gate-Insulator thickness 1nm	Gate-Insulator thickness 1.5nm	Gate-Insulator thickness 2nm
0	0.0000661	0.0000661	0.0000661
0.0833	0.00112	0.00112	0.00112
0.1667	0.0189	0.0189	0.0189
0.2500	0.28	0.271	0.265
0.3333	1.99	1.78	1.64
0.4167	5.93	5.02	4.46
0.5000	11.5	9.62	8.43
0.5833	18.2	15.2	13.3
0.6667	25.4	21.4	18.8
0.7500	32.8	27.9	24.7
0.8333	40.3	34.6	30.8
0.9167	47.9	41.4	37.1
1.0000	55.5	48.3	43.4

4. Conclusion

Here we can see that just by changing the nanotube diameter and gate-insulator thickness the drain current can be changed. For the nanotube diameter 1nm and gate-insulator thickness 1nm we get a huge drain current compare to the other combinations. Simulation results ensure the effect of gate dielectric increment in terms of decrement in drain current. It is due to reduction in the electric field for the same gate voltages. Thus the decay in carrier density and drain current.

This CNTFET also has a huge advantage over the Si-MOSFET. In MOSFET switching happens by altering channel resistivity whereas in CNTFET switching is due to modulation contact resistance. CNTFET generates three to four times of drive current than MOSFET. About quadruple higher transconductance of CNTFETs than MOSFETs comes from the band structure and improved mobility. The average carrier velocity is also very high almost double in CNTFET than that is in MOSFET. Its power consumption is low. Electron mobility is high. Threshold voltage is also low. It has better control over channel formation. There is no direct tunnelling and gate leakage current is also reduced.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Raychowdhury, A., & Roy, K. Carbon nanotube electronics: design of high-performance and low-power digital circuits. IEEE Transactions on Circuits and Systems I: Regular Papers, **54**(11), 2391-2401, 2007, 10.1109/TCSI.2007.907799

[2] A. Sarkar, S. Maity, P. Chakraborty , S. K. Chakraborty , "Characterization of carbon nanotubes and its application in biomedical sensor for prostate cancer detection," American Scientific Publishers, 17, 17-24, 2019, https://doi.org/10.1166/sl.2019.4039

[3] A. Sarkar, M. Sreenath, K. Srinivas and NV Teja, "Investigation of Graphene as a Sensing Layer for Future Prostate Cancer Biosensing Applications," Journal of Physics: Conference Series, **1921**(1), 2021, 10.1088/1742-6596/1921/1/012038

[4] T. W. Odom, J. L. Huang et al.,"Atomic structure and electronic properties of single-walled Carbon Nanotubes," Nature, Nature, **391**(6662), 62-64, 1998, https://doi.org/10.1038/34145

[5] S. J. Tans, A. R. M. Verschueren and C. Dekker, "Room-temperature transistor based on a single carbon nanotube," Nature , **393**(6680), 49-52, 1998, https://doi.org/10.1038/29954

[6] S. Rasmita and R. R. Mishra, "Simulations of carbon nanotube field effect transistors", International Journal of Electronic Engineering Research **1**(2), 117-125, 2009.

[7] N. T. Rouf, A. H. Deep and R. B. Hassan, Current-voltage characteristics of carbon nanotube field effect transistor considering non-ballistic conduction, BRAC University Dhaka-1212, Bangladesh.

[8] S. J. Wind, J. Appenzeller, P. Avouris, "Lateral scaling in carbon nanotube field effect transistors", P Physical Review Letters, **91**(5), 058301, 2003, https://doi.org/10.1103/PhysRevLett.91.05830

[9] T. Dang, L. Anghel, R. Leveugle, "CNTFET Basics and Simulation. International Conference on Design and Test of Integrated Systems in Nanoscale Technology," IEEE Explore 2006.

[10] J. Guo, S. Datta and M. Lundstrom, "Assessment of silicon mos and carbon nanotube fet performance limits using a general theory of ballistic transistors," IEEE 2015.

[11] A. Rahman, J. Wang, J. Guo, Md. S. Hasan, Y. Liu, A. Matsudaira, S. S. Ahmed, S. Datta, M. Lundstrom, FETToy, 10254/nanohub-r220.4, 2006.

[12] M. Radosavljevic, S. Heinze, J. Tersoff, and P. Avouris, "Drain voltage scaling in carbon nanotube transistors", Applied Physics Letters, **83**(12), 2435-2437., 2003, https://doi.org/10.1063/1.1610791

[13] M. Zhang, P. C. H. Chan, Y. Chai, "Novel Local Silicon-Gate Carbon Nanotube Transistors Combining Silicon-on-Insulator Technology for Integration," IEEE transactions on nanotechnology, **8**(2), 260-268,2009, 10.1109/TNANO.2008.2011773

[14] Compano, R. ed. "Technology Roadmap for Nanoelectronics," Microelectronics Advanced Research Initiative, 2000.

[15] P. Sagar P., Handa A., Kumar G., Gupta V. Nanocomposite hydrogel materials for defective cartilage repair and its mechanical tribological behavior, A Review. Paper and Biomaterials, **7**(3), 63-72, 2022, https://doi.org/10.1213/j.issn.2096-2355.2022.03.007

[16] P. Sagar P., Handa A, Kumar G. (2022), Metallurgical, mechanical and tribological behavior of Reinforced magnesium-based composite developed Via Friction stir processing, Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering, **236**(4), 1440-1451, https://doi.org/10.1177/09544089211063099

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 1, 17-29 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Design of an Open Source Anthropomorphic Robotic Hand for Telepresence Robot

Jittaboon Trichada, Traithep Wimonrut, Narongsak Tirasuntarakul, Eakkachai Pengwang*

Institute of Field Robotics, King Mongkut's University of Technology Thonburi, Bangkok, 10140, Thailand

ARTICLE INFO

Article history:
Received: 28 September, 2022
Accepted: 18 December, 2022
Online: 24 January, 2023

Keywords:
Open source
Anthropomorphic Robotic Hand
Low cost

ABSTRACT

Most anthropomorphic robotic hands use a lot of actuators to imitate the number of joints and the movement of the human hand. As a result, the forearm of the robot hand has a large size for the installation of all actuators. This robot hand is designed to reduce the number of actuators, but also retain the number of movable joints like a human hand by using the four-bar linkage mechanism and only flexion-extension movements. This stamen is added in the problem statement according to the reviewer's comment. The special features of this robotic hand are the ability to adjust the link length and the range of rotation for each joint to suit various applications and can fabricate with 3D printing and standard parts with costing about $750. All hardware CAD files and equations are published on the GitHub website, which benefits for researchers to utilize as an open-source approach that their project might be further expanded in the future. The anthropomorphic robotic hand has five fingers, 16 joints, and 12 active Degrees of Freedom (DOFs) with 12 servo motors applied to finger motion and one for wrist motion. The structure of the hand is designed using the average of Asian human hands in combination with the golden ratio. All servo motors are installed in the forearm designed in a ventilated structure with 12V vent exhaust fan motor to stabilize the operating temperature of the robotic hand. Size and weight of the hand included with the forearm are 20×54×16.5 centimeters and 2.2 kilograms respectively. The hand has achieved human-like movement by using a four-bar linkage mechanism and tendon with PTFE tube to guide operation path of the tendon with the lowest friction force. This paper presents the design processes, the experimental set-up, and the evaluation of the finger movements. From the experiment of grasping objects, this hand was able to grasp 10 basic grasp types including 32 different objects, perform 9 common gestures, and lift the object to 450 grams. From this paper, the kinematic equation is proved that the designed finger structure can move exactly as the equation with maximum error of repeatability test around 1.6 degrees.

1. Introduction

This paper is an extended paper of our work initially presented in 2021 4th International Conference on Robot Systems (ICRSA 2021) [1]. The technology that the operator interacts with the robot remotely is called the telepresence robot. This technology allows the human remotely to control the robotic end effector system in a human environment with experience as they locate there. In order to elaborate the sensibility of the operator, the design should be similar in structure and scale to the human. One of the main systems of the telepresence robot is an anthropomorphic robotic hand enabling a user to interact with the remote environment realistically.

The human hand is one of the best grippers in the world which can interact with and perceive the physical environment. However, the anatomy of the human hand is complicated to be implemented in robotics field due to size, proportions and mechanisms. For example, each person has a different size and length of each finger according to their genes [2]. The focus of this paper lies in the average hand size of Asian people, which is similar to Thai people. Thus, this paper using the average hand size of Koreans (167 males) and the Golden ratio [3-5] to design robotic hands.

Generally, the robotic hand is numerous in the design in terms of the number-type of actuator, active hand DOFs, total hand joints, power transmission, sensors, and price. The price of each robot hand varies. The cost of robotic hands has ranged between $1,500 and $150,000, depends on the number and type of actuator, sensor, and application programming interface (API) [6]. In

*Corresponding Author: E. Pengwang, KMUTT, Bangkok, 10140, Thailand, (+66)24709339 eakkachai@fibo.kmutt.ac.th

addition, most robotic hands measure only grasping and performing various gestures, but this paper has added the repeatability experiments of robotic hands and the accuracy verification of the equations provided in the conference paper.

In this article, we designed an anthropomorphic robotic hand based on the anatomical structure that use the kinematic equation to calculate the length of the four-bar linkage (l_4) of four common fingers. Before that, the designer must define basics of the configuration, including the range of motion of each joint ($\theta_1, \theta_2, \theta_3$), and initial degrees ($\alpha_2, \alpha_3$) as shown in Table 1 and Figure 1. Since the average Korean hand size informed the total length of each finger, a golden ratio is required to devided the average length to each phalanx (l_1, l_2, l_3). The purpose of this paper is to design a new open-source robotic hand with a low cost ($750) by using standard parts and 3D printing techniques and publish it on the GitHub website. Moreover, the performances of the design of low-cost robot hand are high performance from the tested as shown in the experiment section. The performances of this hand are proven from five experiments: grasping experiment, gesture experiment, motor temperature experiment, structure experiment, and repeatability experiment.

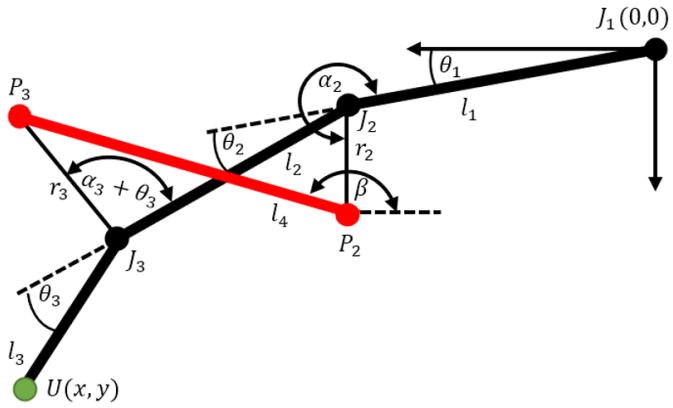


Figure 1: The Simplified Image of Finger and Parameters

Table 1: Parameter of Equations

J_1	Proximal joint	θ_1	ROM of J_1
J_2	Middle joint	θ_2	ROM of J_2
J_3	Distal joint	θ_3	ROM of J_3
L_1	Proximal phalanx	r_2	Distance of P_2J_2
L_2	Middle phalanx	r_3	Distance of P_3J_3
L_3	Distal phalanx	α_2	Initial degree of r_2
L_4	Inner link1	α_3	Initial degree of r_3

2. Design

In fact, each joint of the human hand can move in two directions: flexion-extension and abduction-adduction. The flexion-extension movement is increase or decrease the angle between the bones of the limb at a joint while the abduction-adduction motion is away or toward the midline of the body as shown in Figure 2.

In this article, the anthropomorphic robotic right hand is 200 mm wide, 225 mm long and the forearm is 145 mm wide, 315 mm long. The robot hand contains four common fingers (differs in length of each joint), thumb, palm, wrist, forearm, and skin of fingertips as shown in Figure 3. All fingers were designed for only flexion-extension movement. The wrist part has only one actuated joint, the thumb has three actuated joints, and the other fingers have two actuated joints per finger. Figure 2 depicts the MCP, PIP, and DIP joints on each finger, excluding the thumb. All actuated joints use a cable with PTFE tube to control a position, and the underactuated joint uses a linkage mechanism to drive the DIP joint movement that relates to the PIP joint. All servo motors are installed in the forearm designed in a ventilated structure with a 12V 4000rpm exhaust fan. The total weight is about 2.2 kilograms.

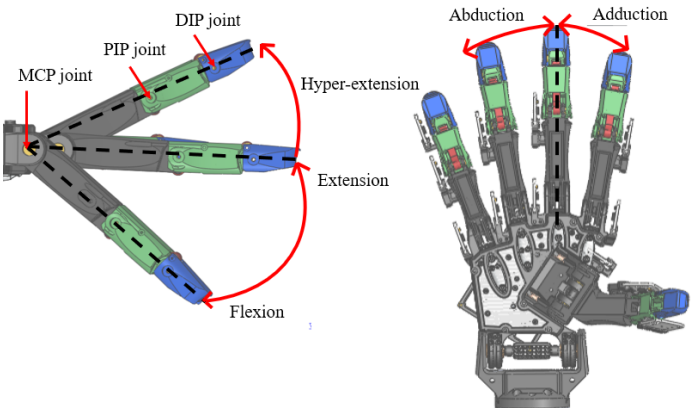


Figure 2: The Finger Movement

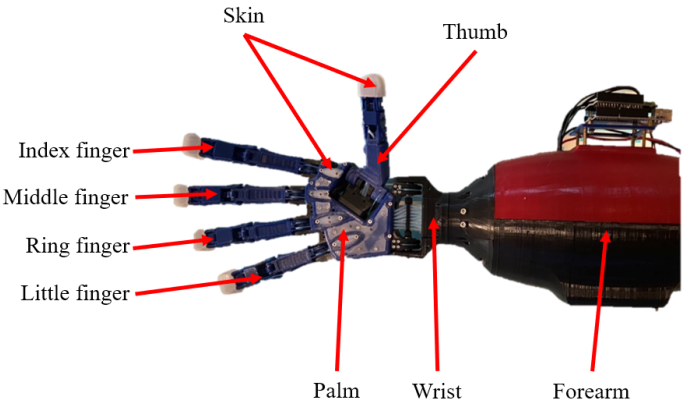


Figure 3: The Robot Hand with Forearm and Skin

2.1. Finger Design

This finger is a development from the finger published in the conference ICRSA 2021. Previously, tendons were routed inside robot fingers to protect tendons from the external environment, such as objects that are caught or touched that can damage the tendon. From Figure 4 (a), the distance and degree of pulling the tendons are not constant when compared with Figure 4 (b).

Because the inner finger has a very tiny space, the radius of the tendon for pulling the joints of the finger was less. To tighten the grip, the tendons used to transmit the force are shifted to the outside near the screw attachment to keep the degrees and pull distance of tendons constant throughout the operation, as shown in Figure 4. Inside the finger, there is a cavity for routing the tendon to the internal finger before attaching the PIP joint of the finger.

Since each joint has a shaft, tendons are unable to pass directly through the MCP joint. This design is enabling the tendons to pass as close as possible to the MCP joint to have the least moment of force caused by the pulling of PIP joint, and PTFE tubes can be inserted to reduce the friction force in pulling the tendons as shown in Figure 5.

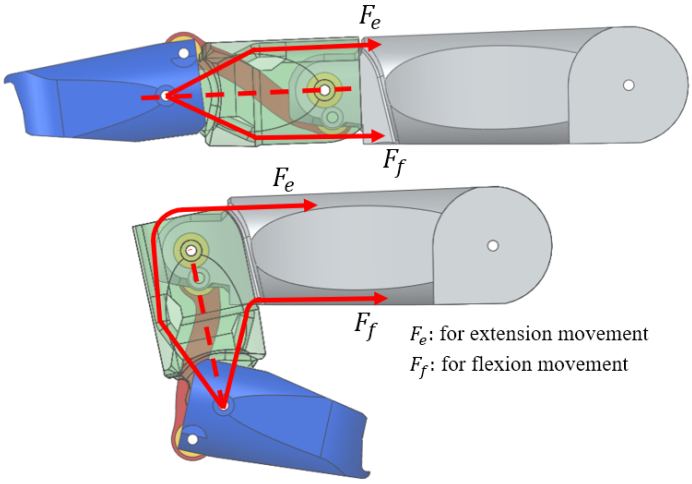


Figure 4 (a): The Index Finger and Routing Tendon Inside the Finger

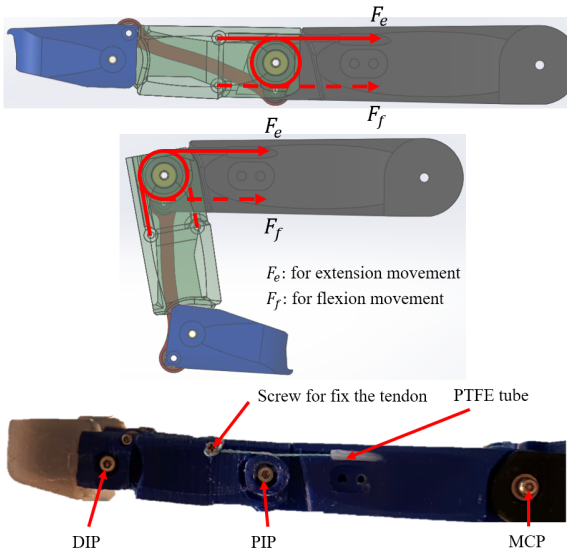


Figure 4 (b): The Index Finger with Tendon Routing Outside and Fingertip's Skin

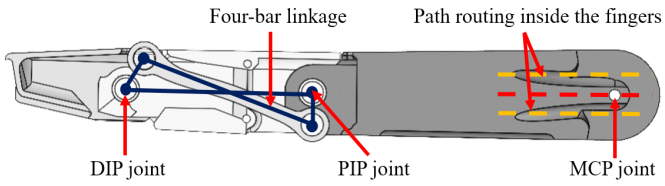


Figure 5: The Four-bar Linkage Mechanism and Path Routing in the Finger

Because the average size of the Korean hand only defines the length of each finger, the golden ratio is required to calculate the length of each phalanx. The distal phalanx of each finger is very small of different lengths (21.15 mm, 22.01 mm, 21.08 mm, 18.45 mm), making it unable to be installed with a four-bar linkage mechanism. This problem can be solved by using an equal length of each distal phalanx. The length of distal phalanx that can be achieved is 21.37 mm. Each finger's length and range of motion are uniformly designed, as shown in Tables 2, 3, and Figure 6 respectively.

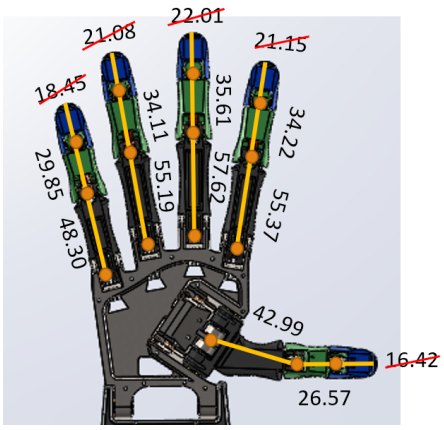


Figure 6: The Length of Each Phalanx

Table 2: Length of Each Phalanx [mm.]

No.	Name	Length (mm.)
1	Distal phalanx-all fingers	21.37
2	Middle Phalanx-Index	34.22
3	Middle Phalanx-Middle	35.61
4	Middle Phalanx-Ring	34.11
5	Middle Phalanx-little	29.85
6	Proximal Phalanx-Thumb	26.57
7	Proximal Phalanx-Index	55.37
8	Proximal Phalanx-Middle	57.62
9	Proximal Phalanx-Ring	55.19
10	Proximal Phalanx-little	48.30
11	Metacarpal Phalanx-Thumb	42.99
12	Distal phalanx wide-all fingers	18.00
13	Distal phalanx thin-all fingers	13.00
14	Middle phalanx wide-all fingers	20.50
15	Middle phalanx thin-all fingers	14.50
16	Proximal phalanx wide-all fingers	20.50
17	Proximal phalanx thin-all fingers	16.00
18	Metacarpal phalanx wide-all fingers	23.50
19	Metacarpal phalanx thin-all fingers	15.50

Table 3: The Range of Motion of Each Joint [degree]

Name	DIP joint	PIP joint	MCP joint	DIP-T joint	MCP-T joint	CMC-T joint
Range of Motion	0 to 80	0 to 90	-10 to 90	0 to 100	0 to 100	-10 to 120

The human hand contains at least 23 degrees of freedom (DOFs) [7]. Human fingers have 3 joints with 4 DOFs: 3 DOFs for flexion-extension movement and 1 DOF for adduction-abduction

movement. In the case of the thumb, there are 3 joints with 5 DOFs: 3 DOFs for flexion-extension movement and 2 DOFs for adduction-abduction movement (Figure 7). The wrist has 2 DOFs for flexing and expanding (Figure 8). The robotic hand has five fingers, 16 joints, and 12 active DOFs (only flexion-extension movement) with 12 servo motors. Four common fingers excluding the thumb use 2 servo motors per finger to control the PIP joint and MCP joints, while the DIP joint moves related to the PIP joint by using the four-bar linkage mechanism as shown in Figure 5. The thumb uses 3 servo motors and 1 servo motor for the wrist. Each movable joint use bearing to reduce the friction force during finger movement. The MCP joint of 4 fingers and the CMC joint of thumb use 2 mm inner diameter bearing. Others moving joints in each finger use 1.5 mm inner diameter bearing. The length of the four-bar linkage of the four fingers was calculated from equations r_3 and l_4 in the conference as shown in Table 4. Each four-bar linkage structure has a different concave curvature to prevent the linkage from a collision with the shaft of each joint while operating (Figure 9).

Table 4: Length of Each Four-bar Linkage [mm.]

Name	Length (mm)
Linkage Index	33.04
Linkage Middle	34.34
Linkage Ring	32.93
Linkage Little	28.98

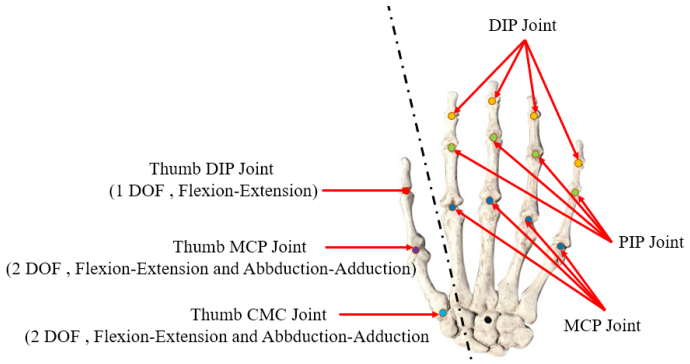


Figure 7: The Anatomy of Human Hand

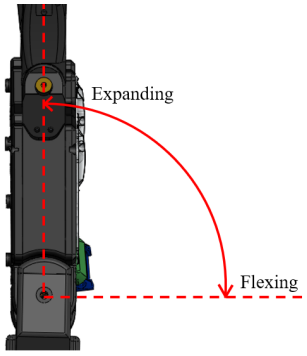


Figure 8: The Wrist Movement

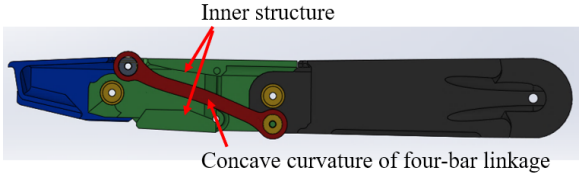


Figure 9: The Curvature of Four-bar Linkage

Many robotic hands have various thumb designs to accommodate the suitability of the hand, such as the number of DOFs, actuators, power transmission, and functionalities; thus, the design of the thumb is rarely defined. The main contribution for this paper is separated into 2 sections: the joint design, and the structure design. The joint design of the thumb is the same design as the controllable joint of four fingers which has a constant both distance and degree for pulling as shown in Figure 10. Normally, the thumb grasps an object by using both abduction-adduction movement and flexion-extension movement. The flexion-extension movement of thumb is used to press objects towards the palm of the hand while the abduction-adduction movement of thumb is used to grasp the various objects. Even though the abduction- adduction motion is very important for grasp [8], this robot hand challenges to design by using only flexion-extension movement to achieve the purpose of this hand in order to reduce the number of motors to be installed on forearm. From the reduction to 1 DOF per joint of the thumb, the importance of thumb angles must be emphasized, which affects the grasping performance of the hand. The motion analysis in SolidWorks program is required to test basic gestures and basic grasps such as handfuls, index-thumb, middle-thumb, cylindrical grasp, and spherical grasp before forming. From the analysis, the best angle for structure design of the thumb is 58.5 degrees in the top view and 28.5 degrees in the front view (Figure 11) to perform as many different gestures and basic grasps as possible in motion analysis of the SolidWorks program.

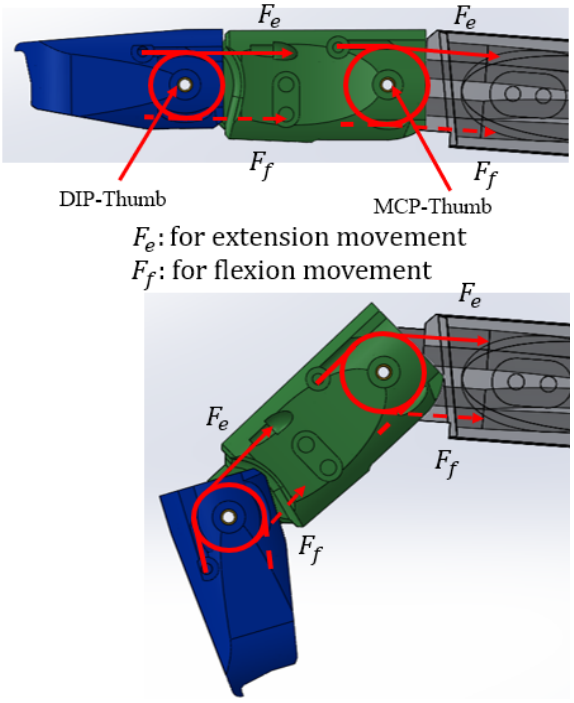


Figure 10: The Joint Design of the Thumb

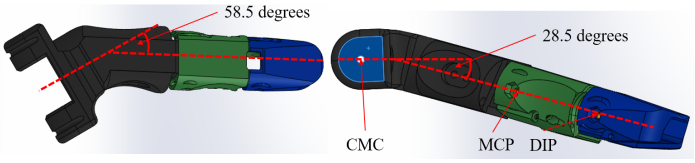


Figure 11: The Attachment Angle of the Thumb

2.2. Palm and Wrist Design

Generally, human hands can grasp objects by using the palm, all fingers, wrist, and skin. The palm is the part that connects the four fingers, thumb, and wrist that is coated by human skin. In the robotic field, the robot finger with no adduction-abduction movement was fixed angle between neighbor MCP joint of each finger at about 12 degrees [9]. The angle of neighbor MCP joint allows the hand to tightly grasp objects, increasing the area of routing tendons inside the hand, and decreasing the bending of PTFE tube around the MCP joint of the index finger in the palm. Each robotic hand has a different angle between neighbor MCP joint depending on the hand. From the analysis, it was found that 10 degrees is most suitable for this hand to gesture and grasp. The palm is designed by defining the middle finger as the reference point with 10 degrees of angle between neighbor MCP joints as shown in Figure 12.

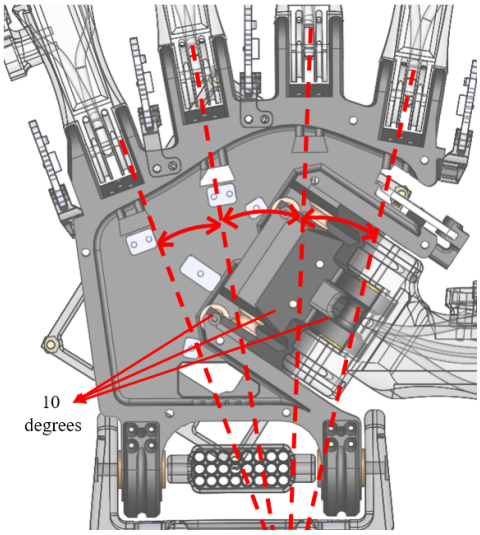


Figure 12: The Palm

In the palm, there is an apparatus that is used to arrange the PTFE tube and tendons from the wrist to each finger as shown in Figure 13. The CMC joint of the thumb was designed to angle the middle finger about 30 degrees (Figure 13) so that PTFE tubes inside the palm are easy to assemble and do not bend too much. The OK pose and check gesture can be performed by using this design. The OK posture is the index finger and thumb converge while the others are fully spread as Figure 14. The check posture is the index finger and thumb are fully spread while the others are fully bent same as the check symbol as shown in Figure 15. These 2 poses are the basic postures to hold objects and make various gestures of the hands. Since the direction of rotation of the CMC joint of the thumb is in a different direction of the tendon movement to other fingers, The bearing must be provided for changing the direction of pulling the tendons of the thumb as shown in Figure 16.

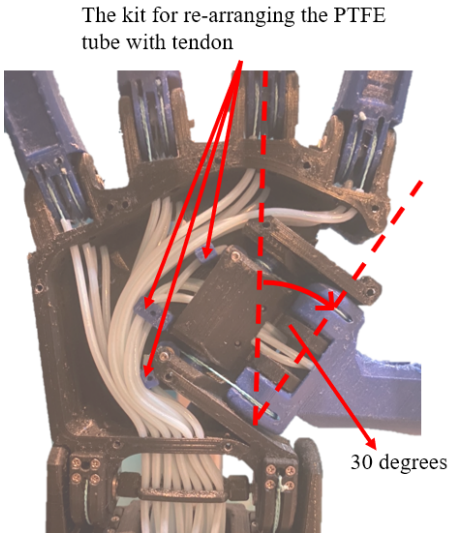


Figure 13: The Re-arranging of the PTFE Tube in the Palm

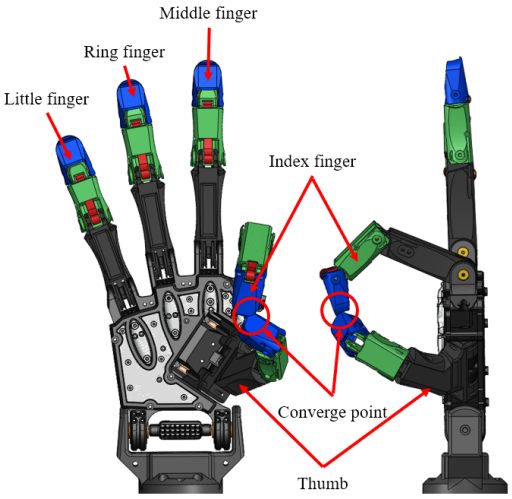


Figure 14: The Robotic Hand Performs OK Gesture

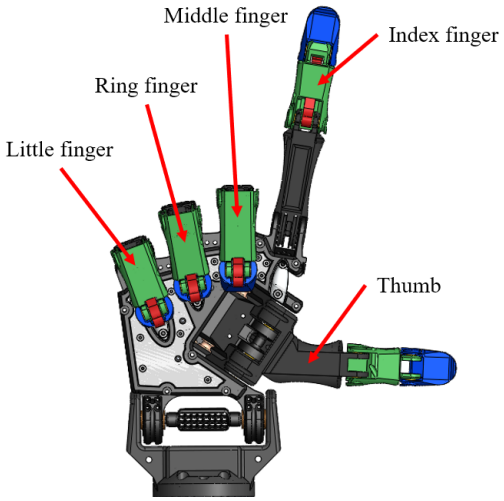


Figure 15: The Robotic Hand Performs Check Gesture

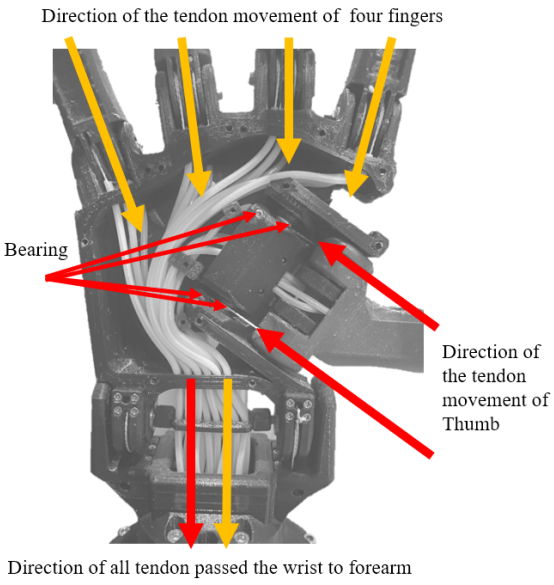


Figure 16: Direction of Tendon Movement in the Palm

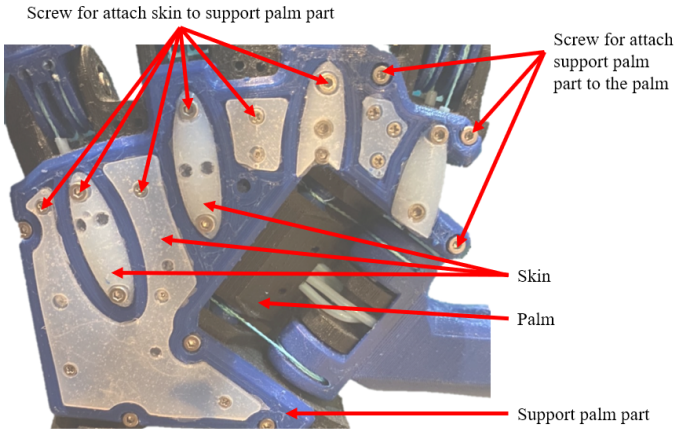


Figure 17: The Support Palm Part

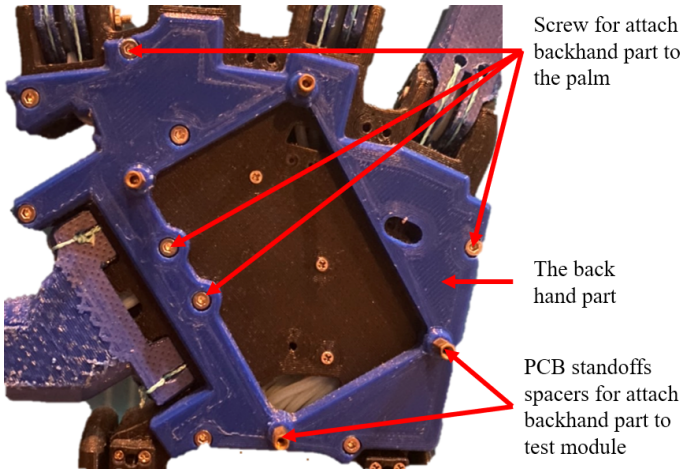


Figure 18: The Back Hand

The support palm part is the connector between palm and skin by using the screw (Figure 17). The backhand part attached behind the palm is used to install the test module (PCB for the connection wire of the encoder, multiplexer, and Arduino Uno board) by using PCB standoffs spacers and screw (Figure 18). The palm is assembled with an extension palm part on the wrist area to increase radius for pulling the tendon (Figure 19).

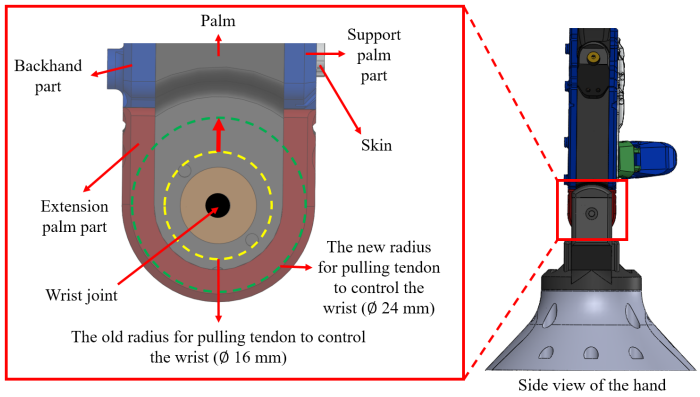


Figure 19: The Extension Palm Part

The wrist area has 3 components which are the main wrist, extension palm, and a support PTFE tube kit as shown in Figure 20. The main wrist is designed by integrating with PTFE tube to decrease friction force for pulling the tendon of wrist joint. In the middle of wrist joint is a support PTFE tube kit for re-arranging and supporting the PTFE tubes with tendons when the tendon is pulled by the motor. This support PTFE tube kit will keep the PTFE tube in place no matter how many degrees the wrist is tilted. The motor that controls the wrist joint will reduce the load caused by the tendon movement of all fingers. Main wrist is a connector between the hand and forearm that is used to re-arranging the PTFE tube same as a support PTFE tube kit and guides the tendon to attach the wrist joint. The range of motion of the wrist joint is from 0 to 180 degrees, the same as the wrist joint of human hand as shown in Figure 21.

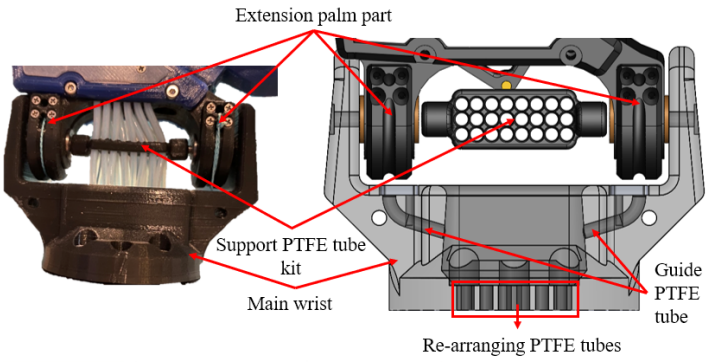


Figure 20: The Wrist

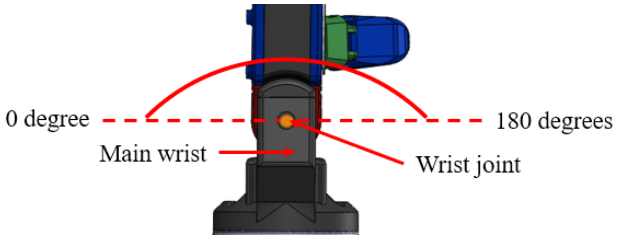


Figure 21: The Range of Motion of Wrist Joint

2.3. Skin of Fingertips and Palm

The entire skin is made of silicone (RA-22AB, RUNGART, Thailand) [10] forming with the use of PLA moles (3D printing). The skin of the palm consists of flat palm skin and half-ellipse palm skin (Figure 22). The thickness of flat palm skins and half-ellipse palm skins are about 1.5 mm and 4 mm respectively as shown in Figure 23. The skin of fingertips is the one of the main parts for grasping an object. Skin tips are used to increase friction force for tightening grip. These skin tips are wearable skin that has the same shape as fingertips with a thickness from the outside of the fingertips of about 2.5 mm as shown in Figure 24.

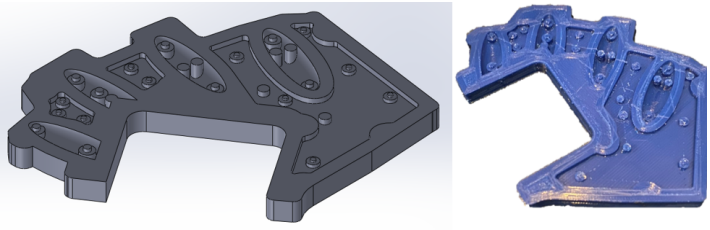


Figure 22: The Mold Skin of Palm

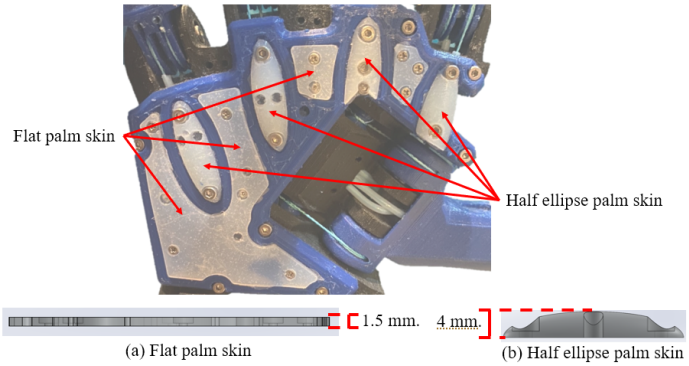


Figure 23: The Thickness of Palm Skin

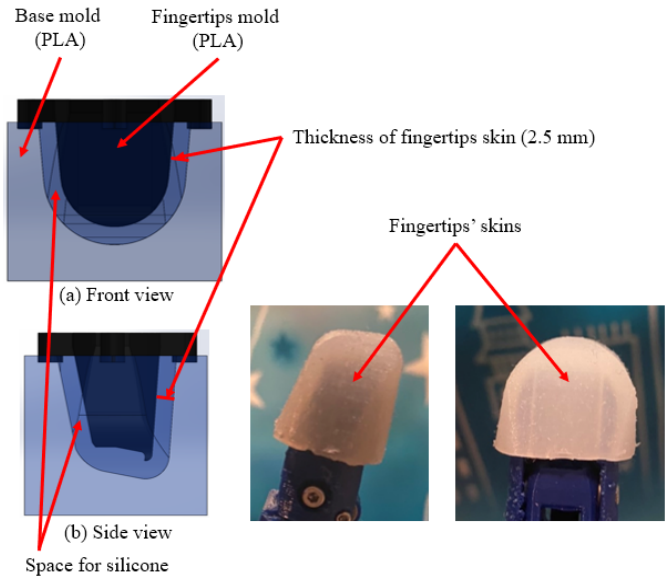


Figure 24: The Mold Skin of Fingertips

2.4. Forearm Design

The forearm contains the main forearm, connecting motor part, PTFE tube guide part, and cover. All servo motors, Arduino Uno board with Dynamixel shield, 12V fan, and PCB boards are installed in the forearm as shown in Figure 25. This Dynamixel servo motor (XL430 W250-t) has engineering plastic gear with an operating temperature from -5 to +72 °C. Long periods of heavy work of the motor will cause the accumulation of high temperature inside the forearm, so the design of the robot forearm must focus on temperature reduction to extend the motor life. The structure of the forearm must be a ventilated structure with a 12V vent exhaust fan to stabilize the operating temperature of the robotic hand. All PTFE tubes with tendons are routed following the PTFE guide from the motor in the forearm to each joint of the finger and wrist.

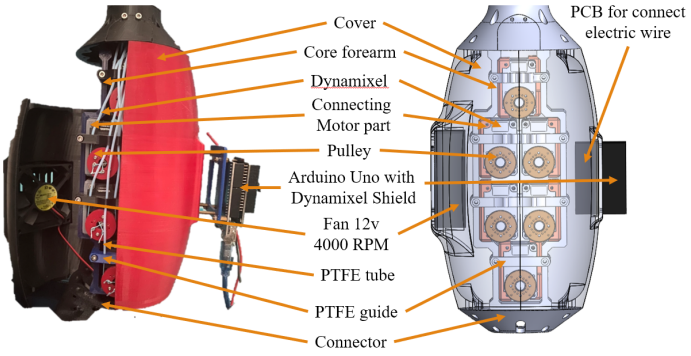


Figure 25: The Forearm

2.5. Actuation, Power Transmission and PLA Material

This hand uses 12 servo motors to control each actuated joint while underactuated joints (DIP joint of four fingers) move to follow the actuated joints (PIP joint of four fingers) by using a four-bar linkage mechanism. The actuators can transmit the power to control each joint by using a tendon with the PTFE tube. The PTFE tube [11] is used to protect the tendon and reduce the friction force between the tendon and the body part (PLA). The actuators of this robotic hand are XL430 W250-T from Dynamixel (TTL connection) because it has a suitable price with high torque and small size (Table 5). In addition, this actuator has precision to control with feedback sensors: position, current load, current temperature, etc. The tendon that is used to transmit the power from the actuators is a fishing tendon from Proberos. This tendon is made from 4 braids of Polyethylene (PE) with high max tension (36.2 Kg) and a small diameter (0.5 mm) costing about $1.5 per 100m as shown in Table 6. Because PLA filament has Young's modulus of about 3.04 GPa but ABS material has Young's modulus of about 1.97 GPa [12]. Therefore, the material for 3D printing parts uses PLA (Polylactic acid) filament with a cost (of $17.99 per Kg.) [13-14].

Table 5: Specification of Servo Motor [15]

Name	Dynamixel (XL430 W250-T)
Stall Torque	1.5 N.m
Stall Current	1.4 A
Weight	57 grams
Dimension	28.5 x 46.5 x 34 mm.

No Load Speed (at 12V)	61 rev/min
Resolution	4096 pulse/rev, 360 degree
Gear Ratio	258.5 : 1
Operating Temperature	-5 to +72 ℃
Connection	TTL
Feedback	Position, Load, Temperature, etc.
Material	Engineering Plastic
Price	$49.90

Table 6: Specification of Tendon [16]

Name	Tendon
Brand	Proberos
Material	Polyethylene (PE)
Number of tendons	4 braids
Outer Diameter	0.5 mm
Max Tension	36.2 Kg.
Price (100 m.)	$1.5

2.6. Cost Analysis

Generally, the range cost of a robotic hand is between $1,500 and $150,000. However, this robotic hand costs about $750 (as shown in Table 7) with 12 servo motors (Dynamixel) that can grasp objects and gestures similar to robotic hands that cost $1500 as proof in the results and experiment section. All parts of this robot hand are made from 3d print, designed for standard components that can be easily purchased locally and replaced. Moreover, robotic hands are designed to use as few motors as possible while keeping the ability to grasp various objects and gestures like other robotic hands as much as possible. The four-bar linkage mechanism is used to control the DIP joint related to the PIP joint in four fingers that can reduce one actuator per finger with the same number of movable joints as shown in the design section. Because all actuators are installed in the forearm, the low number of actuators in use saves costs and reduces the size and weight of the forearm. Thus, the price of this hand will be cheaper than the others.

Table 7: Price of Actuator and Material

Name	Amount	Total Price ($)
Dynamixel XL430 W250-T	12	598.80
PLA (eSUN) 1 Kg.	2	35.98
Arduino Uno	1	20
Dynamixel shield	1	19
Fan 12V 4000 RPM	1	1.5
Power supply 12V 20A 240W	1	9
LCD meter and shunt (20A)	1	9
Electric wire AWG24 (30m.)	2	5.50
Tendon	1	1.5
Bolt and screw	-	15
Bearing, etc.	-	35
Total		750.28

3. Experiment and Results

All experiments are intended to prove the various performances of robotic hands compared to other expensive robotic hands such as grasping objects and gestures. The development of the anthropomorphic robotic hand in this paper has five experiments: grasping experiment, gesture experiment, motor temperature experiment, structure experiment, and repeatability experiment.

One of the performance experiments of the anthropomorphic robotic hand is the grasping experiment that uses various objects in daily life to test the grasping of the robot hand. The robotic hand grasp objects that are different in shape, weight, and size by using various grasping gestures as shown in the result of the grasping experiment.

The gesture experiment is the test of the robot hand to perform basic hand gestures and symbols that were chosen from daily hand posture. These two experiments were intended to test the ability to grip objects and perform gestures as designed.

The third experiment is the operating motor temperature test to determine whether the added structure and fan can reduce the temperature of the motor while operating. The motor temperature experiment uses Arduino Uno to read the current temperature from the feedback sensor of each motor.

The fourth experiment is the structure test of the degrees of dip and pip joints of the index finger, whether the DIP joint moves along with the PIP joint by using the four-bar linkage mechanism is similar to the equation used in the design.

The last experiment is the repeatability test of the robotic hand which shows how many errors each joint has. The structure experiment and repeatability experiment use magnetic encoders (AS5600) and an Arduino Uno board to read the current degree of each joint.

All experiments test only the joints of the fingers excluding the wrist. The controllable joints of this robot hand are 11 joints (3 joints for the thumb and 2 joints for each finger), therefore the maximum magnetic encoder used to read the angle of each joint is 11 positions. The Arduino Uno board connects to each magnetic encoder board by using I2C communication. All encoders cannot connect to the Arduino Uno board directly because each encoder has the same address (0x36). The I2C multiplexer (TCA9548A) is required to expand the I2C bus port and control multiple I2C devices with the same I2C address. One multiplexer can connect to 8 devices, so 2 multiplexers are enough. The address of the multiplexer can select a value from 0x70 to 0x77 by adjusting the values of the A0, A1, and A2 pins. The robot hand with an encoder module for the test is shown in Figure 26.

The average error values of the structure experiment and repeatability experiment are calculated from the following equation.

$$Average\ Error\ Value = \frac{\sum_{i=0}^{n}(X_n - T)}{N}$$

Table 8: The Meaning of Variable

Variable	Meaning
X_n	Position value from encoder (n^{th})
T	Target position value
N	Total of test

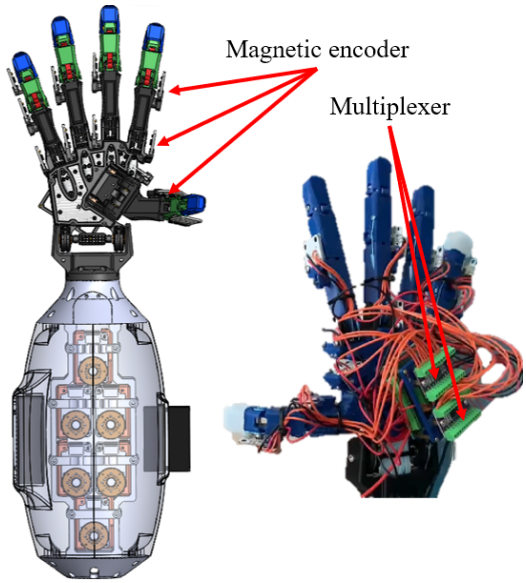


Figure 26: The Robotic Hand with Encoder Module

3.1. Result of Grasping Experiment

Generally, there are two kinds of grasping objects: power grasp and precision grasp. The precision grasp (tripod, two fingers, disk, and tip pinch) uses only fingertips with skin to hold small lightweight objects, while the power grasp (spherical, cylindrical, lateral pinch, lumbrical, large diameter, and platform) uses every part of the hand (fingertips, phalanx, palm, and skin) to grasp huge heavyweight objects. This experiment is the gripping test of the robot hand with power grasp and precision grasp by using ten grasping gestures: spherical (1.01-1.02), tripod (2.01-2.04), two fingers (3.01-3.02), cylindrical (4.01-4.05), lateral pinch (5.01-5.11), lumbrical (6.01), disk (7.01), large diameter (8.01-8.04), tip pinch (9.01), and platform (10.01) as shown in Tables 9 and Figure 27 respectively.

Because this robotic hand is controlled by humans to be used for handling various objects in daily life, The items used in the test must be found in daily life (differ in shape, weight, and size). There are 32 different objects that are used to test such as baseball, glue, pen, table tennis, bottle 600 ml, screwdriver, power bank, key, lighter, book, disk, and plaster. This test only focuses on that each item can be held in that posture without getting out of hand. The robot hand successfully grasped selected 32 different objects with 10 basic postures and can grasp objects up to 450 grams (bottle 600 ml) in cylindrical gripping gesture. The bottle made from plastic (PE) with a slippery skin and large diameter is grasped by cylindrical gesture (power grasp) to test the maximum weight that this hand can hold and to test whether robotic hands can handle things (not structural polishing). Holding a slippery object in this pose is a real gripping efficiency test of the robotic hand because the object may slip out of hand if the grasp is not tight enough. The proposed anthropomorphic design allows our robotic hand to grasp objects in a suitable gripping posture.

Table 9: Grasping Poses and Objects (a)

Grasping Pose	Objects		
	Name	Dimension(mm)	Weight(g)
	Baseball (1.01)	Ø 73	150
Spherical (1)	Tennis ball (1.02)	Ø 65	55
Tripod (2)	Glue (2.01)	Ø 20	11
	Pencil (2.02)	Ø 7.8	4
	Pen (2.03)	Ø 9.8	6
	Marker (2.04)	Ø 10	7

Table 9: Grasping Poses and Objects (b)

Grasping Pose	Objects		
	Name	Dimension(mm)	Weight(g)
Two Fingers (3)	Table tennis ball (3.01)	Ø 39.5	2
	Golf ball (3.02)	Ø 42.5	45
Cylindrical (4)	Bottle 600 ml (4.01)	Ø 60	450
	Bottle skin care (4.02)	Ø 50	72
	Huge screwdriver (4.03)	Ø 33.5	96
	Trowel (4.04)	Ø 32	27
	Power bank (4.05)	Ø 41.5	133
Lateral Pinch (5)	Key (5.01)	Thick 4.9	6
	Smart key (5.02)	Thick 0.8	4
	Metal key (5.03)	Thick 2.2	39
	Card reader (5.04)	Thick 8.5	3
	Tweezers (5.05)	Thick 10	15
	Small screwdriver (5.06)	Ø 7.1	14
	Pen (5.07)	Ø 9.8	27
	Hand drill (5.08)	Ø 8.15	41
	Lighter (5.09)	Thick 11	13
	Tape (5.10)	Thick 18.5	22
	Utility knife (5.11)	Ø 9	15
Lumbrical (6)	Book (6.01)	148.5 × 210 Thick 12	110
Disk (7)	Disk (7.01)	Ø 19 Thick 1.25	17
Large Diameter (8)	Wire strippers (8.01)	104 × 15	174
	Combination pliers (8.02)	90 × 16.5	200
	Diagonal cutter (8.03)	93 × 11.5	25
	Screwdriver box (8.04)	67.5 × 17.25	263
Tip Pinch (9)	Wound closure plaster (9.01)	Thick 1	1
Platform (10)	Document pouch (10.01)	297 × 210 Thick 8	250

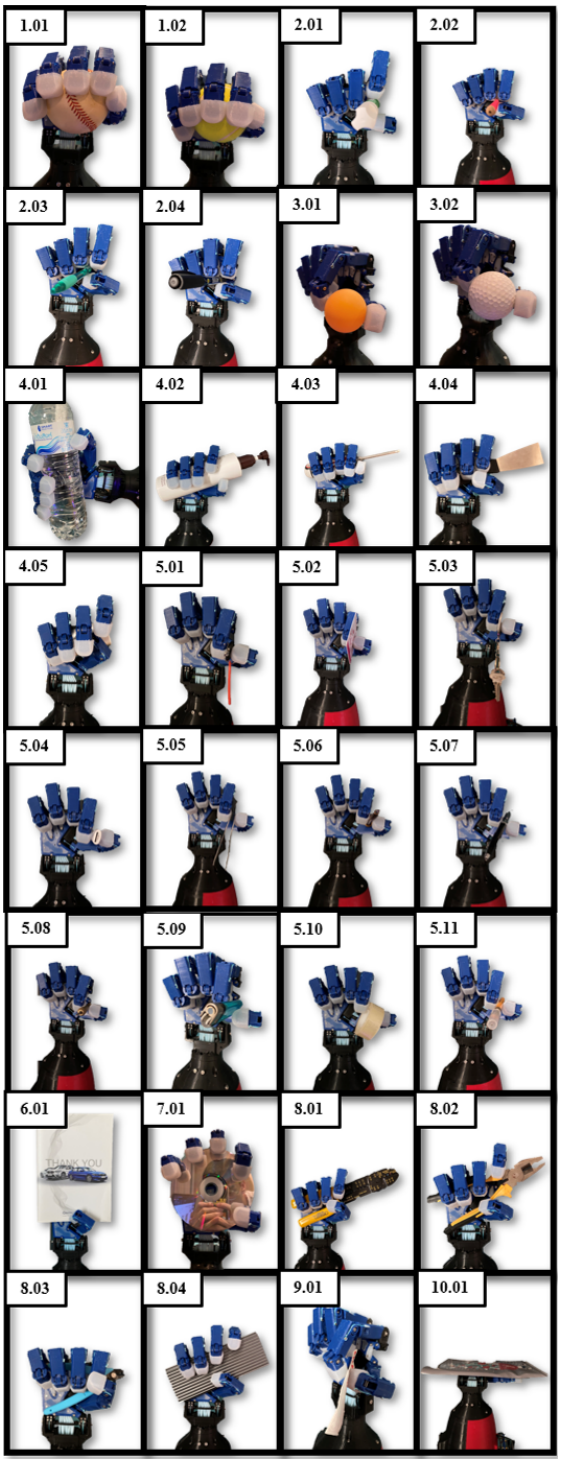


Figure 27: Robotic Hand Grasping Various Objects

3.2. Result of Gesture Experiment

This experiment is a test of the basic hand gestures and symbols that are chosen from frequently used in daily life. The robotic hand successfully posed 9 common gestures including high-five (1), peace (2), ok (3), index pointing (4), grasp (5), promise (6), love (7), check (8), and good job (9) as shown in Figure 28. The purpose of this robotic hand design does not focus on the adduction-abduction movement but to reduce the number of motors. Consequently, this hand cannot perform gestures that use the abduction-adduction movement such as fingers crossed, fig sign, and Vulcan salute.

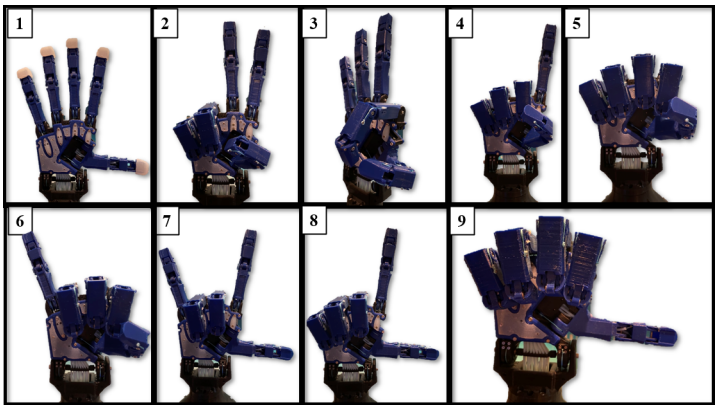


Figure 28: Gesture of Robot Hand

3.3. Operating Temperature of Motor Experiment

This experiment was to prove that an extra fan can reduce the temperature of the motor while operating by monitoring the motor temperature directly from the motor feedback sensor. The fan is installed to the cover of the forearm as shown in Figure 25. This is the performance experiment of a designed ventilated structure with a fan (12V 4000RPM). Usually, the motor can operate in the temperature range between -5 to +72 °C, but the operating motor temperature at 30 percent torque (enough for grasping objects) is between 55.0 °C to 68.0 °C (Figure 29) that close to the maximum temperature of the motor. This experiment uses Arduino Uno to control 12 motors and get feedback (Temperature) from motors in real-time (20 times). After installation and test, the range temperature of the operating motor is between 46.0 °C to 56.0 °C as shown in Figure 30. From the above, this experiment can prove that the fan can reduce the average temperature of the operating motor from 61.5 °C to 51.0 °C.

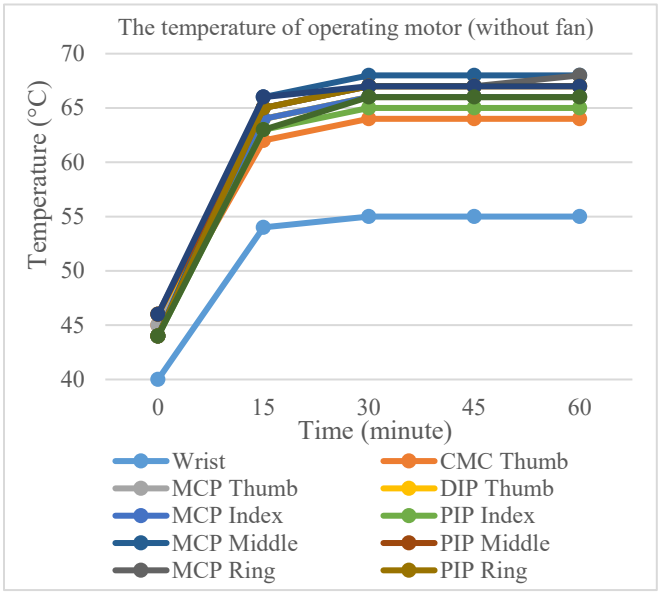


Figure 29: The Temperature of Each Operating Motor in Solid Structure

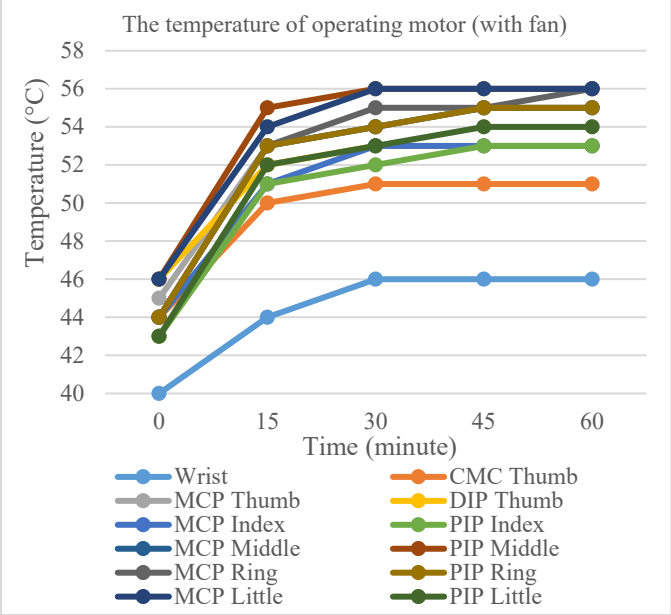


Figure 30: The Temperature of Each Operating Motor in Ventilated Structure

3.4. Structural of Four Common Fingers Experiment

The kinematic equation is used to design a four-bar linkage mechanism to move dip joints related to pip joints by using 1 actuator to control. This experiment is a structural experiment that tests the movement of dip joints related to pip joints (index finger) compared with the equation used in the design. This test uses magnetic encoder and Arduino Uno to check the degrees of each joint and control servo motors. The test will control the PIP joint and observe DIP joint of four common fingers around 100 times per position and compare with the kinematic equations from [1]. While the PIP joint is controlled by servo motor which moves from 0.0 degree to 78.0 degrees and back to 0.0 degree (100 times), the DIP joint moves from 0.0 degree to 68.1 degrees following the PIP joint by using a four-bar linkage mechanism. When the PIP joint moves from 0.0 degree to 78.0 degrees, the DIP joint will move from 0 to 69.3 degrees by calculating from the equation. After testing, the structure can move according to the equation with an error of fewer than 1.6 degrees and the PIP joint of index finger has an error of about 0.1 degrees as shown in Tables 10 and 11.

Table 10: Structural Test (Degree)

No	Encoder (statistic method)				Equation	Avg. Error (DIP)
	PIP Joint (Deg)		DIP Joint (Deg)		DIP Joint (Deg)	
	Average	SD	Average	SD		
1	78.1	0.2	68.1	0.2	69.3	1.2
2	31.1	0.2	29.1	0.1	27.5	1.6

Table 11: The Average Error of PIP Joint (Degree)

No	Encoder (statistic method)		Target	Avg Error (PIP)
	PIP Joint (Deg)		PIP Joint (Deg)	
	Average	SD		
1	78.1	0.2	78.0	0.1
2	31.1	0.2	31.0	0.1

3.5. Repeatability Experiment

The repeatability test of robotic hands using Arduino Uno boards to control 2 positions of the motor and read the position of each joint from magnetic encoder modules (12-bit resolution or about 0.08° per count). The Arduino Uno board can read the degree of each joint in real-time via a magnetic encoder. This board control motor moves forward and backward to the same position around 200 times per cycle. To conclude the data in the table, all information is expressed as the statistical method (max, min, standard deviation, average). This test is divided into two experiments which are the repeatability test of the index finger, and the repeatability test of the robot hand. From all of the experiments, this robot hand has a maximum error of repeatability of about 1.2 degrees.

First, the repeatability test of the index finger uses 3 magnetic encoders with Arduino Uno to measure the degree of MCP, PIP, and DIP joint of the index finger. This test has three sets of moving positions and each set has two positions that are selected from the range of motion of each joint. From the result, we found that the maximum error of the repeatability test of the index finger is 0.2 degrees as shown in Table 12.

Table 12: Repeatability Test of Index Finger (Degree)

Number of Sets		Statistical Method				Target	Avg Error
		Min	Max	Avg	SD		
MCP (1)	Pos 1	0.0	0.6	0.1	0.1	0.0	0.1
	Pos 2	84.0	85.1	84.1	0.2	84.0	0.1
PIP (1)	Pos 1	0.0	2.0	0.1	0.2	0.0	0.1
	Pos 2	78.0	79.8	78.2	0.4	78.0	0.2
MCP (2)	Pos 1	0.0	0.5	0.1	0.1	0.0	0.1
	Pos 2	55.0	56.1	55.1	0.2	55.0	0.1
PIP (2)	Pos 1	0.0	0.5	0.1	0.1	0.0	0.1
	Pos 2	67.0	69.0	68.2	0.3	68.0	0.2
MCP (3)	Pos 1	0.0	0.5	0.1	0.1	0.0	0.1
	Pos 2	70.0	72.0	70.2	0.4	70.0	0.2
PIP (3)	Pos 1	0.0	0.5	0.1	0.1	0.0	0.1
	Pos 2	34.8	36.2	35.1	0.2	35.0	0.1

Second, the repeatability experiment of the robot hand is testing the error of all controllable joints to find the maximum error of the repeatability test of the robot hand by using 11 magnetic encoders and Arduino Uno. This test has settings and methods the same as the repeatability test of the index finger. The maximum error of this test is 1.2 degrees at CMC joint of thumb as shown in Tables 13 and 14. The PTFE tubes with tendons are routed through the CMC joint to control the MCP and DIP joints of the thumb while the other joints have only tendons that route through. The maximum error of the repeatability test is on the CMC joint.

Table 13: Repeatability Test of Robot Hand (Degree)

Finger Names	Name of Joint		Statistical Method (Deg)			
			Min	Max	Avg.	SD
Thumb	CMC	Pos 1	0.0	2.0	0.1	0.2
		Pos 2	98.0	100.0	99.2	0.3
	MCP	Pos 1	0.0	0.5	0.1	0.1
		Pos 2	76.0	77.0	76.1	0.1
	DIP	Pos 1	0.0	2.0	0.1	0.2

		Pos 2	39.0	40.0	39.1	0.2
Index	MCP	Pos 1	0.0	0.5	0.1	0.1
		Pos 2	70.0	71.3	70.3	0.5
	PIP	Pos 1	0.0	0.6	0.1	0.1
		Pos 2	78.0	79.5	78.5	0.4
Middle	MCP	Pos 1	0.0	0.5	0.1	0.1
		Pos 2	98.0	99.2	98.4	0.2
	PIP	Pos 1	0.0	0.5	0.1	0.1
		Pos 2	89.0	90.0	89.4	0.2
Ring	MCP	Pos 1	0.0	1.2	0.1	0.2
		Pos 2	98.0	98.5	98.2	0.1
	PIP	Pos 1	0.0	0.5	0.1	0.1
		Pos 2	89.0	90.6	89.2	0.3
Little	MCP	Pos 1	0.0	0.5	0.1	0.1
		Pos 2	80.0	81.9	80.4	0.6
	PIP	Pos 1	0.0	0.5	0.1	0.1
		Pos 2	86.0	87.2	86.2	0.3

Table 14: Target and Average Error of the Repeatability Test of Robot Hand (Degree)

Finger Names	**Name of Joint**		**Target (Deg)**	**Avg (Deg)**	**Avg Error (Deg)**
Thumb	CMC	Pos 1	0.0	0.1	0.1
		Pos 2	98.0	99.2	1.2
	MCP	Pos 1	0.0	0.1	0.1
		Pos 2	76.0	76.1	0.1
	DIP	Pos 1	0.0	0.1	0.1
		Pos 2	39.0	39.1	0.1
Index	MCP	Pos 1	0.0	0.1	0.1
		Pos 2	70.0	70.3	0.3
	PIP	Pos 1	0.0	0.1	0.1
		Pos 2	78.0	78.5	0.5
Middle	MCP	Pos 1	0.0	0.1	0.1
		Pos 2	98.0	98.4	0.4
	PIP	Pos 1	0.0	0.1	0.1
		Pos 2	89.0	89.4	0.4
Ring	MCP	Pos 1	0.0	0.1	0.1
		Pos 2	98.0	98.2	0.2
	PIP	Pos 1	0.0	0.1	0.1
		Pos 2	89.0	89.2	0.2
Little	MCP	Pos 1	0.0	0.1	0.1
		Pos 2	80.0	80.4	0.4
	PIP	Pos 1	0.0	0.1	0.1
		Pos 2	86.0	86.2	0.2

4. Conclusions

From the experiment, this anthropomorphic robot hand can grasp selected 32 different objects commonly found in daily life with 10 basic gripping postures and can perform 9 basic gestures. The other gestures that this hand cannot perform use the abduction-adduction movement such as fingers crossed, fig sign, and Vulcan salute. This robot hand can increase grasping posture and hand gesture by adding the abduction-adduction motion with the smallest actuators into the MCP joint of each finger. In addition, the robot hand can grasp an object up to 450 grams. When grasping huge objects, we usually use the cylindrical grasp (power grasp posture) as in the grasping experiment section. From the above results, it can be found that the motor can sufficiently transmit force and torque to the fingers and fingertips in order to grasp the 450 grams object. By using a Lateral grasp, the anthropomorphic hand can pick up small objects such as keys and utility knives with fingertips. In addition, the proposed robot hand has sufficient force and rigidity to grasp various objects while the cost is lower than other designs. The equations used in the design proven that the structure can move according to the equation with an error value of about 1.6 degrees. In the repeatability experiment, this robot's hand has a maximum error of repeatability of about 1.2 degrees.

We have designed and prototyped an open-source anthropomorphic robotic hand for teleoperated robots with a detailed design process for further developers. We use 3D printing and common components for assembling. The four-bar linkage mechanism is used to mimic the relative motion between DIP and PIP joints same as the human finger, while also reducing the number of motors. We experimentally that our proposed robotic hand design has good repeatability in finger motions and grasping daily objects. This paper explains how to design a robot hand, it can be adjusted to any desired size by using the equation given above.

Design of an Open-Source Anthropomorphic Robotic Hand for Telepresence Robot is available for study and development, which can be found at the following site. https://github.com/Jittaboon-tri/Anthropomorphic-Robotic-Hand

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We want to give acknowledgements to AI for all projects for the financial support in our research, Institute of Field Robotics (FIBO), and King Mongkut's University of Technology Thonburi (KMUTT) and Fundamental Fund (Basic Research Fund) for supporting fund.

References

[1] J. Trichada, T. Wimonrut, N. Tirasuntarakul, T. Choopojcharoen, B. Sakulkueakulsuk, "Design of an open source anthropomorphic robotic finger for telepresence robot," ACM International Conference Proceeding Series, 62–66, 2021, doi:10.1145/3467691.3467704.

[2] S.C. Jee, M.H. Yun, "An anthropometric survey of korean hand and hand shape types," International Journal of Industrial Ergonomics, 53, 10–18, 2016, doi:10.1016/j.ergon.2015.10.004.

[3] V. Doroshenko, O. Mul, O. Kravchenko, "Mathematical relations for harmonization with technical and decorative casting nature," Boundary Field Problems and Computer Simulation, 55(December), 44–49, 2016, doi:10.7250/bfpcs.2016.007.

[4] P.G. Narasimha-shenoi, Golden ratio in human anatomy, Thesis, Government College Chittur, 2014, doi:10.13140/2.1.2265.9526.

[5] D. Persaud, J.P. O, "Fibonacci series, golden proportions, and the human biology," Austin Journal of Surgery, 2(5), 1–6, 2015, ISSN : 2381-9030.

[6] S.A. Powell, A review of anthropomorphic robotic hand technology and data glove based control, Masters Thesis, Virginia Polytechnic Institute and State University 2016.

[7] M. Controzzi, C. Cipriani, B. Jehenne, M. Donati, M.C. Carrozza, "Bio-inspired mechanical design of a tendon-driven dexterous prosthetic hand," 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10, (i), 499–502, 2010, doi:10.1109/IEMBS.2010.5627148.

[8] D. Choi, D.-W. Lee, W. Shon, H.-G. Lee, "Design of 5 D.O.F robot hand

with an artificial skin for an android robot," The Future of Humanoid Robots - Research and Applications, (January), 2012, doi:10.5772/26282.
[9] W.S. You, Y.H. Lee, H.S. Oh, G. Kang, H.R. Choi, "Design of a 3D-printable, robust anthropomorphic robot hand including intermetacarpal joints," Intelligent Service Robotics, 12(1), 1–16, 2019, doi:10.1007/s11370-018-0267-8.
[10] Material property of RA-22AB, https://www.resinrungart.com/ra_22ab.html, Jan. 2022.
[11] The coefficient of the PTFE tube,https://www.fluorotherm.com/technical-information/ materials-overview/ptfe- properties/, Jan. 2022.
[12] Z. Su, K. Inaba, A. Karmakar, A. Das, "Characterization of mechanical property of pla-abs functionally graded material fabricated by fused deposition modeling," Proceedings of ASME 2021 Gas Turbine India Conference, GTINDIA 2021, (December), 2021, doi:10.1115/GTINDIA2021-76025.
[13] Material price of polymaker, https://us.polymaker.com, Jan. 2022.
[14] Material price of eSUN, https://www.esun3d.com, Jan. 2022.
[15] Specification of Dynamixel servo motor,https://www.robotis.us/dynamixel-xl430-w250-t/, Jan. 2022.
[16] Speciafication of tendon,https://www.lazada.co.th/products/1-2-proberos-x4-100m-bluegreenyellowredgrey-pe-4-5-100-thailand-fishing-mall-fishing-line-i2161246830-s7196994582.html, Jan 2022.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 4, 18-27 (2023)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Localization of Impulsive Sound Source in Shallow Waters using a Selective Modal Analysis Algorithm

Faraz Talebpour, Saeed Mozaffari, Mehrdad Saif, Shahpour Alirezaee*

Department of Electrical and Computer Engineering University of Windsor, Windsor, Ontario, Canada

ARTICLE INFO

Article history:
Received: 19 December, 2022
Accepted: 22 June, 2023
Online: 21 July, 2023

Keywords:
Passive Underwater Localization
Underwater Signal Processing
Passive Environmental Monitoring
Modal Analysis

ABSTRACT

Passive remote monitoring applications of underwater signal processing in a shallow water environment are an impactful area of research for environmental and marine-life monitoring. The majority of the sound source localization techniques require carefully placed synchronized hydrophone arrays, which can be complicated and hard to maintain. In this paper, we utilized the modal dispersions of a signal to derive a localization method for a noisy, shallow water environment. Our proposed algorithm employs modal selection to process the most noise-resistive dispersion curves, improving the accuracy and noise-resistivity of the existing methods. Moreover, we proposed a 2D localization method with multiple unsynchronized hydrophones and minimal hardware requirements and limitations. Furthermore, we analyzed the effects of underwater ambient noise on the accuracy of the proposed method, using simulated and real recorded explosion and whale sounds, and compared our algorithm's localization performance with others. Simulation results show increased localization accuracy of 30m for the recorded explosion sound and 360m for the Whale sound.

1 Introduction

This paper extends our previous work presented in CCECE 2022 [1] by introducing a selective-modal algorithm architecture for localizing impulsive sound sources in shallow waters. Our proposed algorithm improves performance in lower signal-to-noise ratio (SNR) scenarios by selecting the best modal pairs. In this paper, we provide a more detailed explanation of the localization formulas, propose a 2D unsynchronized localization scheme, analyze the performance of our algorithms using real recorded signals, and compare them with existing works. This paper extends our previous work presented in CCECE 2022 [1] by introducing a selective-modal algorithm architecture for localizing impulsive sound sources in shallow waters. Our proposed algorithm improves performance in lower signal-to-noise ratio (SNR) scenarios by selecting the best modal pairs. In this paper, we provide a more detailed explanation of the localization formulas, propose a 2D unsynchronized localization scheme, analyze the performance of our algorithms using real recorded signals, and compare them with existing works.

The field of underwater acoustics encompasses the primary modality of underwater sensing and communication, which is sound. Early research in underwater signal processing focused on mathematical models and the behavior of acoustic sounds in the underwater environment [2]. Over time, advancements in adaptive signal processing and sensor technology led to practical applications in underwater signal processing. Sonar systems, particularly underwater sonars, have undergone rapid developments in the past two decades, driven by increased processing capability and the implementation of more computationally intensive techniques. The underwater environment presents unique challenges, including increased human-made noise due to the growing number of vessels in the ocean. Marine mammals heavily rely on vocalization for communication and locating other mammals, making them sensitive to sounds generated by human activities such as geophysical explorations, offshore extraction, shipping, and active sonar applications[3]. As a result, researchers have been motivated to develop remote monitoring techniques to study marine mammal behavior and monitor environmental changes. Underwater localization techniques can be broadly classified into passive and active categories. Passive sonar processes received signals without signal transmission, while active sonar involves both signal transmission and reception [4]–[5]. Researchers have proposed various passive underwater localization methods, including time-frequency difference of arrival (TDOA), received signal strength (RSS), and modal-based analysis [6]–[7]. The

*Corresponding Author: Shahpour Alirezaee, CEI, University of Windsor, Windsor,ON, CA, 519-253-3000 ext. 7472 & alirezae@uwindsor.ca

underwater medium is a dynamic multi-path channel where sound-waves travel through multiple paths with different speeds [8, 9]. TDOA algorithms utilize time differences between received signals, while RSS algorithms focus on received signal power. However, implementing TDOA-based techniques often requires synchronized hydrophone arrays and prior information, resulting in increased costs, complexity, and high error levels in low SNR environments. In [10], the authors conducted experiments under real test conditions with sensor nodes and observed that the sensors constantly move due to varying water surface conditions, resulting in unsynchronized sensor nodes. To address this issue, the authors in [11] proposed a self-calibration technique utilizing a shift-keying pulse and composite transducers. Similarly, in [12], it was demonstrated that the use of maximum likelihood estimators (MLE) in TDOA methods led to non-linearity problems. In response, the authors in [13] formulated TDOA target motion analysis as a least-square optimization problem, solving it in polynomial time. Furthermore, [14] investigated the performance of TDOA techniques under different noise levels and highlighted the significant impact of white noise on the accuracy of TDOA algorithms.

To improve the accuracy and noise resistivity, [7] introduced a hybrid localization technique based on the direction of arrival (DOA) and received signal strength (RSS) using a vector and an isotropic acoustic hydrophone. Phased array-based localization was proposed in [15] to enhance noise resistivity. However, TDOA-based methods, while accurate, often require arrays of synchronized hydrophones and prior information, resulting in higher implementation costs, increased complexity, and reduced accuracy in low SNR environments. In [16], the authors suggested the utilization of the Kronecker product operation to extract the two-dimensional power distribution matrix from the beam power function, reducing the number of required hydrophones and improving noise resistivity.

Despite extensive efforts in the field, achieving sensor node synchronization and fulfilling the multi-hydrophone requirements of TDOA-based techniques can still pose significant challenges and incur high costs. To overcome these limitations, modal analysis-based localization was introduced as a solution, eliminating the need for source prior information, multiple hydrophones, and hydrophone synchronization [16, 17]. In the underwater environment, acoustic waves consist of multiple modes that travel through water with varying velocities. As a result of these differing velocities, the modes disperse during propagation through the water channel [6, 18]. In [19] , the authors proposed a modal analysis-based approach specifically designed for localizing mammal sounds. Furthermore, in [11], accuracy was enhanced by expanding the localization frequency range and considering additional modes during the localization process. Additionally, [20] proposed a nonlinear-based warping technique for modal filtering.

In this paper, we build upon our previous work published in [1] and introduce novel advancements to the field of underwater localization. Specifically, we extend our research by incorporating the utilization of multiple hydrophones for two-dimensional localization. Unlike previous approaches, our proposed techniques are independent and standalone, enabling each hydrophone to perform separate target localization in an unsynchronized manner.

To lay the groundwork for our methodology, we begin by introducing a shallow underwater channel model based on the theory of normal modes in Section 2. Additionally, we present a comprehensive model for the channel's ambient noise and derive the modal functions necessary for modal analysis.

In Section 3, we take a significant step forward by deriving a selective noise-resistive modal-based localization method that exhibits improved resistance to noise. This novel approach addresses a crucial challenge in underwater localization and enhances the accuracy of our algorithm.

To evaluate the performance of our proposed method, we present the obtained results in Section 4 and highlight the significance of modal selection for achieving superior performance. Furthermore, we thoroughly investigate the impact of noise on the accuracy of our algorithm within the 30dB < SNR < 45dB range, providing insightful comparisons with existing approaches.

In addition, we conduct an in-depth analysis and comparison of the accuracy and noise resistivity of our proposed method with other techniques using real recorded explosions and north Atlantic sounds. By doing so, we establish a comprehensive understanding of the strengths and limitations of our approach in realistic scenarios.

Finally, we evaluate the performance of our proposed 2D Localization and tracking method by comparing it with state-of-the-art techniques, demonstrating the advancements we have made in the field of underwater localization.

2 Normal Mode Propagation

Normal mode theory is suitable for modeling shallow underwater environments with respect to normal-Modes propagation. While modal-based channel models are not the most accurate model currently available, they can accurately model shallow underwater environments for passive sound source localization and monitoring applications.

2.1 Underwater Acoustic Propagation

Let us consider the model description presented in Figure 1 where an acoustic sound source is located at (x_s, y_s, z_s) that produces a continuous-time signal. After propagation, the signal is picked up by a hydrophone placed on a buoy at (x_h, y_h, z_h). For ease of use, we have considered the hydrophone on the right end of Figure 1 as the point of origin in the Cartesian and cylindrical coordinates. The displacement caused by the propagating source is time-harmonic, governed by Helmholtz law, and is given as [6, 21, 17, 22, 23, 24]

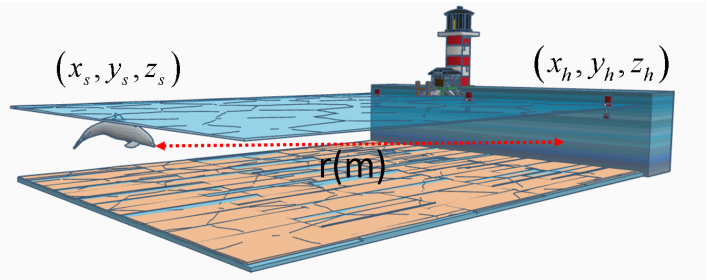


Figure 1: Model description

$$\left[\nabla^2 + K^2(\vec{r})\right] p(\vec{r}) = -4\pi f(\vec{r}) \tag{1}$$

$K(\vec{r})$ is the medium wave number at radial frequency ω,∇ gradient operator, and $p(\vec{r})$ is the pressure. We can further simplify this equation to form the Helmholtz equation in two dimensions, as the sound speed and density depends only on depth z

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial p}{\partial r}\right) + \rho(z)\frac{\partial}{\partial z}\left(\frac{1}{\rho(z)}\frac{\partial p}{\partial z}\right) + \frac{\omega^2}{c^2}p = \frac{\delta(r)\,\delta(z-z_s)}{-2\pi r} \tag{2}$$

where r is the distance to the source, ρ is the medium density, c is the propagation speed, δ is the Dirac delta, and ω is angular velocity [6]. Using the separation of variables, we look for a depth-related pressure solution in the form of $p(r,z) = \phi(r)\psi(z)$, which will result in

$$\frac{1}{\phi}\left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \phi}{\partial r}\right)\right] + \frac{1}{\psi}\left[\rho(z)\frac{\partial}{\partial z}\left(\frac{1}{\rho(z)}\frac{\partial \psi}{\partial z}\right) + \frac{\omega^2}{c^2}\psi\right] = 0 \tag{3}$$

where ϕ is the volume displacement, and Ψ is the general modal depth function. Terms in square brackets of the equation (3) are functions of $r$and $z$respectively. To satisfy the equation (3), each term should be equal to a constant [24]. Considering the Pekeris waveguide -where water is considered equal columns with varying speeds of propagation- We can drive the modal equation by considering the K_{rm}^2 as a separation constant [25].

$$\begin{aligned} &\rho(z)\tfrac{\partial}{\partial z}\left(\tfrac{1}{\rho(z)}\tfrac{\partial \psi_m(z)}{\partial z}\right) + \left[\tfrac{\omega^2}{c^2} - K_{rm}^2\right]\psi_m(z) = 0 \\ &\psi(0) = 0 \quad , \quad \tfrac{d\psi}{dz}\Big|_{z=D} = 0 \end{aligned} \tag{4}$$

where $\psi_m(z)$ is the particular modal function $\psi(z)$ obtained with horizontal wave-number K_{rm} as separation constant. The boundary condition of the equation (4) considers each water column a pressure release surface ($z = 0$) and a perfectly rigid seabed at $z = D$ ($D < 100$) which translates to no changes in the volume at surface and seabed resulting in $d\psi/dz = 0$ [25].
Equation (4) is a classical Sturm-Liouville eigenvalue problem [24]. Applying the orthogonality of the modal Sturm-Liouville problem, we can write

$$\int_0^D \frac{\psi_m(z)\,\psi_n(z)}{\rho(z)}dz = 0 \qquad m \neq n \tag{5}$$

Equation (3), the solutions of modal equations are arbitrary to multicaptive constants; therefore, we can further simplify the results using equation (5) as

$$\int_0^D \frac{\psi_m^2(z)}{\rho(z)}dz = 1 \tag{6}$$

Moreover, modes transmit as a complete set, resulting in an arbitrary function as a sum of all normal modes, which will yield the pressure function as:

$$p(r,z) = \sum_{m=1}^{\infty} \phi_m(r)\,\psi_m(z) \tag{7}$$

Substituting equation (7) in equation (2) provides :

$$\begin{aligned} &\textstyle\sum_{m=1}^{\infty}\left\{\frac{1}{r}\frac{d}{dr}\left(r\frac{d\phi_m(r)}{dr}\right)\psi_m(z) + \phi_m(r)\left[\rho(z)\frac{d}{dz}\left(\frac{1}{\rho(z)}\frac{d\psi_m(z)}{dz}\right) + \frac{\omega^2}{c^2}\psi_m(z)\right]\right\} \\ &= -\frac{\delta(r)\delta(z-z_s)}{2\pi r} \end{aligned} \tag{8}$$

After applying the operator equation (9)

$$\int_0^D (\cdot)\frac{\psi_n(z)}{\rho(z)}dz \tag{9}$$

Furthermore, considering the orthogonality property stated in the equation (5), only n terms of the sum remain.

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{d\phi_n(r)}{dr}\right) + K_{rn}^2\phi_n(r) = -\frac{\delta(r)\,\psi_n(z_s)}{2\pi r\rho(z_s)} \tag{10}$$

the solution to the equation (10) is provided in terms of the Hankel function as:

$$\phi_n(r) = \frac{1}{4\rho(z_s)}\left(\psi_n(z_s)\,H_0^{(1,2)}(K_{rn}r)\right) \tag{11}$$

The signal's energy radiates outwards, and therefore the solution will be $H_0^{(1)}$. Considering the radiation conditions, after substituting (11) in (7), we can derive the pressure equation based on the modal function as

$$p(r,z) = \frac{1}{4\rho(z_s)}\sum_{m=1}^{\infty}\psi_m(z_s)\,\psi_m(z)\,H_0^{(1)}(K_{rm}r) \tag{12}$$

we can further simply equation (12) by using the asymptotic approximation to the Henkal's function, yielding:

$$p(r,z) \simeq \frac{i}{\rho(z_s)\sqrt{8\pi r}}e^{-\frac{i\pi}{4}}\sum_{m=1}^{\infty}\psi_m(z_s)\,\psi_m(z)\frac{e^{iK_{rm}r}}{\sqrt{K_{rm}}} \tag{13}$$

provides us with the pressure function, based solely on modal functions and depth.

2.2 Solution to Wave Equation

We must simplify the displacement equations further to perform channel modeling in simulation software. The non-homogeneous differential equation (1) can be solved using the Green's function method and expanded as the displacement equation as [6, 23]

$$\begin{aligned} &\rho(z)\frac{d}{dz}\left[\frac{1}{\rho(z)}\frac{dg(z)}{dz}\right] + \left[\frac{\omega^2}{c^2} - K_r^2(z)\right]g(z) = \frac{\delta(z-z_s)}{-2\pi} \\ &\delta(z-z_s) = \sum_m a_m\psi(z_s) \\ &\delta(z-z_s) = \sum_m \frac{\psi_m(z_s)\,\psi_m(z)}{\rho(z_s)}, a_m = \frac{\psi_m(z_s)}{\rho(z_s)} \end{aligned} \tag{14}$$

substituting $g(z) = \sum_m a_m\psi_m(z)$ In equation (14) provides the depth related modal function as:

$$\begin{aligned} &\sum_{m=1}^{\infty} b_m\left[\left(K_{rm}^2 - K_r^2\right)\psi_m(z)\right] = \\ &\frac{1}{-2\pi}\sum_m \frac{\psi_m(z_s)\,\psi_m(z)}{\rho(z_s)}, K_z^2 = K_{rm}^2 - K_r^2 \end{aligned} \tag{15}$$

where K_{rm}, K_z, K_r are the angular, vertical and horizontal wavenumbers [26, 6]. applying Green's solution to equation (15) would provide the general modal function:

$$g(z) = -\frac{1}{2\pi\rho(z_s)} \sum_m \frac{\psi_m(z_s)\psi_m(z)}{K_r^2 - K_{rm}^2} \tag{16}$$

with general solutions and eigenfunctions as follows

$$\begin{aligned} &\psi_m(z) = A\sin(K_z z) + B\cos(K_z z) \\ &K_z = \sqrt{(K_{rm})^2 - K_r^2} \qquad K_{rm} = \frac{\omega}{c} \\ &\psi_m(z) = \sqrt{2(\rho/D)}\sin(K_z z) \\ &v_m(\omega) = \frac{\omega}{K_{rm}} \\ &\frac{\omega}{c_{Seabed}} < \text{Number of Modes} < \frac{\omega}{c_{Water}} \end{aligned} \tag{17}$$

where $v_m(\omega)$ is the velocity of mode m at angular frequency ω.

2.3 Underwater Ambient Noise

Noise in a shallow underwater environment can be categorized into two main types, ambient noise caused by the channel characteristics and artificial noises created by external sources such as ships and marine life. Many studies consider the noise a simple added white noise; however, underwater ambient noise can be more accurately modeled as colored noise. The underwater channel's behavior is best described as a low-pass filter. It can be modeled as a white noise sequence filtered using a Butterworth IIR low-pass filter with 30dB attenuation in stopband and normalized stopband frequency of 0.05 $half cyle/sample$ per sample ad 0.9 $half cyle/sample$ Respectively [27]. Figure 2 presents the signal and noise in the time domain with SNR=45dB.

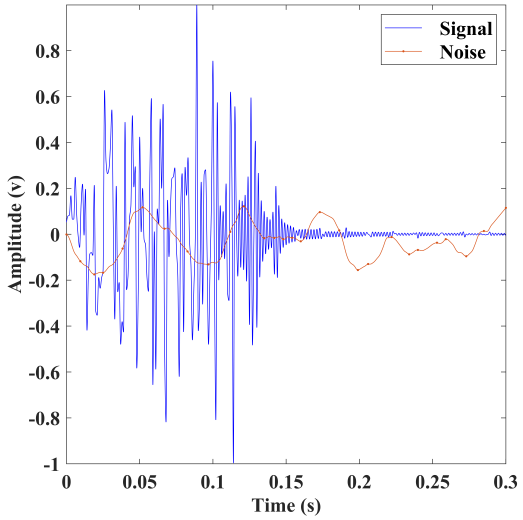


Figure 2: Signal (blue), Noise (red)

3 Modal Analysis based localization

In the previous section, we introduced the channel model and modal functions. In this section, we will derive the necessary equations for the localization of impulsive sound sources using modal functions. The modal-based localization methods are based on the dispersion of the natural frequencies as they propagate underwater.

3.1 Modal Dispersion

As stated earlier, modes travel at different speeds (equation (17)), resulting in dispersion at the receiver. Let us consider the simulation scenario of Figure.1, where the Normal mode theory with ambient noise is used to model the channel. Considering an impulsive sound source at a depth of D_s=20m, 4000 meters away from the hydrophone, ($\rho_{(Seabed)}$=1000(Kg/m^3), $\rho_{(Water)}$=1000(Kg/m^3), $c_{(Seabed)}$=1500(m/s), $c_{(Water)}$=1600(m/s)), the propagated signal will have the time-frequency (TF) representation provided in Figure.3, which illustrates the dispersion caused by the difference in propagation speeds. One can employ the dispersion of modes to localize the sound source through modal analysis after filtering them.

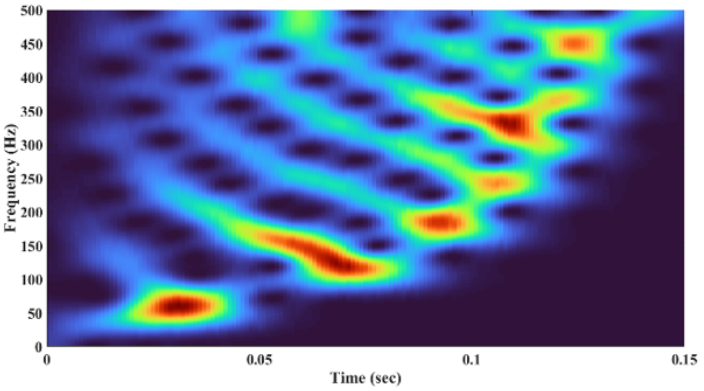


Figure 3: TF analysis, Modal dispersion , $f_{(Max)}$=600Hz

As the TF analysis graph illustrates, the dispersion curve's frequencies overlap between the modes and render conventional filtering techniques inert. The overlapped frequencies are the product of the nonlinear phase characteristics in the equation (16).To address this issue, considering the pressure signal in the time domain as

$$P(t) = \sum_m \psi_m(t)\, e^{2j\pi\nu_c(m)\zeta(t)} \tag{18}$$

Where $\zeta(t)$is the dispersity function $\zeta(t)$ in the time domain is given as

$$\zeta(t) = \sqrt{t^2 - t_r^2} = \sqrt{t^2 - (r/v_g)^2} \tag{19}$$

Using $\zeta(t)$, we can warp the signal by linearizing the phase using a warping function [17]:

$$\begin{cases} \zeta = \sqrt{t^2 - (r/v_g)^2} \\ \zeta\zeta^{-1} = 1 \end{cases} \rightarrow \zeta^{-1}(t) = \sqrt{t^2 + \left(r/v_g\right)^2} \tag{20}$$

Applying the warping function ζ^{-1} linearizes the phase. The TF graph of the linearized signal is presented in Figure 4.

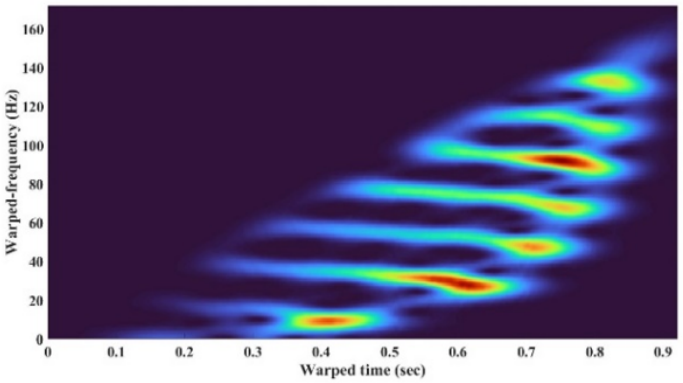


Figure 4: TF graph of warped signal

3.2 Localization Algorithm

Since the modal dispersion is directly related to the speed of propagation and distance, we can develop localization algorithms based on the TDOA concept. After filtering each of the dispersion curves; in an accurate channel model, the following expression will be true

$$\underset{Estimated}{\tau_n}(r,c) - \underset{Measured}{\tau_n} \approx 0 \qquad \forall n \tag{21}$$

Where τ_n is the dispersion curve. Measured τ_n can be obtained by warping and filtering each of the modals and by substituting the relationship between velocity and distance in the equation (17), the estimated τ_n can be obtained as:

$$\underset{Estimated}{\tau_n}(r,c) = \frac{r}{v_g(f,n)} \tag{22}$$

Where τ_n is the estimated dispersion curve for mode n transmitted over the range R with seabed sound speed c and group velocity $\upsilon_g(f,n)$. To localize the signal, we are looking for a range r that minimizes the statement (21). In other words

$$[\hat{r}] = \arg\min_{[\hat{r}]}((\underset{Estimated}{\tau_m}(r,c_{seabed}) - \underset{Estimated}{\tau_n}(r,c_{seabed})) - \ldots (\underset{Measured}{\tau_m} - \underset{Measured}{\tau_n})) \tag{23}$$

Where m and n can be any of the modes, summing over all frequency bins will yield

$$\sum_n \sum_m \sum_f \left[\left(\underset{Estimated}{\Delta\tau_{n,m}}(r,c)\right) - \left(\underset{Measured}{\Delta\tau_{n,m}}\right)\right] \approx 0 \qquad \forall n,m \tag{24}$$

Equation (24) results in a $m \times n$ matrix of dispersion curve differences and are used to derive the following cost function

$$\eta(r,c,n,f) = \sum_r \sum_n \sum_m \sum_f \left[\left(\left(\underset{Estimated}{\Delta\tau_{n,m}}(r)\right) - \left(\underset{Measured}{\Delta\tau_{n,m}}\right)\right)\right]^2 \tag{25}$$

We employed a grid search algorithm to minimize the cost function η for values of r.

Algorithm 1 presents our proposed method where μ_r defines the localization step size in the search boundary $[r_{min}, r_{max}]$ and ε is the accuracy of the estimated range. The Localization is performed in two steps; first, seabed and water parameters are defined based on the environment, and search boundaries for range and propagation speeds in seabed and sea are set. Next, the tensor of order 3, as shown in Figure 5, is formed to find the pairs of dispersion curves with the best performance (lowest value). Then, the cost function is formed only for the selected pairs of modes. Using a grid search algorithm, the location of the source can be estimated.

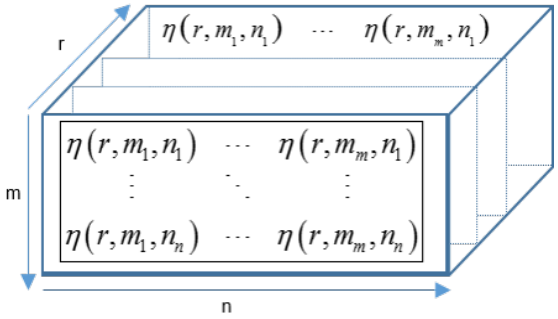


Figure 5: Cost function $[\ ]_{r\times m\times n}$

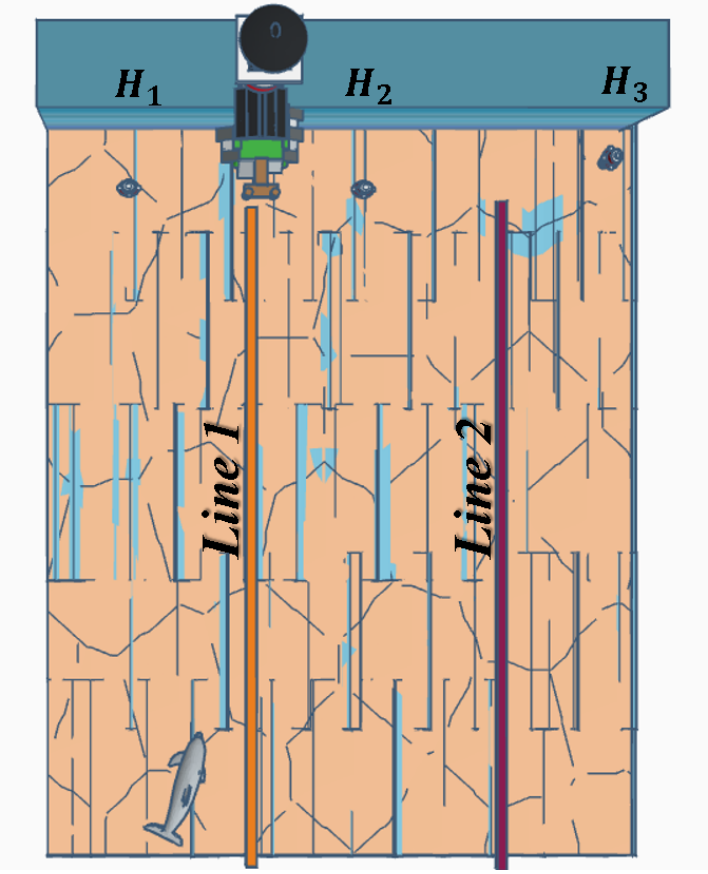


Figure 6: Model Description, multi hydrophone (H1,H2,H3)

Algorithm 1: Proposed localization Algorithm

Result: r
Initialization $\hat{r} = (r_{\min} : r_{\min} + \mu_r : r_{\max})$, $\rho_{water}, \rho_{seabed}$;
Warp Input signal;
Extract τ_n using TF ;
while $(r_{\min} < \hat{r} < r_{\max})$ **do**
  Form MPC;
  Select best modal pair;
  Minimize cost function;
  if $\Delta r, \leq \varepsilon$ **then**
    return:$[\hat{r}]$;
  else
    Change μ_r;
  end
end

3.3 2D Localization

While most modal-based localization methods proposed by literature perform ranging, we propose a method for unsynchronized 2D localization with minimal hardware requirements. In the case of 2D-localization requirements, buoys (each with a single hydrophone) can transmit the received signals to a base station on shore or a vessel to be analyzed in a central processor. Although utilizing multiple hydrophones would require sensor synchronization in other methods, the proposed modal-based localization analyzes modes picked up by each hydrophone separately. Moreover, given the high-range localization capabilities, buoys can be placed far apart, reducing implementation costs. Figure 6 illustrates the model description for 2D localization, where lines $Line1$ and $Line2$ are assumed at coordinates $[(x_{H2} - x_{H1})/2], [(x_{H2} - x_{H3})/2]$. Given the distances of each buoy, each hydrophone's average power of received modes is different. Hydrophones with the highest levels of received signal power are closest to the target. In the model description presented in Figure.6 ; $P(B_{H3}) < P(B_{H2}) < P(B_{H1})$ places the estimated latitude of the source $Line1_{H1H2} > x_s$. Based on the estimated location of the source and three calculated ranges from each buoy, we can perform 2D triangulation and track an object without needing a synchronized sensor array.

4 Results and Discussions

This section includes numerical experiments to illustrate the proposed localization method and discusses the effects of ambient noise on the accuracy of the proposed algorithm. Modal analysis is suitable for processing underwater signals in long distances ($r > 1000m$) based on only one hydrophone without synchronization.

4.1 Simulated Sound

We consider an impulsive sound source is placed at Cartesian coordinates (4000,45,0) with a maximum frequency of 500 Hz. An Omni-directional hydrophone is located at (0,15,0). We assume the speed of propagation in the seabed $c_b = 1600m/s$, speed of propagation in water c_w=1500m/s, density in water ρ_w=1000kg/m^3, and density in the seabed ρ_b=1500kg/m^3. The performance is evaluated based on the cost function's mean square error (MSE) and the estimated range's Root Mean Square Error (RMSE). Moreover, the result of this study is compared with those of [17], which has used the same approach in localization.

Figure 7 (a),(b), and (c) illustrates the RMSE of the cost function for SNR=45dB,35dB, and 30dB values for each modal pair. As we can see, considering the low-pass filter nature of the ambient noise, the noise than others would more influence pairs of first and last modes. This is mainly due to both filter boundaries' relatively low stop-band attenuation. This effect can compromise localization accuracy in low SNR environments. To address this issue, we proposed employing the cost-function MSE matrix of Figure 7, using the equation (25) to identify the best and most noise-resistant pairs of modes (lowest values), resulting in the lowest MSE. After identifying the best modal pairs (2 pairs in this study), we can find the estimated location of the acoustic sound source through the equation (23).

Figures 7 (a), (b), and (c) depict the Root Mean Square Error (RMSE) of the cost function for SNR values of 45dB, 35dB, and 30dB, respectively, for each pair of modes. It can be observed that, due to the low-pass filter characteristics of ambient noise, certain modal pairs are more influenced by noise compared to others. This effect is particularly prominent in the first and last mode pairs, primarily because of the relatively low stop-band attenuation at the boundaries of the filter. In low SNR environments, this influence can significantly compromise localization accuracy. Furthermore, Figure 7 demonstrates that the choice of modal pairs significantly impacts the error levels, as different pairs yield varying levels of error. The study presented in [17] solely employs modal pairs with sequential wavenumbers numbers, disregarding the performance of different pairs. To address this issue, we propose utilizing the MSE matrix of Figure 33 as the cost function, employing equation (25) to identify the most noise-resistant and optimal pairs of modes (with the lowest values). This selection process leads to lower MSE and enables us to determine the estimated location of the acoustic sound source using equation (23).
Figure 8a showcases the Root Mean Square Error (RMSE) of our proposed cost function for range estimation at different SNR levels, and it compares these results with the localization outcomes presented in [17]. Figures 8a and 8b clearly demonstrate that our proposed method exhibits superior performance in both low and high SNR environments. This improvement can be attributed to the fact that the localization method employed in [17] does not incorporate mode pair evaluation or selection. Instead, they utilize pairs of modes with consecutive mode numbers in their localization algorithm. However, as indicated in Figure 7, sequential mode numbers do not necessarily yield better localization results. By performing mode evaluation and selection, as shown in Figures 8a and 8b, the localization algorithm becomes more resilient to high levels of noise and achieves greater accuracy.

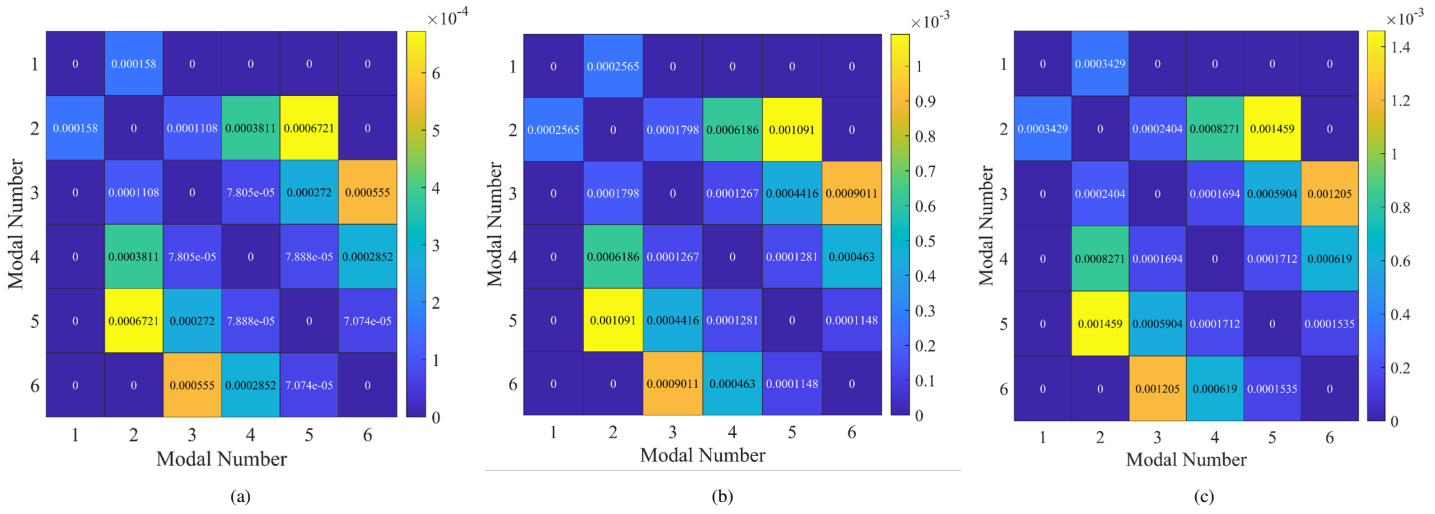


Figure 7: Cost function MSE for (a)SNR:45dB, (b)35dB, (c)30dB

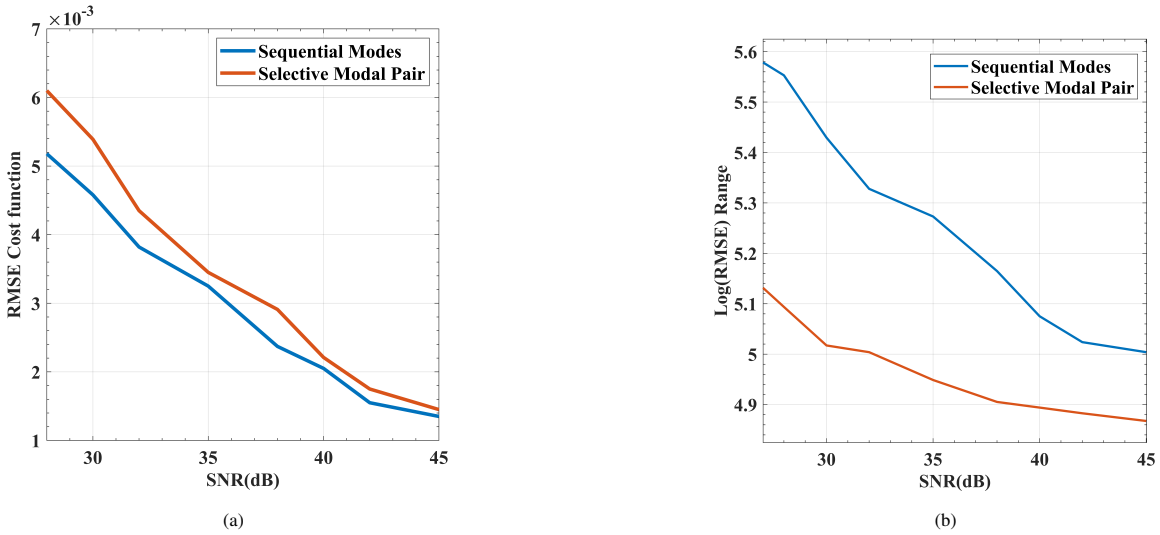


Figure 8: Time Frequency Representation (TFR) of (a) Cost Function RMSE (b) Log(RMSE) of estimated range for: $28dB < SNR < 45dB$

4.2 *Recorded North Atlantic Whale and Explosion Sound Localization*

In this section, we conduct a comprehensive evaluation of our proposed selective weighted algorithm using two distinct sound sources: the sound of a North Atlantic Right Whale and an explosion sound. Figure 9 depicts the time series and time-frequency (TF) analysis of these signals transmitted over different distances: 4.5 Km (z=20m) for the explosion sound and 8.7 Km (z=66m) for the whale sound. The TF analysis reveals that the noisy signal representing the explosion has a maximum frequency of 450 Hz, while the whale sound exhibits a lower maximum frequency of approximately 350 Hz. Furthermore, it is evident that certain modes are more susceptible to interference, highlighting the significance of modal selection and weighting functions in our approach.

We proceeded to localize the two signals and compared our results with our previous work and other existing methods. Table 1 presents the localization outcomes, demonstrating notable improvements compared to other proposed methods. Our Selective-modal based localization (SMP) approach achieved an error rate of 2.6% for both the recorded explosion sound and whale sound, while the Sequential Pair-Mode Analysis (SM) method yielded error rates of 3.11% and 6.2% for the respective signals. The superior performance of our proposed SMP method can be attributed to employing a larger number of dispersion curves (as opposed to only six sequential dispersion curves in SM) and performing initial modal selection.

Despite these improvements, it is important to note, as indicated in the TF analysis of Figure 9 and discussed in Section 3, that noise and channel effects vary across different modes. Consequently, each mode exhibits different weights and importance in the localization process, a consideration that is addressed in our approach.

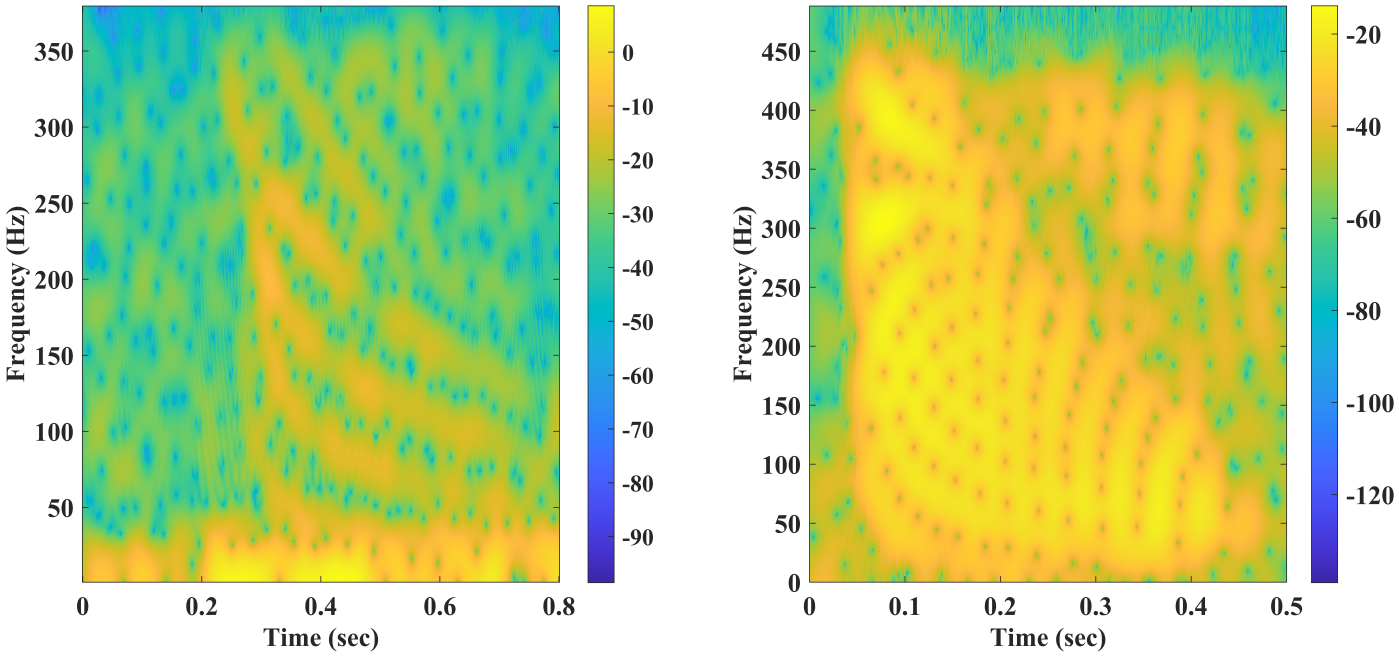


Figure 9: TF analysis: (a) North Atlantic Whale (b) Underwater Explosion

Table 1: Localization of Recorded North Atlantic whale (r=8700m) and explosion sound(r=4500)

Signal Source	Method	Number of Modes Used	Range (m)	Error (%)	References
Explosion Sound	Sequential Pair-Mode Analysis (SM)	6	4351	3.11	[17]
	Selective modal-Pair Analysis	9	4383	2.6	Proposed
North Atlantic Whale	Sequential Pair-Mode Analysis	4	9240	6.2	[17]
	Mode analysis	2	9225	6.03	[28]
	Downhill simplex algorithm	2	8884	2.11	[29]
	TOA	2	8950	2.87	[30]
	Selective modal-Pair Analysis	4	8881	2.06	Proposed

4.3 2D Localization

In this section, we conduct a comparative analysis of the 2D tracking performance of our localization algorithm in relation to other methods. Using the model description outlined in Figure 6, we employed a simulated non-stationary impulsive sound source that closely resembles the characteristics of a traveling whale following a sinusoidal path along the (x, y) axis.

Our 2D localization approach involves estimating the range of the sound source to each buoy, followed by triangulation based on the approximate direction of arrival and the intersection point of circles with a radius of r_h. The localization results for both the Sequential modes (SM) and our proposed Selective-modal based localization (SMP) are depicted in Figure 9, along with the true location of the sound source. It is evident from the results that SMP exhibits a closer adherence to the true range line compared to SM. This improved performance can be attributed to the modal selection function we introduced in this paper, which enables more accurate localization of the sound source.

5 Conclusion

In this study, we presented a passive impulsive sound source localization approach specifically designed for shallow underwater environments. Our method utilized the normal mode channel model and ambient noise to achieve accurate localization. A key contribution of this paper is the introduction of a localization scheme that incorporates modal pair selection, enabling enhanced noise resistance and improved accuracy.

Additionally, we proposed a 2D localization technique suitable for unsynchronized hydrophones, which aligns with the requirements of existing remote monitoring systems. To evaluate the performance of our algorithm, we conducted extensive analyses under various signal-to-noise ratio (SNR) conditions, comparing its noise resistance capabilities with other methods.

Furthermore, we validated our algorithm by testing it with actual recorded whale and explosion sounds. The results demonstrated its effectiveness in accurately tracking impulsive sound sources in a 2D space. Overall, our proposed approach showcases advancements in impulsive sound source localization and offers notable improvements over existing techniques.

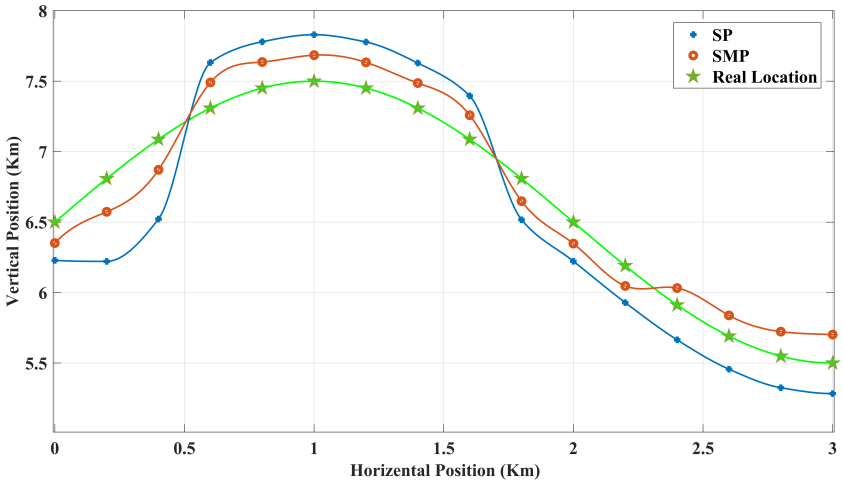


Figure 10: 2D localization and tracking of an impulsive sound source

References

[1] F. Talebpour, S. Mozaffari, M. Saif, S. Alirezaee, "Multi-Modal Signal Analysis for Underwater Acoustic Sound Processing," in 2022 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 300–305, IEEE.

[2] X. Su, I. Ullah, X. Liu, D. Choi, "A Review of Underwater Localization Techniques, Algorithms, and Challenges," Journal of Sensors, **2020**, 1–24, 2020, doi:10.1155/2020/6403161.

[3] R. Vaccaro, "The past, present, and the future of underwater acoustic signal processing," IEEE signal processing magezine, **15**(4), 21–51, 1998.

[4] V. Kavoosi, M. J. Dehghani, R. Javidan, "Underwater Acoustic Source Positioning by Isotropic and Vector Hydrophone Combination," Journal of Sound and Vibration, **501**, 116031, 2021, doi:10.1016/j.jsv.2021.116031.

[5] K. F. Bou-Hamdan, A. H. Abbas, "Utilizing ultrasonic waves in the investigation of contact stresses, areas, and embedment of spheres in manufactured materials replicating proppants and brittle rocks," Arabian Journal for Science and Engineering, **47**(9), 11635–11650, 2022.

[6] D. A. Abraham, Underwater Acoustic Signal Processing: Modeling, Detection, and Estimation, Springer, 2019.

[7] D. Neupane, J. Seok, "A Review on Deep Learning-Based Approaches for Automatic Sonar Target Recognition," Electronics, **9**(11), 1972, 2020, doi: 10.3390/electronics9111972.

[8] W. A. Kuperman, J. F. Lynch, "Shallow-Water Acoustics," Physics Today, **57**(10), 55–61, 2004, doi:10.1063/1.1825269.

[9] S. M. Wiggins, M. A. McDonald, L. Munger, S. E. Moore, J. A. Hildebrand, "Waveguide propagation allows range estimates for North Pacific right whales in the Bering Sea," Canadian acoustics, **32**, 146–154, 2004.

[10] M. Sanguineti, J. Alessi, M. Brunoldi, G. Cannarile, O. Cavalleri, R. Cerruti, N. Falzoi, F. Gaberscek, C. Gili, G. Gnone, D. Grosso, C. Guidi, A. Mandich, C. Melchiorre, A. Pesce, M. Petrillo, M. G. Taiuti, B. Valettini, G. Viano, "An automated passive acoustic monitoring system for real time sperm whale (Physeter macrocephalus) threat prevention in the Mediterranean Sea," Applied Acoustics, **172**, 107650, 2021, doi:10.1016/j.apacoust.2020.107650.

[11] L. An, L. Chen, "A real-time array calibration method for underwater acoustic flexible sensor array," Applied Acoustics, **97**, 54–64, 2015, doi: 10.1016/j.apacoust.2015.04.008.

[12] F. B. Jensen, W. A. Kuperman, M. B. Porter, H. Schmidt, A. Tolstoy, Computational ocean acoustics, volume 794, Springer, 2011.

[13] S. Khazaie, X. Wang, P. Sagaut, "Localization of random acoustic sources in an inhomogeneous medium," Journal of Sound and Vibration, **384**, 75–93, 2016, doi:10.1016/j.jsv.2016.08.004.

[14] G. Wang, S. Cai, Y. Li, M. Jin, "Second-Order Cone Relaxation for TOA-Based Source Localization With Unknown Start Transmission Time," IEEE Transactions on Vehicular Technology, **63**(6), 2973–2977, 2014, doi:10.1109/tvt.2013.2294452.

[15] Y. Zou, H. Liu, Q. Wan, "An Iterative Method for Moving Target Localization Using TDOA and FDOA Measurements," IEEE Access, **6**, 2746–2754, 2018, doi:10.1109/access.2017.2785182.

[16] P. Wu, S. Su, Z. Zuo, X. Guo, B. Sun, X. Wen, "Time Difference of Arrival (TDoA) Localization Combining Weighted Least Squares and Firefly Algorithm," Sensors (Basel), **19**(11), 2554, 2019, doi:10.3390/s19112554.

[17] J. Bonnel, A. Thode, D. Wright, R. Chapman, "Nonlinear time-warping made simple: A step-by-step tutorial on underwater acoustic modal separation with a single hydrophone," J Acoust Soc Am, **147**(3), 1897, 2020, doi:10.1121/10.0000937.

[18] E. Xu, Z. Ding, S. Dasgupta, "Source Localization in Wireless Sensor Networks From Signal Time-of-Arrival Measurements," IEEE Transactions on Signal Processing, **59**(6), 2887–2897, 2011, doi:10.1109/tsp.2011.2116012.

[19] R. Diamant, L. Lampe, "Underwater Localization with Time-Synchronization and Propagation Speed Uncertainties," IEEE Transactions on Mobile Computing, **12**(7), 1257–1269, 2013, doi:10.1109/tmc.2012.100.

[20] A. Thode, J. Bonnel, M. Thieury, A. Fagan, C. M. Verlinden, D. Wright, J. Crance, C. L. Berchok, "Using nonlinear time warping to estimate North Pacific right whale calling depths and propagation environment in the Bering Sea," Journal of the Acoustical Society of America, **142**(4), 2711–2712, 2017.

[21] H. Jia, X. Li, "Underwater reverberation suppression based on non-negative matrix factorisation," Journal of Sound and Vibration, **506**, 116166, 2021, doi:10.1016/j.jsv.2021.116166.

[22] Y. Tian, M. Liu, S. Zhang, T. Zhou, "Underwater multi-target passive detection based on transient signals using adaptive empirical mode decomposition," Applied Acoustics, **190**, 108641, 2022.

[23] C. T. Tindle, A. Stamp, K. Guthrie, "Virtual modes and the surface boundary condition in underwater acoustics," Journal of Sound Vibration, **49**(2), 231–240, 1976.

[24] E. Costa, L. Godinho, W. Mansur, Peters, "Prediction of Acoustic Wave Propagation in Underwater Step Problems via the Method of Fundamental Solutions," European Acoustics Association, 2016.

[25] C. L. Pekeris, Theory of propagation of explosive sound in shallow water, Geological Society of America, 1948.

[26] R. P. Hodges, Underwater acoustics: Analysis, design and performance of sonar, John Wiley and Sons, 2011.

[27] Q. Yang, K. Yang, "Seasonal comparison of underwater ambient noise observed in the deep area of the South China Sea," Applied Acoustics, **172**, 107672, 2021, doi:10.1016/j.apacoust.2020.107672.

[28] C. Gervaise, S. Vallez, Y. Stephan, Y. Simard, "Robust 2d localization of low-frequency calls in shallow waters using modal propagation modelling," Canadian Acoustics, **36**(1), 153–159, 2008.

[29] F. Desharnais, M. Côté, C. J. Calnan, G. R. Ebbeson, D. J. Thomson, N. E. Collison, C. A. Gillard, "Right whale localisation using a downhill simplex inversion scheme," Canadian Acoustics, **32**(2), 137–145, 2004.

[30] M. H. Laurinolli, A. E. Hay, "Localisation of right whale sounds in the workshop Bay of Fundy dataset by spectrogram cross-correlation and hyperbolic fixing," Canadian Acoustics, **32**(2), 132–136, 2004.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 3, 19-28 (2023)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Inferring Student Needs Based on Facial Expression in Video Images

Yu Yan*, Eric Wallace Cooper, Richard Lee

Information Systems Science and Engineering, College of Information Science and Engineering, Ritsumeikan University, Kusatsu, 525-8577, Japan

ARTICLE INFO

Article history:
Received: 28 February, 2023
Accepted: 26 April, 2023
Online: 15 May, 2023

Keywords:
Student needs
Online education
Facial Action Units
Exploratory Factor Analysis
Random Forests

ABSTRACT

Limited interactive communication modes between students and teachers in online environments may lead to teachers misinterpreting or overlooking student needs during online teaching. Students learning online may also hesitate to make their needs known even when latent desires in teaching flow, pacing, and review, may be beneficial to the quality of the learning experience. The objective of the study is to construct and test models to infer student needs based on the facial expressions of students while they are learning online. Several Random Forest models were constructed to infer the reported conditions and tested using facial expression data extracted from the videos as action units in Facial Action Coding System (FACS). Exploratory Factor Analysis (EFA) was adopted to extract and combine the highly-related facial action units for building the training and testing data. The testing of the inference model yielded a result of 0.028 on the mean average error (MAE). This result suggests these methods would contribute to the development of improved online learning systems that assist teachers in understanding in real-time how students are responding to a lecture or other classroom experience.

1 Introduction

With the sudden expansion of online teaching from 2020, due in large part to measures intended to prevent opportunities for transmission of the SARS-Cov-2 virus, teachers and students were suddenly confronted by many of the difficulties associated with online learning environments. Among those difficulties is significantly fewer points at which a student can indicate feedback to a teacher, or a teacher can take a quick reading of the room to investigate facial expressions, sounds, body movements, and the like. The objective of this research is to investigate the feasibility of using video of students to monitor their facial movements in order to infer needs the students may wish to communicate to the teacher without unnecessarily interrupting the flow of the class. Here, student needs refers to implicit or latent requests about class flow and pace, such as increased or decrease in teaching speed, review of material introduced in this lecture, requests for breaks, and the like. The concept, as envisioned here, uses only local video monitoring and therefore can avoid privacy issues, as well as video and audio resolution difficulties, involved when classrooms rely on direct video of student faces during learning. The present study extends a study originally presented in the 10th International Conference on Information and Education Technology (ICIET 2022) by describing an investigation of new methods of inferring student needs in an online teaching scenario and the results of testing these methods on the experiment data presented in the previous work [1]

During the 2020 and 2021 academic years, schools of all types but especially institutions of higher learning, greatly expanded the use of online platforms to deliver lectures and other teaching activities. Platforms such as Zoom [2], Skype [3], and Google Hangouts [4] were quickly adopted to allow student and faculty participation from home in order to minimize contact and quell the spread of the Covid-19 pandemic. Teachers who were mostly accustomed to lecturing, and students who had for the most part attended class, in a conventional classroom, quickly discovered that many common, established objectives of communication between student and teacher [5] may become far more difficult when learning online [6], [7]. For example, in a conventional classroom, teachers may more easily gauge student needs, such as, by periodically checking their facial expressions. In addition to monitor resolution and video quality questions, the direct use of video has also been a privacy concern for students studying in their own abodes or in shared spaces [8].

This study provides a model for automatically inferring student needs based on their facial expressions during online lectures. The Facial Action Coding System (FACS) [9] provides a methodology to identify human facial emotions by collecting a group of facial Action Units (AUs), which are collections of facial muscle movements. In this study, FACS is used to identify a student need from facial

*Corresponding Author: Yu Yan, Ritsumeikan University, 1-chome-1-1, Nojihigashi, Kusatsu, Shiga, Japan, yuyan@fc.ritsumei.ac.jp

muscle movement at a certain moment. The inputs of the model are the intensity levels of AUs. Exploratory Factor Analysis (EFA) [10] is performed to reduce the number of inputs and determine the most effective combinations of AUs. Finally, Random Forests (RF) [11], a popular machine learning method for classification models were adopted to implement the inference models.

This article differs from the previously published conference paper titled "Inference of Student Needs in an Online Learning Environment Based on Facial Expression" in the following aspects. The introduction was significantly changed and expanded to discuss the relevance of this system for online learning support. The inference model was implemented with ten RF models to infer each student need rather than using one neural network model to infer ten student needs at once as in the conference paper, showing significantly improved accuracy. The experiment settings were significantly expanded, adding detail to the experimental methods, as well as deeper analysis and discussion of the experimental results. This paper further confirms the feasibility previously presented methods with added accuracy reported and further details for implementation in future online learning systems.

2 Related Work

Previous studies have investigated methods of automatically assessing student emotions during classroom activities. For example, in [12], the authors describe a method to provide teachers with emotional signals from their students based on measurements of electrodermal activity. Such methods, while offering immediate and relevant signals, also require shipping and local set up of equipment that is not typically part of the student's hardware and software. This paper, on the other hand, describes methods that use hardware and software typically part of most student online equipment available in laptops (as in the present work) or other devices commonly used to participate in online learning.

Some approaches use the camera for eye tracking in an effort to use gaze detection signals, as such signals are thought to be relevant to both emotional responses and learning activities. For example, the authors in [13] propose a system that collects data on eye movements, such as blinking or the duration of a gaze in a single location, to determine how well a student is progressing in the visual contents during a lesson. Signals reporting on student concentration during a lesson, while certainly providing what could be processed into useful information for instructors, do not on their own give the teacher some understanding of how to respond in a positive manner. In other words, simply telling the teacher that students may not be concentrating does not necessarily assist the teacher in responding to that situation. Facial expressions allow acquisition of more specific states of emotional reaction than concentration alone.

Several systems have been developed to track student facial landmarks, head positions, facial actions, and eye movements in order to infer a student's emotional state. For example, the systems described in [14] and [15] propose models to measure student engagement during video engagement based on facial expression data. The system described in [14] uses the Microsoft Kinect camera to acquire input data for a model to infer labels attached by a separate set of video observers. The work in [15] uses OpenPose [16] in a similar manner to that in [13]. In such systems, video observers may not be able to pick up relevant but subtle cues in student expressions. Additionally, as noted above, tracking student emotions or concentration during learning does not necessarily yield information directly applicable to assisting teachers grasp student needs.

Therefore, the contributions of this study can be listed as follows:

- To propose a new approach to the improvement of online learning environments by providing teachers specific information about student status in terms of specific needs.
- To propose a novel model of inferring student needs based on their facial expressions and, according to the model proposed, an online education platform using commonly available device tools and processing power.
- To find and identify associations between specific needs and facial emotions.

3 Proposed Inference Model

A student needs inference model was proposed. Figure. 1 shows each component of the model. The following shows a detailed explanation:

- *Facial expression video recordings*: A front-facing camera continuously records facial expressions while students attend real-time online lectures or watch video lectures. The tracking system simultaneously transfers the facial expression recordings to the FACS server.

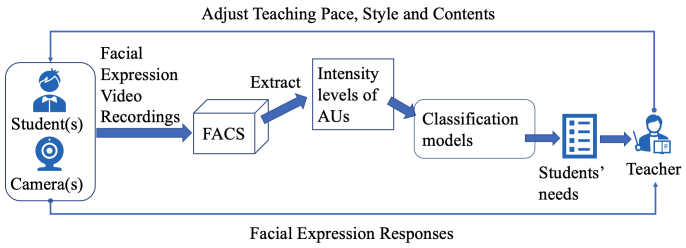


Figure 1: Model architecture for a student needs tracking system

- *Facial Action Coding System (FACS):* FACS detects human facial emotions like surprise or fear using forty-six facial Action Units (AUs), each of which represents a collection of human facial muscle movements. For example, levator palpebrae superioris, superior tarsal muscle facial muscle movements align to AU05's definition of "Upper lid raiser". Student needs depend only partly on their emotions. Therefore, the FACS method is used in this model to pre-process the video recordings. The input data of the classification models is composed of the intensity levels of the AUs from the pre-processing results. Here, an intensity level of each facial AU represents the confidence level of that facial AU.
- *Classification models:* In order to infer a list of student needs at a given moment, pre-trained supervised classification models are also necessary. These models are based on the intensity levels of the AUs.

- *Average of each need across all students:* In the end, the tracking system reports the average value of each need in the list, across all students, at a given moment. Ideally, the teacher will adjust the teaching pace, style and contents based on the reports from the tracking system.

This model can be used to build a student need tracking system for either real-time or on-demand online education. In the latter case, a teacher can adjust the teaching style and contents according to the overall output of the system.

4 Experiment Settings

This experiment simulates a situation in which students observe a college-level class given in an online video format with web cameras facing the students during the learning. Students were asked to watch educational videos that ranged from 8 to 9 minutes in length. Figure 2 shows the experiment flow for one viewing session. The experiment system automatically pauses the video every two minutes and requests that the student complete a survey on their current needs. After completing the survey the participant presses the "Play video" button on the in order to resume watching the video lesson. Other controls for the flow of the video were disabled in order to resemble real-time streaming participation. The web camera facing the student recorded facial expressions while the participants were watching the videos.

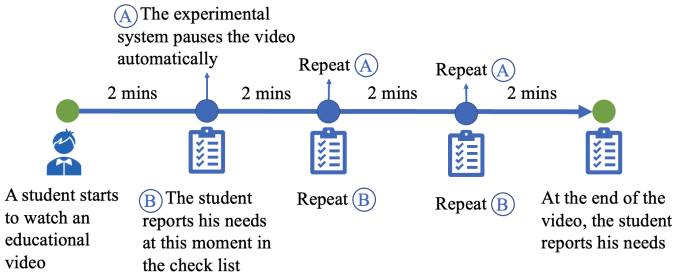


Figure 2: Human interaction experiment flowchart

The videos used in the experiment were three college-level educational videos, two selected from the Massachusetts Institute of Technology (MIT) OpenCourse [17] lectures on propositional logic and computing mathematics, and one of a recorded lecture given as a part of a C Language programming lecture at Ritsumeikan University, Japan, in which the topic is function declarations. This course is intended for students with no C programming experience and therefore precedes at a slower pace than the MIT courses selected. Due to privacy and consent considerations, the faces of students and teachers, as well as the voices of the students were omitted from this video. The three videos used in the experiments are each about fundamental computer science topics typically given to computer science students an intermediate level of difficulty in the curriculum. All three of the videos lectures employed a typical teaching format for such college courses, for example using slides to introduce and detail each main topic.

Seven students participated in the experiment. All of the students were in their 4th year of an undergraduate information systems engineering curriculum.

4.1 Creation of a Student Needs Survey

Based on previous studies and investigations of student needs, in [18] and [6], with consideration for their findings on needs that would be of practical use during teaching, the survey asked students about ten specific needs or requests for the teacher. Table 1 shows the list of student needs surveyed after each video session. These student needs were also intended to be helpful for learning and may be classified by their general teaching objectives:

- Needs that allow the teacher to adjust teaching pace in order to allow enough time for teachers and students to progress in their learning activities, which includes the needs numbered 01, 02, 03, 05, 06, 07, 08 and 10.
- Needs that inform the teacher on teaching style in order to increase engagement and understanding during the lecture and subsequent activities, which includes the need numbered 09.
- Needs that give the teacher feedback on adjust teaching contents so that the teaching materials may be more effectively paired with subsequent lectures, as in the need numbered 04.

Table 1: List of student needs investigated

Student Needs No.	Student Needs Descriptions
01	Please teach faster
02	Please teach slower
03	Please wait a moment
04	Please skip this part
05	Please go back to the last part
06	Please explain more
07	Please let me ask a question
08	Please let me take a break
09	Please make the class more interesting
10	I don't need anything; please continue

4.2 Facial Action Unit Extraction using OpenFace Technology

The FACS toolkit, OpenFace [19] was used to extract AU intensity levels (or AU values) in each recorded facial expression video. Seventeen AU features have been extracted as shown in Table 2. Although OpenFace can not extract all forty-six AUs as mentioned in Section 3, the seventeen AU features extracted were enough for the purpose of this study.

Additionally, an AU correlation matrix, which is given in Table 3, was also calculated to show the linear correlations among the AU features using all of the recorded videos. The AU correlation matrix calculation method is by Equation 1, which is based on the Pearson correlation matrix [20].

$$r = \frac{Cov(x, y)}{\sigma_x \sigma_y} \tag{1}$$

where, r is the correlation coefficient between two AU features: x and y, which are two sets of AU values.

The AU correlation matrix shows there are many features with correlation coefficients greater than 0.30 and some of the correlation coefficients are relatively high, such as the correlation coefficient (0.78) between AU01 and AU02. Therefore, factor analysis models may be implemented to identify interrelationships among the AU features. In this study, an Exploratory Factor Analysis (EFA) model was adopted to find the factor analysis clusters, which were also used to construct the sample data for training and testing the classification models.

Table 2: Action Unit feature list extracted by OpenFace

Action Unit No.	Action Unit Descriptions
AU01	Inner Brow Raiser
AU02	Outer Brow Raiser
AU04	Brow Lowerer
AU05	Upper Lid Raiser
AU06	Cheek Raiser
AU07	Lid Tightner
AU09	Nose Wrinkler
AU10	Upper Lid Raiser
AU12	Lid Corner
AU14	Dimpler
AU15	Lip Corner Depressor
AU17	Chin Raiser
AU20	Lip Stretcher
AU23	Lip Tightener
AU25	Lips Part
AU26	Jaw Drop
AU45	Blink

4.3 Construction of Sample Data using Exploratory Factor Analysis (EFA)

In order to implement EFA, three major steps were conducted:

1. *Assessment of the factorability of the AU features:* Both the Kaiser-Meyer-Olkin (KMO) test [21] (given by Equation 2) and Bartlett's Test of Sphericity (BTS) [21] (given by Equation 3) were performed on the AU correlation matrix.

$$KMO_j = \frac{\sum_{i \neq j} R_{ij}^2}{\sum_{i \neq j} R_{ij}^2 + \sum_{i \neq j} U_{ij}^2} \tag{2}$$

where, KMO_j is the KMO value for the given AU dataset; R is the AU correlation matrix shown in Table 3; i and j indicate the indices of the AU correlation matrix; and U is the partial covariance matrix.

$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right) \times ln|R| \tag{3}$$

where, p is the number of variables; n is the total sample size in the given AU dataset; and R is the AU correlation matrix shown in Table 3.

The KMO statistic was equal to $0.63 > 0.60$, which indicates that the collected AU features are adequate and it is appropriate to use EFA for the data. The BTS was highly significant with a test statistic of $146,276.26$ and an associated degree of significance, $p < 0.0001$, which shows that the AU correlation matrix has significant correlations among at least some of the features. Hence, the hypothesis that the AU correlation matrix is an identity matrix is rejected, also indicating that an EFA model is worthwhile for the AU features.

2. *Factor extraction:* In this study, Kaiser's (Eigenvalue) Criterion [22] and the Scree Test [23] were used to determine the number of the initial unrotated factors to be extracted. The eigenvalues associated with each component represent the total amount of variance that can be explained by this component. They were plotted based on the Scree Test. Six remarkable factors having an eigenvalue greater than one were retained. In the end, the Varimax rotation method [24] was adopted to implement the factor extractions.

3. *Sample data construction:* Six remarkable factors were used to describe the sample data for classification models in six dimensions. Each factor score for each dimension was calculated from the factor loadings extracted from EFA. The calculation is a weighted average as shown in Equation 4 .

$$Score(f) = \frac{\sum_{i=1}^{n} AU_i l_i}{\sum_{i=1}^{n} l_a} \tag{4}$$

where, f is the fth factor; n is the number of AU features involved; AU_i is the ith AU value in each sample; l_i is the corresponding factor loading for the ith AU feature.

4.4 Classification Models

Random Forests (RF) were adopted to build the student needs inference model. Figure 3 shows the architecture of the inference model. The inputs of the inference model were the floating factor scores, and the outputs were the ten student needs listed in Table 1. The factor loadings for each corresponding AU feature and the AU values for the final 300 frames in each video clip were used to generate the factor scores. Ten RF models were trained to target each of the ten student needs for classifying whether this sample data corresponds to the need or not, indicated with an integer value "zero" (not corresponding) or "one" (is corresponding). In the end, the outputs from the ten models were combined as the output of the inference model.

4.5 Construction of the Experiment

A web application was built on the frontend to show and control the flow of the educational videos and survey to students. The survey results were stored in a local web server. A web 720p front-facing camera was also used to record the facial expression videos.

Table 3: Linear correlation matrix for AU features: correlation coefficients above 0.30 are highlighted; columns and rows are corresponding to AU No. shown in Table 2

	01	02	04	05	06	07	09	10	12	14	15	17	20	23	25	26	45
01	1																
02	0.78	1															
04	0.16	0.13	1														
05	0.01	0.12	-0.10	1													
06	0.15	0.08	0.55	-0.11	1												
07	0.19	0.08	0.73	-0.13	0.70	1											
09	-0.09	0.00	0.00	0.06	-0.05	0.02	1										
10	0.24	0.20	0.16	0.02	0.45	0.24	0.01	1									
12	0.08	0.11	0.08	0.01	0.57	0.17	-0.04	0.68	1								
14	-0.08	0.02	0.17	-0.13	0.33	0.15	-0.04	0.31	0.37	1							
15	0.01	-0.01	-0.07	0.05	0.02	-0.07	0.02	0.12	0.03	-0.05	1						
17	-0.07	-0.04	-0.08	0.08	-0.01	0.00	0.05	0.15	0.06	0.03	0.36	1					
20	0.33	0.35	0.07	0.11	0.15	0.06	0.08	0.30	0.21	0.07	0.43	0.30	1				
23	-0.10	0.03	-0.13	0.36	-0.07	-0.08	0.07	-0.04	-0.01	0.07	0.01	0.42	0.11	1			
25	0.32	0.27	-0.03	0.11	0.04	-0.07	0.00	0.18	0.14	-0.04	0.07	-0.08	0.13	-0.02	1		
26	-0.01	0.12	-0.04	0.04	0.08	-0.02	-0.02	0.15	0.25	0.17	0.00	0.22	0.08	0.21	0.03	1	
45	0.31	0.21	0.01	-0.07	-0.01	0.02	0.08	0.01	0.03	-0.08	0.02	-0.04	0.04	-0.04	0.16	-0.06	1

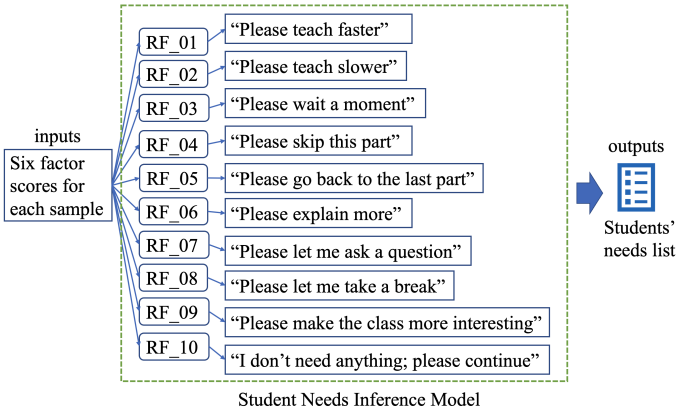


Figure 3: Student needs inference model architecure

Table 4: An example of labeled partial clips: the student need numbers are corresponding to those in Table 1

Student Needs No.	Clip No.			
	A2-1	A2-2	B2-1	C2-4
01	1	1	0	1
02	1	0	0	0
03	0	1	0	0
04	0	0	0	0
05	0	0	0	1
06	0	0	1	0
07	0	0	0	0
08	0	0	0	0
09	0	0	0	0
10	0	0	0	0

5 Experimental Results and Discussion

The experiment data includes a total of 84 segments of participant facial expression data, each recorded at the end of a two-minute viewing session. Two of the seven participants completed all three educational videos. Four participants watched two of the video lectures. One subject watched one video lecture. Each educational video includes four clips; for each clip, one survey result was produced. OpenFace recorded all AU intensity levels (ranging from 0.0 to 5.0) for each video frame of the participants face for an average of 3485 frames for each two-minute session. As mentioned in Section 4, the final 300 frames of each clip (ten seconds) prior to each survey response were used as the input data for training and testing the classification models. Therefore, the 84 data sets of 300 video frames each resulted in a total of 25,200 sets of AU intensity levels as input data in the models. The labels for the supervised classification models tested were each student need selected immediately after the given video clip. Table 4 shows an example of labeled clips; "A2-1" and "B2-1" represent two clips from two participants, where a "one" indicates that particular student need was selected in the survey and a "zero" indicates it was not selected.

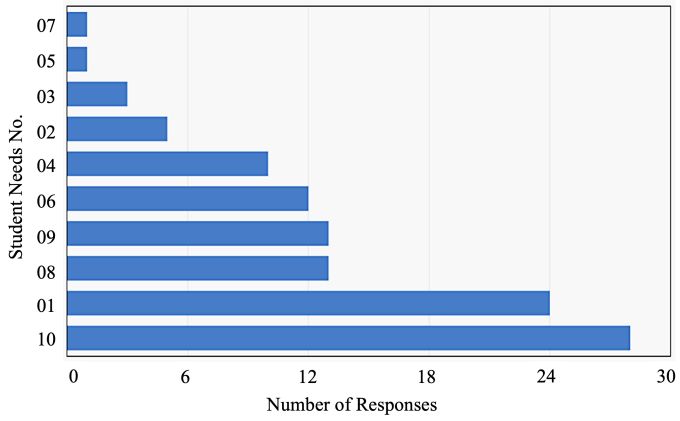


Figure 4: Distribution of student needs responses: the vertical labels are student needs numbers corresponding to those in Table 1.

5.1 Student Needs Survey Results

Figure. 4 shows the frequency for each response in the student needs survey data for all 84 sessions. The most frequently selected need was No. 10 ("I don't need anything; please continue") at 35% of the total responses. The second most frequently selected was

No. 01 (“Please teach faster”) at 29%. These results indicate that the level of difficulty in these lessons may have been slightly low but not to the degree that it would interfere with the collection of data about the other surveyed needs. The overall distribution of the frequencies lends support to the validity of the inclusion of these particular needs in the survey, with each one selected by at least one participant and no need selected more than half of the time.

Table 5: Eigenvalues (EV) and total variance explained:

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
01	3.289	19.347	19.347
02	2.254	13.257	32.604
03	2.027	11.925	44.529
04	1.416	8.331	52.860
05	1.217	7.478	60.337
06	1.042	6.130	66.467
07	0.997	5.863	72.331
08	0.844	4.965	77.296
09	0.769	4.525	81.821
10	0.712	4.186	86.007
11	0.623	3.667	89.674
12	0.454	2.669	92.342
13	0.407	2.394	94.736
14	0.319	1.876	96.612
15	0.270	1.588	98.200
16	0.158	0.930	99.130
17	0.148	0.870	100.000

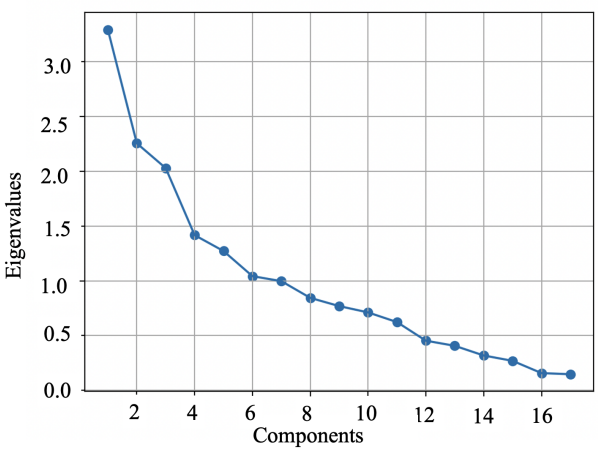


Figure 5: Scree Plot

5.2 Exploratory Factor Analysis Results

Based on the presumption that isolated facial actions taken out of context would be difficult for observers to interpret, EFA was used to determine effective combinations of the AU sets, as mentioned in Section 4. Table 5 shows the eigenvalues and total variance explained. Figure 5 shows the results of a Scree test, plotting the seventeen components on the x-axis and the respective eigenvalues for each number of components on the y-axis. Following the general rule of including the distinct linear components of the AU features with eigenvalues of greater than one, six of the seventeen components are included after extraction and rotation. The validity of this decision is further supported by the fact that the proportion of the total variance explained by the factors retained(66.5%) is greater than 50% [25]. Therefore, six factors were used for the subsequent EFA.

The present study performed EFA based on the Varimax rotation method. Table 6 shows factor loadings after EFA extraction, the mean and standard deviation of each corresponding AU features of all 25,200 frames. Here, the range of the intensity level of an AU is from 0.0 to 5.0 and is measured by OpenFace. AU features with loading values **less than** 0.40 are in grey, which indicates that they are not able to represent the corresponding factor.

Table 6: Factor loadings for each AU types

Action Units	Mean	SD	Factor Loadings
Factor 1: Scowling			
Brow Lowerer	1.002	1.045	0.765
Cheek Raiser	0.331	0.518	0.668
Lid Tightner	0.871	1.081	0.943
Factor 2: Squinting			
Upper Lid Raiser	0.451	0.501	0.687
Lid Corner	0.496	0.646	0.945
Dimpler	0.907	0.675	0.432
Factor 3: Blinking			
Inner Brow Raiser	0.299	0.682	0.957
Outer Brow Raiser	0.133	0.338	0.809
Lips Part	0.270	0.351	0.356
Blink	0.264	0.442	0.304
Factor 4: Frowning			
Nose Wrinkler	0.056	0.164	0.074
Lip Corner Depressor	0.137	0.293	0.796
Lip Stretcher	0.151	0.315	0.529
Factor 5: Raising			
Upper Lid Raiser	0.088	0.252	0.978
Factor 6: Pursing			
Chin Raiser	0.370	0.457	0.575
Lip Tightener	0.121	0.272	0.668
Jaw Drop	0.320	0.391	0.351

In addition, the top three highest mean values are highlighted. The AUs with the three highest mean intensity levels are “Brow Lowerer”, “Dimpler” and “Lid Tightener”, indicating that the brow, lid and dimple are the most significant signals. Each extracted factor was also given a name. These names do not necessarily reflect the emotions that might typically be expressed with these face actions. They are only intended as convenient labels for the discussion and analysis.

The factor analysis results appear to include principles of physical vicinity. For example, Factor 3 (Blinking) is associated with “Inner Brow Raiser” and “Outer Brow Raiser”. When a person elevates their inner brow, he is very likely to raise the outer brow as well. The sample data used for the inference was, therefore, the combinations between factor loadings and the actual AU values of

each frame in each facial expression video. Here, only AU features with loading values that are not in grey in Table 6 were used.

5.3 Classification Model Evaluations

The training and testing data for all of the classification models were based on frames rather than clips. They were randomly split in a training-to-testing ratio of eight to two from the 25,200 frames. During training, for all of the models, 5-fold Cross-Validation [26] was done to reduce overfitting.

Random Forest is a popular machine learning procedure which can be used to develop prediction models. In the random forest settings, many classification and regression trees are constructed using randomly selected training datasets and random subsets of predictor variables for modeling outcomes. Results from each tree are aggregated to give a prediction for each observation [27]. In this study, "sklearn.ensemble.RandomForestClassifier" was used, which is a class of the "sklearn" machine learning package to train and test the RF models. In addition, considering the training time and overall accuracy, the basic parameter settings of all of the RF models are {number of estimators: 100 (default); max depth: 40}.

Figure 6 shows the confusion matrices for each trained RF model for each student need using the test data. As shown in Figure 6, in terms of all RF model evaluation results, the performance on true negatives was higher than that of the true positives due in part to an imbalance of the training data where there were far fewer positives than negatives in each category. For example, the lowest ratio of positives to negatives was approximately 0.01, and the highest ratio was less than 0.50. The highest two true positives were 89.3% and 86.0% on the student needs "I don't need anything, please continue" and "Please teach faster", indicating the model has better prediction abilities when the amount of training data is higher. However, while the amount of the training data may affect the model performance, it is not the only factor that affects the performance. For example, when comparing the performance on the student needs "Please wait a moment" and "Please go back to the last part". The first need has a lower true positive rate than the second one, even though the amount of training data of the second need is twice the first one. In addition, the performance on false positive is worse than the false negative for all of the models, due in part to the issue of imbalance.

Figure 7 and Figure 8 show the validation curve and learning curve for each trained RF model for each student need using the test data. The left figure in each subfigure shows the validation curve, indicating the appropriate max depth for each RF model. The right figure shows the learning curve, summarizing the whole learning process during the 5-folder cross validation. The shading around the lines represents a 95% Confidence Interval (CI) [28] (given by the Equation 5) of each data point. From the validation curves, the maximum depth for nearly all of the RF models was between 20 and 30, indicating that a max depth in between 20 and 30 is appropriate and the most efficient for this task. The learning curves show increased accuracy with an increase in the number of cross-validations for most of the RF models.

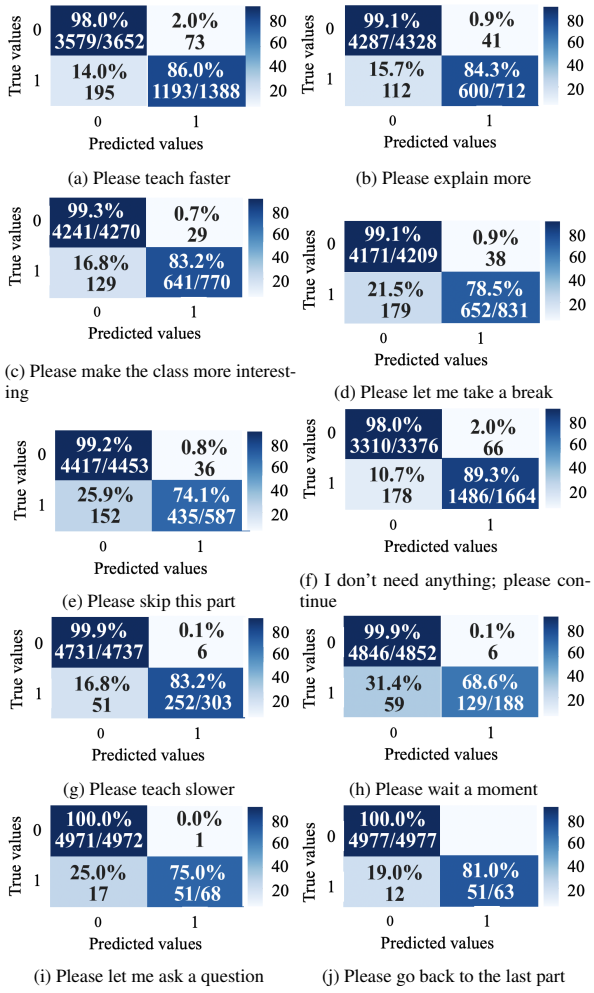
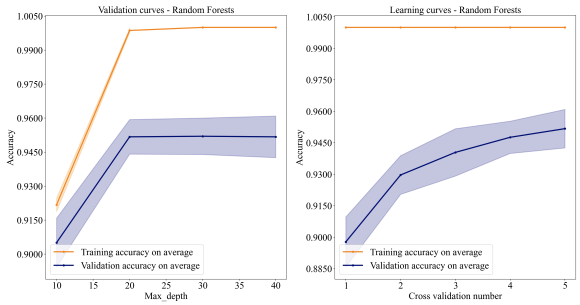


Figure 6: Confusion matrices for each trained RF model using the test dataset

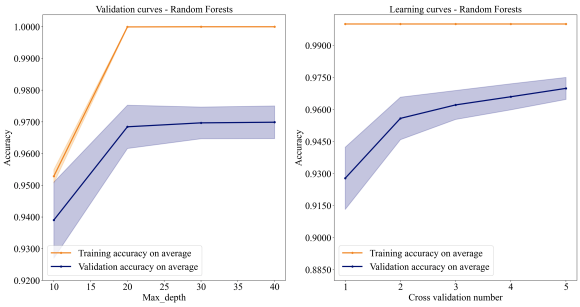
$$CI = \overline{X} \pm Z \times \frac{\sigma}{\sqrt{n}} \tag{5}$$

where, X is the mean of the training or validation scores; Z is the z-statistic for the confidence level (for 95%, Z = 1.96 approximately); n is the sample size.

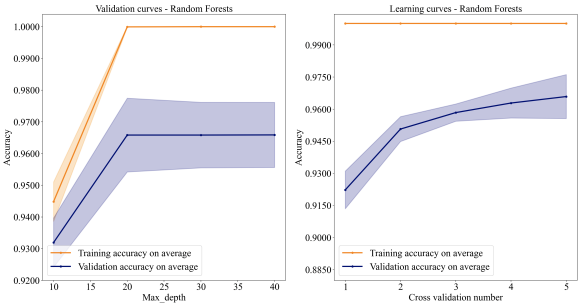
Finally, all of the RF model outputs were combined into one student needs list, including all ten possible needs. Table 7 shows the over-all evaluation of the inference model. Student need numbers: 08 ("Please let me take a break"), 09("Please make the class more interesting") and 10("I don't need anything; please continue") received relatively lower average errors than other needs. This may be because those needs are more related to emotions. Student need numbers: 06 ("Please explain more") and 01("Please teach faster") received higher mean average errors. This may be because of individual differences. For example, some people hide their emotions when they are thinking. The over-all MAE was 0.0283, which indicates the inference model could correctly infer more than nine students needs out of ten for each test video frame.



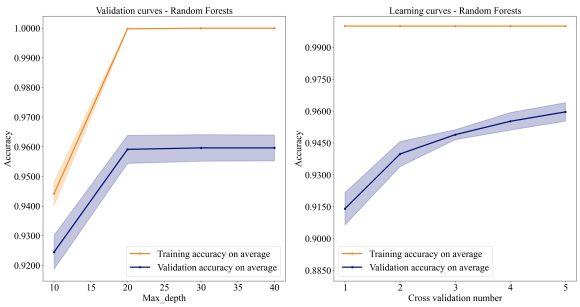
(a) Please teach faster



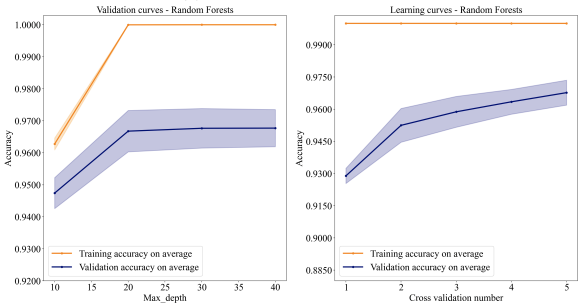
(b) Please explain more



(c) Please make the class more interesting

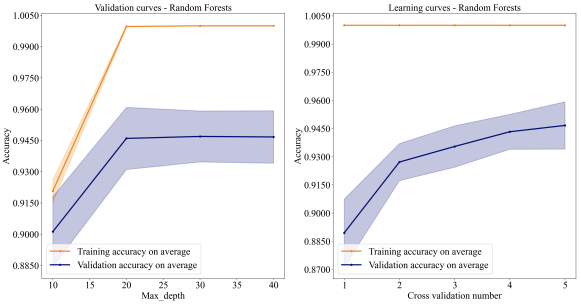


(d) Please let me take a break

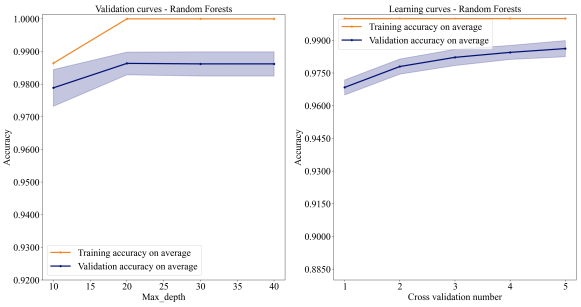


(e) Please skip this part

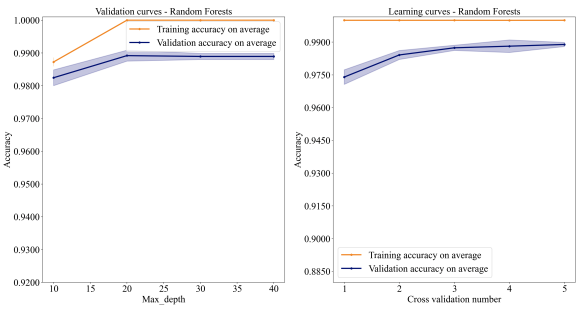
Figure 7: The validation curves (left) and learning curves (right) of RF models (RF01 to RF05)



(a) I don't need anything; please continue

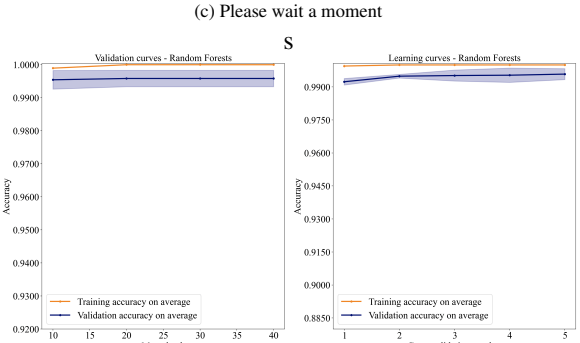


(b) Please teach slower

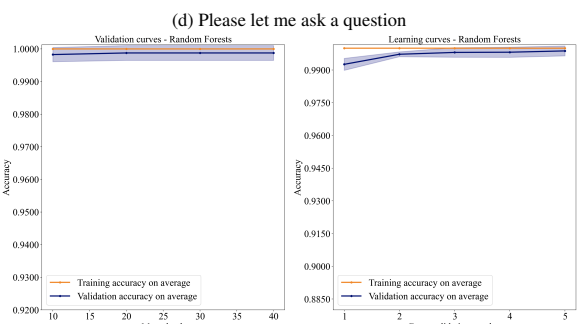


(c) Please wait a moment

S



(d) Please let me ask a question



(e) Please go back to the last part

Figure 8: The validation curves (left) and learning curves (right) of each RF models (RF06 to RF10)

Table 7: Mean average error (MAE) of the inference model for predicting each student need: the student need numbers are corresponding to the ones in Table 1

Student Needs No.	MAE for Each Need (1×10^{-2})
01	5.595
02	2.778
03	3.472
04	4.345
05	3.671
06	5.238
07	1.488
08	1.032
09	0.456
10	0.198
MAE of all sample data (1×10^{-2}): 2.827	

6 Conclusion

This study demonstrates the feasibility of inferring student needs from real-time video data in online learning situations based on models trained on video and survey data collected from students learning from video learning materials. The video data was extracted as facial points which were further encoded using facial expression modeling methods, all of which can be collected in real-time during online teaching with the permission of the student. The survey was collected at the end of each two-minute video segment to label the needs of the student in terms of what the student would prefer in terms of change (or no change) in the teaching. This survey and the system itself is intended to allow students to give meaningful feedback that teachers may use to assist in decisions about class flow and student interaction, including teaching speed, class breaks, review of new material, etc. Facial actions were further classified using factor analysis to result in a final set of input parameters for the inference based on what may be described as facial expressions.

Of several models tested to infer the student needs from the facial expression data, the Random Forest models performed best. The results show each Random Forest is trained to classify a single need performs very well, especially excelling at excluding individual cases of need at which the model consistently performs with greater than 98% accuracy. When indicating a need, the model accurately classifies at a greater than 80% accuracy, a rate that could be extremely useful for teachers who, in online scenarios, typically cannot closely monitor each face or stop the flow of a lecture to inquire about feedback. More importantly, this accuracy in a class of even a small group could amalgamate several students inferred needs to give teachers a more accurate reading of the general mood about the class without needing any video transmission, or allowing complete anonymity.

The proposed inference model can also be tested in real learning situation given a flexible method of interaction and, as similar data is collected on the models' accuracy in various situations, and the system accuracy improved, the interface may be adjusted to give teachers more confidence in relying on such systems to assist them in deciding who to respond to lecture flow and, eventually, to in responding to individual students. At the same time, as teachers become more comfortable with using such systems to determine the timing of specific feedback, they will also become more aware of the influence of their pacing and feedback on teaching in general.

Acknowledgment This study was done using funds from Japan KAKENHI grant No. 21K17865.

References

[1] Y. Yan, J. C. Lee, E. W. Cooper, "Inference of student needs in an online learning environment based on facial expression, " in the 10th International Conference on Information and Education Technology (ICIET), 113-117, 2022, doi: 10.1109/ICIET55102.2022.9779022.

[2] B. B. Wiyono, H. Indreswari, A. P. Putra, "The utilization of 'Google Meet' and 'Zoom Meetings' to support the lecturing process during the pandemic of COVID-19," in the International Conference on Computing, Electronics & Communications Engineering (iCCECE), 25-29, 2022, doi: 10.1109/iCCECE52344.2021.9534847.

[3] A. Karabulut, A. Correia, "Skype, Elluminate, Adobe Connect, Ivisit: a comparison of web-based video conferencing systems for learning and teaching, " in the Society for information technology & teacher education international conference, 484-484, 2008.

[4] H. Pratama, M. N. A. Azman, G. K. Kassymova, S. S. Duisenbayeva, "The trend in using online meeting applications for learning during the period of pandemic COVID-19: A literature review," Journal of Innovation in Educational and Cultural Research, **1**(2), 58-68, 2020, doi:10.46843/jiecr.v1i2.15.

[5] E. A. Skinner, M. J. Belmont, "Motivation in the classroom: reciprocal effects of teacher behavior and student engagement across the school year," Journal of Education Psychology, **85**(4), 571-581, 1993, doi:10.1037/0022-0663.85.4.571.

[6] Y. Nailufar, S. Safruddin, M. I. Zain, "Analysis of teacher difficulties in online learning on mathematics subjects," Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram, **9**(2), 280-288, 2021, doi:10.33394/j-ps.v9i2.4376.

[7] X. Lu, M. Wang, J. Fang, H. Liao, "Investigation on the difficulties and challenges of teachers online teaching in primary and middle schools of guangxi middle school," in the International Conference on Computer Vision, Image and Deep Learning (CVIDL), 542–545, 2020, doi:10.1109/CVIDL51233.2020.00-31.

[8] M. H. Rajab, M. Soheib,"Privacy concerns over the use of webcams in online medical education during the COVID-19 pandemic," Cureus, **13**(2), 2021, doi:10.7759/cureus.13536.

[9] P. Ekman, W. V. Friezen, "Facial Action Coding System (FACS)," Environmental Psychology & Nonverbal Behavior, 1978, doi:10.1037/t27734-000.

[10] A. G. Yong, S. Pearce, "A beginner's guide to factor analysis: focusing on exploratory factor analysis," Tutorials in quantitative methods for psychology, **9**(2), 79-94, 2013.

[11] G. Biau, E. Scornet, "A random forest guided tour," Test, **25**, 197-227, 2016, doi: 10.1007/s11749-016-0481-7.

[12] P. W. Kim, "Real-time bio-signal-processing of students based on an intelligent algorithm for internet of things to assess engagement levels in a classroom," Future Generation Computer Systems, **86**, 716– 722, 2018, doi: 10.1016/j.future.2018.04.093.

[13] M. Porta, S. Ricotti, C. J. Perez, "Emotional e-learning through eye tracking," in IEEE Global Engineering Education Conference (EDUCON), 1–6, 2012, doi:10.1109/EDUCON.2012.6201145.

[14] B. M. Booth, A. M. Ali, S. S. Narayanan, I. Bennett, A. A. Farag, "Toward active and unobtrusive engagement assessment of distance learners," in the Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), 470–476, 2017.

[15] X. Zheng, S. Hasegawa, M.-T. Tran, K. Ota, T. Unoki, "Estimation of learners' engagement using face and body features by transfer learning," in the International Conference on Human-Computer Interaction, 541–552, 2021, doi:10.1109/ACII.2017.8273641.

[16] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, **43**(1), 172-186, 2021, doi:10.1109/TPAMI.2019.2929257.

[17] D. Cecilia, S. Carson, K. James, J. Lazarus, "MIT OpenCourseWare: unlocking knowledge, empowering minds, " Science, **329**(5991), 525-526, 2010, doi:10.1126/science.11826962.

[18] D. Patel, D. Ghosh, S. Zhao,"Teach me fast: how to optimize online lecture video speeding for learning in less time?, " in the Sixth International Symposium of Chinese (CHI),160–163, 2018, doi: 10.1145/3202667.3202696.

[19] T. Baltrusaitis, A. Zadeh, Y. C. Lim, L. P. Morency, "Openface2.0: facial behavior analysis toolkit," in the 13th International Conference on automatic face & gesture recognition (FG 2018), 59–66, 2018, doi:10.1109/FG.2018.00019.

[20] J. Benesty, J. Chen, Y. Huang, I. Cohen, "Pearson correlation coefficient," Noise Reduction in Speech Processing, **2**, 2009, doi:10.1007/978-3-642-00296-0_5.

[21] C. D. Dziuban, E. C. Shirkey, "When is a correlation matrix appropriate for factor analysis? some decision rules," Psychological Bulletin, **81**(6), 358–361, 1974, doi:10.1037/h0036316.

[22] K. A. Yeomans, P. A. Golder, "The Guttman-Kaiser Criterion as a predictor of the number of common factors," Journal of the Royal Statistical Society. Series D (The Statistician), **31**(3), 221–229, 1982, doi:10.2307/2987988.

[23] B. Raymond, Cattell, "The Scree Test For The Number Of Factors," Multivariate Behavioral Research, **1**(2), 245-276, 1966, doi:10.1207/s15327906mbr0102_10.

[24] E. E. Cureton, S. A. Mulaik, "The weighted varimax rotation and the promax rotation," Psychometrika, **40**, 183–195, 1975, doi:10.1007/BF02291565.

[25] N. Shrestha, "Factor analysis as a tool for survey analysis," American Journal of Applied Mathematics and Statistics, **9**(1), 4-11, 2021, doi:10.12691/ajams-9-1-2.

[26] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation, " Statistics and Computing, **21**, 137–146, 2011, doi:10.1007/s11222-009-9153-8.

[27] J. L. Speiser, M. E. Miller, J. Tooze, E. lp, "A comparison of random forest variable selection methods for classification prediction modeling", Expert Systems with Applications, **134**(15), 93-101, 2019, doi: 10.1016/j.eswa.2019.05.028.

[28] H. W. David, L. Stanley, "Confidence Interval Estimation of Interaction", Epidemiology, **3**(5), 452-456, 1992.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 2, 24-29 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Temperature-Compensated Overcharge Protection Measurement Technology

Jin Uk Yeon[1], Ji Whan Noh[2], Innyeal Oh[*,1]

[1]Sunmoon University, Department of Advanced Automotive Engineering, Asan, Chungnam, 31460, Republic of Korea

[2]Korea Institute of Machinery & Materials(KIMM), Yuseong-gu, Daejeon, 34103, Republic of Korea

ARTICLE INFO

Article history:
Received: 04 November, 2022
Accepted: 05 February, 2023
Online: 11 March, 2023

Keywords:
Battery
Battery Management System
Isolation
Indirect Voltage Measurement
LED

ABSTRACT

Recently, many problems have been caused by battery fires. The existing BMS(Battery Managment System) measured the voltage of each cell of the battery through the physical connection between the battery and the control module. However, if a battery with up to 1000 VDC becomes inoperable due to an external factor, the battery is damaged, and accordingly, a large current of the battery breaks the control unit of the BMS with 5 VDC to 24 VDC, putting the BMS inoperable. If the battery is operated when the bms is in trouble, it poses a risk of battery fire.Recently, as bms technology was announced with a wireless function, battery information could be easily transferred from the outside, so that convenience was maximized, but stability is still weak. This paper physically separated the battery and control module by measuring the battery voltage depending on the strength of the LED by connecting the battery and LED. and furthermore, the measurement error should be less than 1 mV even when the temperature changes. In addition, it was designed to operate at a low output level of 200 μW to 360 μW using the sub-threshold section of the LED.

1. Introduction

Recently, with the development of battery technology, it is widely used in the field of electric vehicles, such as Uninterruptible Power Supply (UPS) & Energy Storage System (ESS), etc., demand for a system using lithium-ion batteries has increased rapidly. Accordingly, the demand for battery management system (BMS), which has a function of controlling and protecting a battery, is increasing [1]. However, as battery fires have recently occurred, many issues have been raised. As a result of the analysis of electric vehicle fire accidents in 2018, about 16 accidents were reported. Fire accidents involving electric vehicles occur during a collision, but they occur during charging, driving, and even parking. Therefore, this study proposes a system to identify the main causes of electric vehicle fire and prevent fire accidents. The battery fire is caused by various factor (same, lack of choice words), such as temperature and humidity of the surrounding environment, battery fire caused by overcharging, etc. [2]. However, most of the causes are that the overvoltage of the battery affects the control unit that uses a low voltage, which destroys the control unit, and accordingly, the control unit cannot react with it, leading to a battery fire. As a representative function of BMS, it monitors the voltage between cells of the battery to support an overcharge protection function and safely performs charging and discharging through battery cell balancing. However, it is reported that the control unit is destroyed due to an imbalance between a battery capable of representing 1,000 VDC or higher and a control unit operating between 1.2 V and 2.4 V, and a fire accident of battery is ensued due to BMS failure. As a fundamental solution, a new BMS system that supports the safe operation of the overcharge protection monitoring system by isolating the high voltage battery and the control module is proposed. The overcharge protection system of the existing battery system is shown in Figure 1. In the structure of Figure 1, for voltage measurement, switch 1 is closed to charge C1, switch 1 is opened, switch 2 is closed to charge C2, and the control unit is ADC(Analog to Digital Converter) and measured with C2 voltage [3]. In this measurement method, when a problem of a high voltage unit including a battery pack occurs, insulation is destroyed in a low voltage unit. Therefore, like the existing BMS, a circuit that monitors the battery voltage without physically connecting the control unit and the battery pack is required. This study applied an indirect measurement technology that measures the voltage without the physical connection between the control part and the battery pack. A circuit that monitors the voltage between battery cells while separating the high voltage and low voltage parts was proposed and implemented using this technology. Systems with detachable structures have some cases mentioned in existing communication circuits [4]. There is also a patent for monitoring voltage by applying a photo sensor, but it was not a technology used for isolation structures [5]. This study

*Corresponding Author: Innyeal Oh, innyealoh@sunmoon.ac.kr

https://dx.doi.org/10.25046/aj080203

proposes and implements an isolated battery cell monitoring system by applying LEDs and photo sensors.

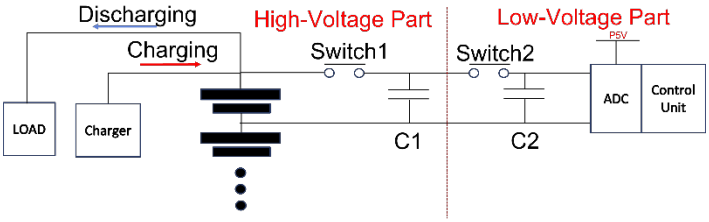


Figure 1: Existing Overcharge Monitoring Circuit [6]

2. Overcharge protection device design

2.1. Battery cell Indirect monitoring

In this paper, an indirect measurement technology of measuring the voltage through a medium without physical connection between a battery pack including a high voltage unit and a control unit is proposed and implemented. Indirect measurement technology is a technology that measures a change in the voltage of a battery by detecting a change in the brightness of an LED. To this end, the light of the LED was prevented from being emitted through the light blocker, and the light was measured using a photo sensor. A block diagram of indirect measurement technology is shown in Figure 2.

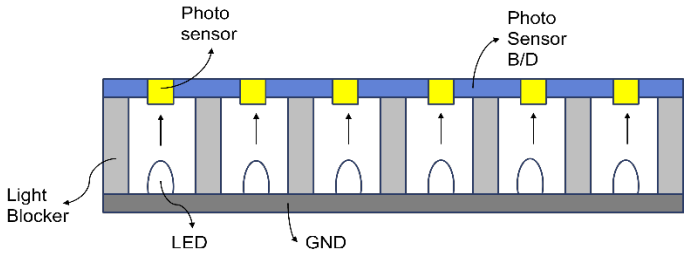


Figure 2: Block diagram of indirect measurement technology [7]

Normally, lithium-ion batteries are known to have an operating range of 3V to 4.3V. However, to measure a voltage using the indirect measurement technology in this paper, an alternative is required because the operating voltage is not matched with the lithium-ion battery. Therefore, this paper proposes an indirect measurement technology that uses a diode array to drop the voltage of the battery and measures the voltage through the corresponding change in the brightness of the LED. Figure 3 shows an indirect measurement circuit.

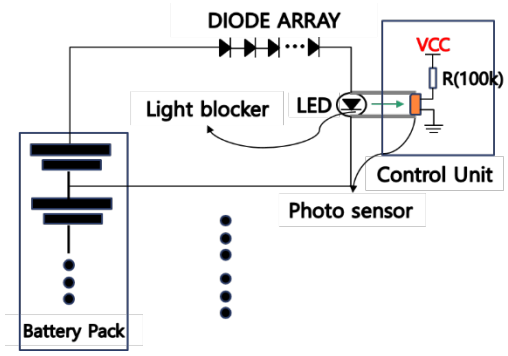


Figure 3: Indirect measurement circuit [7-8]

According to the structure shown in Figure 3, the LED connected to the battery is transmitted to the photo sensor through the light block. The resistance value of the photo sensor changes according to the brightness of light. It is connected to the VCC to measure the voltage applied to both ends of the photo sensor. Since the indirect measurement technology proposed in this paper measures the voltage using a medium, the brightness value of the existing LED is measured and stored in a memory included in the control unit. The voltage is estimated by comparing the brightness value stored in the memory with the currently measured brightness value.

2.2. Isolated LED & Photo sensor characteristics

2.2.1. Sensing characteristics for LED

The indirect battery voltage measurement method proposed in this paper measures the brightness of an LED whose brightness changes depending on battery voltage using an illuminance sensor called Cadmium sulfide (CdS). The resistance of CdS varies with the intensity of brightness. Therefore, the voltage at both ends of CdS is generated through the VCC connected to the front end of CdS, and the voltage is measured through the Analog to Digital Converter (ADC) according to the change in the resistance of CdS. In this paper, the resistance change of CdS measured through ADC is defined as Brightness. Figure 4 shows the graph of the relationship between the battery voltage and brightness measured at room temperature.

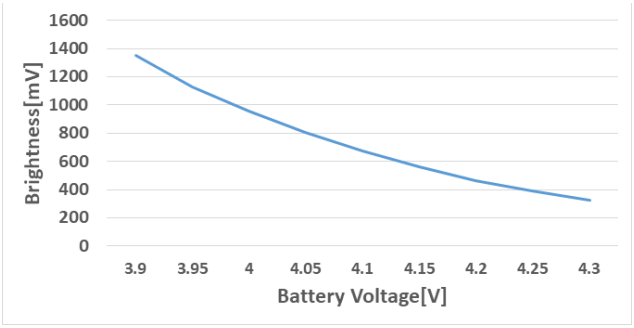


Figure 4: Relationship between Battery Voltage and Brightness

In the graph shown in Fig. 4, it can be found that the relationship between battery voltage and brightness has an inversely proportional relationship. The lithium-ion polymer used for the test in this paper has an operating range of 3.9V to 4.3V. Therefore, it can be confirmed that Brightness has an operation part of 1.3V to 0.3V from 3.9V to 4.3V, which are battery operation parts. In this paper, the voltage was estimated using the aforementioned relationship between the battery voltage and Brightens.

2.2.2 Light source wavelength

The peak spectral response of CdS used in this paper is between 500 nm and 600 nm. Therefore, it is necessary to adjust the spectrum band of the LED used to the spectrum band of the Cds used in this paper. Graphs of the spectrum band of CdS used in this paper and the spectrum band according to the color of LEDs are shown in Figures 5 and 6, respectively.

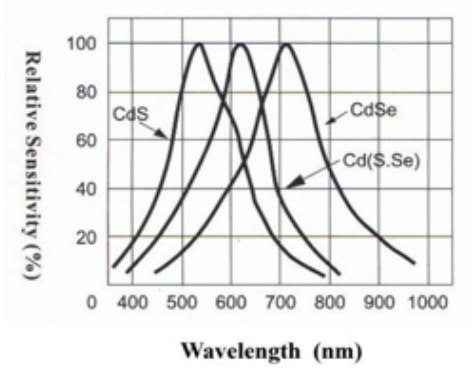


Figure 5: Spectrum band of the CdS [9]

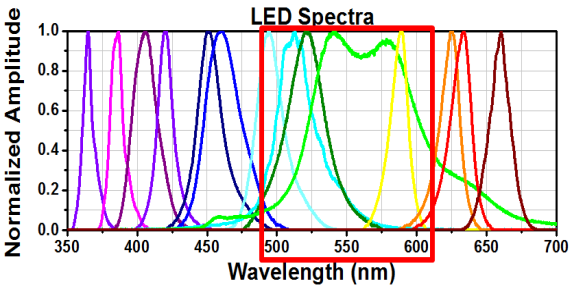


Figure 6: Color of LED according to wavelength [10]

In the graph in Figure 5, it can be seen that the wavelength of CdS is the most sensitive in the characteristics of about 500 nm to 600 nm. On the other hand, the red square box of the relationship between the wavelength length of the LED and the color of the LED in Figure 6 represents LEDs in the wavelength band of 500 nm to 600 nm, which is the wavelength of CdS. The LEDs in the red square box are green LED, blue LED, and yellow LED. Therefore, the LEDs suitable for the wavelength band of CdS, the photo sensor used in this paper, can be said to be Green, Yellow, and Blue LEDs. However, the indirect measurement technology in this paper is a circuit for cell monitoring of batteries, so it shall be possible to accurately measure the battery voltage. On the other hand, the intensity of light in an LED is determined by the current flowing through the LED. Therefore, the circuit including the LED should show a difference in current consumption that can be recognized according to the battery voltage.

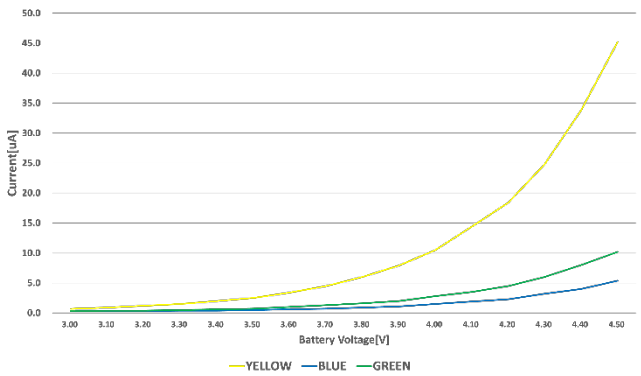


Figure 7: Current consumption according to the LED color

The current consumption graph of each LED is shown in Figure 7. Figure 7 shows the graph of measuring current by connecting 6 diodes of a diode array with LEDs in series. As shown in the figure above, it can be seen that the current of the Blue LED is 0.3 μA ~ 5.4 μA, Green LED is 0.3 μA ~ 10.2 μA, Yellow LED is 0.7 μA ~ 45.2 μA in the 3V to 4.5V Voltage range. Therefore it was confirmed that th Blue LED were not suitable for use in the indirect measurement circuit of this paper.

2.3. Low power behavior design

Since the battery voltage measurement technology operates using battery power, high power consumption leads to degradation of battery performance. Therefore, there is a need for a method to reduce power consumption due to the characteristics of indirect measurement technology that constitute a closed circuit at all times. Therefore, this paper proposes to use the sub-threshold section of LED as an operation section to reduce power consumption of indirect measurement technology. The motion graph of the sub-threshold is shown in Figure 8.

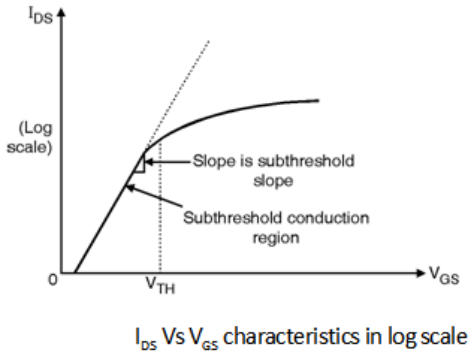
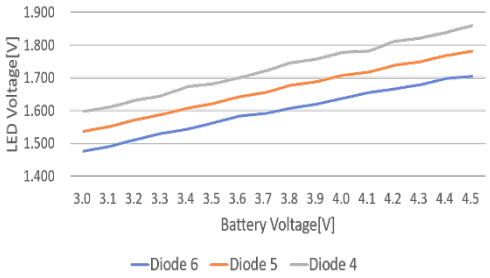


Figure 8: V-I characteristics of LEDs

The structure illustrated in Fig. 8 shows that a leakage current is generated even when a voltage smaller than an operating voltage is applied in a section operated at a point lower than the threshold voltage. The brightness of the light of the LED is proportional to the current and there is a current characteristic curve that changes according to the applied voltage. The general LED usage section is after the operating voltage, and current consumption is high in this section. However, it was confirmed that leakage current occurred even in a low voltage section of 0.7 V or less, and thus current consumption was low and LED was emitting light. It is a known technology in the field of transistor and is being used as a technology called sub-threshold swing. A recent study is also being conducted on the analysis of low-power CMOS inverters using sub-threshold swing [11]. This paper measured the voltage by adjusting the number of diode arrays and lowering the operating range of LEDs to the sub-threshold region. Figure 9 shows the voltage applied to the LEDs according to the number of diodes and the threshold voltage of the LEDs.

The LED used in this paper is the Yellow LED, and the wavelength is 590 nm. Therefore, as shown in Figure 9, the threshold voltage of the LED is about 1.75V to 1.97V [10]. If the voltage of the LED exceeds the threshold voltage, the current rises rapidly, so it should be designed so that the voltage does not exceed the threshold voltage in all sections. Therefore, in this paper, the number of diode arrays was selected and designed so that it does

not exceed the Threshold Voltage in all sections, as shown in the structure of the graph in Figure 9.



LED Wave length	Threshold Voltage
430	2.69 ± 0.3
565	1.91 ± 0.08
590	1.86 ± 0.11
627	1.83 ± 0.13
700	1.85 ± 0.05
850	1.3 ± 0.18

Figure 9: Voltage of the LED according to the number of diodes & Threshold Voltage of LED [12]

2.4. Temperature compensation

In a universal electronic circuit, the intensity of the current consumed by the elements of the circuit varies with temperature. Therefore, the indirect measurement circuit proposed in this paper also changes the current consumed according to the temperature. On the other hand, LED is a device in which the intensity of light changes according to the intensity of current consumed, so the value of brightness changes according to temperature. The change in Brightness according to temperature is shown in Figure 10.

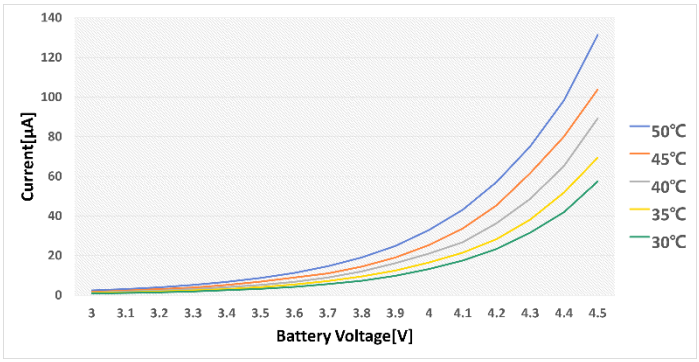


Figure 10: Brightness change according to temperature change of LED

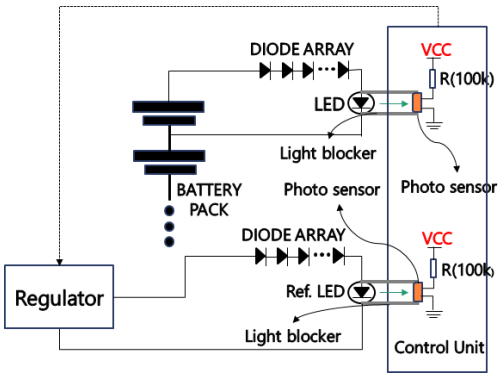


Figure 11. Temperature compensation circuit [8]

The graph illustrated in Fig. 10 shows that the difference in brightness at the same battery voltage at a low temperature and low voltage tends to be large. Therefore, an appropriate temperature compensation circuit is required. This paper measured a temperature corresponding to brightness by connecting a circuit, such as an indirect measurement circuit, with a regulator, and implemented a temperature compensation circuit by applying an offset to the brightness of each LED. The temperature compensation circuit of this paper is shown in Figure 11.

In the circuit shown in Figure 11, a regulator is the power supply that generates the corresponding voltage regardless of external factors, such as temperature, humidity, etc. Therefore, the Ref connected to the regulator. The LED emits corresponding light regardless of external factors. Measure the light using the photo sensor included in the control unit and estimate the temperature using the look-up table stored in the memory included in the control unit. The look-up table is schematized and shown in Figure 12.

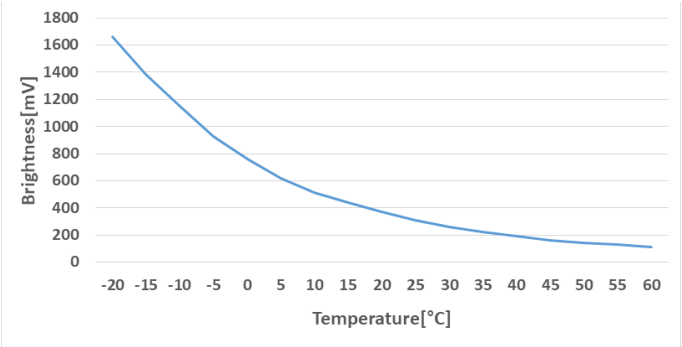


Figure 12: Change of Ref. LED's Brightness

2.5. Structure of Light blocker

With the indirect measurement technology proposed in this paper, it is very important to have a system that estimates a battery voltage by measuring the intensity of light of an LED according to a change in the battery voltage. In this paper, it is defined as a light blocker, a structure that prevents LED light from being emitted. The designed light blocker is shown in Figure 13.

Figure 13: Light blocker

The light blocker in Figure 13 is a structure designed to monitor 36 battery cells. Therefore, it is designed as a structure that can absorb light well so that it does not affect brightness measured according to external light. The light blocker is used in conjunction with the designed Printed Circuit Board (PCB) and can be secured to the PCB through a hole located at the apex of the light blocker. Therefore, it is possible to design so that light does not fade out.

3. Result & Measurement

3.1. Structure of the designed overcharge protection device

Indirect measurement systems are sensitive to the influence of external light. Therefore, it is important to solve structurally so that

external light does not affect the indirect measurement system. Therefore, in this paper, the external influence was greatly reduced by designing the PCB of the system in a multilayer structure. The structure of the designed PCB is shown in Figure 14.

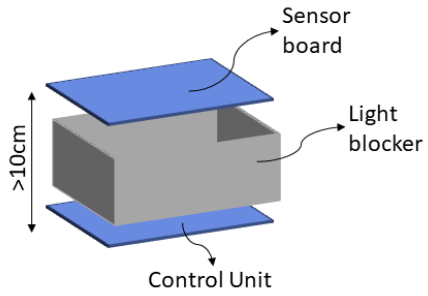


Figure 14: Structure of the designed PCB

In the structure illustrated in Fig. 15, the distance between the sensor board and the control unit was designed to be 10cm or less, so that the light of LED could be transmitted intact, and a size of PCB was designed to be small through the multilayer structure.

3.2. Measurement of overcharge protection device

The designed indirect measurement system is shown in Figure 15. The structure of Figure 15 is 36 LEDs installed in the control unit and 1 Ref. It is a system configured by combining a light blocker on an LED to block light, a photo sensor substrate on it, and a diode array substrate to a control unit. The board shown in Figure 16 is designed to measure the voltage of a total of 36 battery cells. The memory included in the control unit stores a lookup table for the brightness and voltage of the LED.

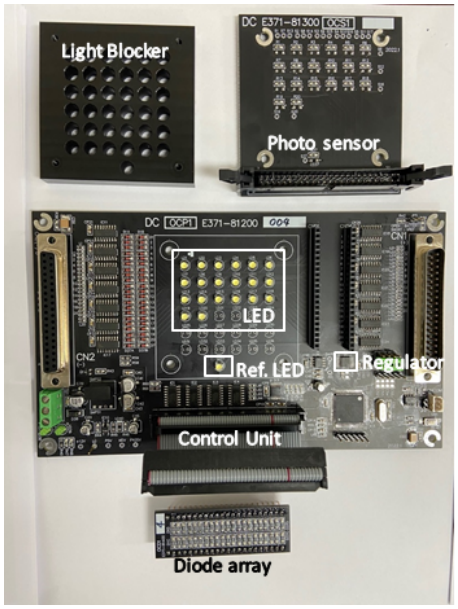


Figure 15: Designed Board

Table 1 shows the average of the values measured more than 10 times by applying a voltage from 3.9V to 4.3V to the LED channel of the designed board. It has an error value of 1 mV at a temperature change of -20°C to 60°C and the electric power amount is about 200 μW to about 360 μW.

Table 1: Results of Overcharge protection board

Battery Voltage[V]	Measured Voltage[V]	Power[㎼]	Current[㎂]
3.919	3.918	211.905	54.067
3.971	3.970	219.975	55.400
4.022	4.021	231.112	57.467
4.071	4.070	244.936	60.033
4.103	4.102	254.523	62.033
4.174	4.172	282.696	67.733
4.204	4.203	398.204	70.933
4.272	4.271	341.333	79.900
4.290	4.289	355.669	82.900

In addition, the comparison results with other papers are shown in Table 2. [4] In the case of , it is not a voltage measurement method used in BMS, but an electric line. Although it is a different application field, voltage measurement technologies in the form of isolation are compared. [13] In the case of , voltage measurement was carried out without physical connection using a wireless transceiver, and the indirect measurement method proposed in this paper measured the voltage using the brightness of light of an LED.

Table 2: Results of comparison with other papers

Paper	Error	Power	temp	Physical connection
[4]	- 40~30[mV]	-	- 20~40[°C]	O
[13]	5[%]	-	-	X
This work	1[mV]	200~350[㎼]	- 20~60[°C]	X

4. Conclusion

This paper proposes and implements an indirect measurement technology that measures voltage without physical connection between battery pack and control unit. Recently, BMS functions for safe battery use have been emphasized [14]. However, it was announced in 2022 as representative of Infineon's BMS (TLE9012DQU) and TI's wireless BMS (BQ79616-Q1), a group that presents advanced BMS technology. Infineon announced wired BMS technology, while TI and Linear Technology implemented wireless capabilities in BMS. The above technologies and this proposed technology are compared and analyzed and shown in the Table 3.

Table 3: Results of comparison with other BMS technology

No	Factor	Infineon BMS (TLE9012DQU) [15]	TI BMS (BQ79616-Q1) [16]	This work		Remark
				Characteristics	Explanation	
1	Physical isolation of high and low	Wired connection	Wired connection	Complete physical isolation	Safed Isolated BMS	Advantage

	voltage parts				Operation	
2	Communication interface with the master board	Wired serial communication	Wireless communication	Wired serial communication	Ease of data delivery	Disadvantage
3	Battery power consumption	20mW	20mW + wireless comm	200~350 [uW]	Low power operation	Advantage

As shown in the Table 3, the existing wireless technology maximizes convenience by providing wireless communication function to the battery so that battery information can be easily received wirelessly from the outside, but it is not implemented as a solution to a battery accident. However, the BMS technology proposed in this proposal is an original technology with a completely different isolation structure and has the advantage of minimizing battery consumption through safe battery operation and low power consumption operation of BMS while implementing BMS. The implemented system can measure the voltage of a total of 36 battery cells, and the voltage used is greatly reduced using the sub-threshold section. It was verified that the voltage measurement error was 1mV or less and the power amount was approximately 200μW to 360μW. In addition, a solution to the error according to temperature was presented using the LED, which is the reference point. Therefore, it can be confirmed that the technology proposed in this paper can have a safe battery monitoring system based on a complete isolation structure and a low power operating structure at the same time.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by a Korea Evaluation Institute of Industrial Technology (KEIT) grant funded by the Korean government (MOTIE) (No. 20022473, Development of 5KWh High-Safety Expandable Battery Module for Electric Vans and Electric Utility Cart)

References

[1] "Battery Management System," https://www.innopolis.or.kr/board/view?linkId=44443&menuId=MENU00999, INNOPOLIS, 2020.03.27

[2] D. Guo, Lei Sun, "The Causes of Fire Explosion of Lithium Ion Battery for Energy Storage," 2018 2nd IEEE Conference on Energy Internet and Energy System Intergration(EI2), Oct. 2018, DOI: 10.1109/EI2.2018.8582017

[3] T. Yue, Liji Wu, "High-precision voltage measurement IP core for battery management SOC of electric vehicles," 2014 12th IEEE International Conference on Solid-State and Integrated Circuit Technology(ICSICT), Oct. 2014, DOI: 10.1109/ICSICT.2014.7021570

[4] A. Hande, S.Kamalasadan, A.Srivastava, "A Selective Voltage Measurement System for Series Connected Battery Packs," Proceeding of the IEEE SoutheastCon 2006, May 2006

[5] D.I. Kim, J.I. Kim "Battery management system," Korean Patent Office,10-2019-0055524, 2021.04.29.

[6] M. Lelie, T.Braun, M. Knips, H.Nordmann, F.Ringbeck, H.Zappen, D.U Sauer, "Battery Management System Hardware Concepts: An Overview," Appl. Sci 534, 2018, 8, DOI: 10.3390/app8040534

[7] J.W. Noh, S.H. Ahn, H.S Kang, J.U. Yeon "Voltage Measurement Method, Voltage Measurement Device, and Battery System," Korean Patent Office, 10-2020-0111200, 2020.09.01.

[8] J. Yeon, KunSik Kim, Innyeal Oh,"Temperature-compensated Overcharge protection Indirect measurement circuit," 2022 Thirteenth International Conference on Ubiquitous and Future Networks(ICUFN),2022.07

[9] TOKEN, "CDS LIGHT-DEPENDENT," wep-site,PGM5 CDS Phtoresistors, 2010

[10] M.W. Davidson, "Fundamentals of Light-Emitting Diodes (LEDs)," Carl Zeiss Microscopy Online Campus,

[11] R Sindhu, Shilpa Mehta, "Sub-threshold Inverter for Low Power Consumption," 2018 Second International Conference on Inventive Communication and Computational Technologies(ICICCT), Apr. 2018.

[12] F. Santonocito, Antonio Tornabene, Dominique Persano-Adorno, "From led light signboards to the Planck's constant," 2018 Journal of Physics Conference Series, Sept 2018

[13] X. Zhang, Bin Yue, Jian Huang, Yuchuan Ruan, Peng Zhang, "Reserch on Non-contact Voltage Measurement Technology," 2019 IEEE 2nd International Conference on Automation, Electronics and Electrical Engineering(AUTEEE), Nov. 2019, DOI: 10.1109/AUTEEE48671.2019.9033249

[14] K.W. See, Guofa Wang, Yong Zhang, Yunpeng Wang, Lingyu Meng, Xinyu Gu, Neng Zhang, K. C. Lim, L. Zhao & Bin Xie, "Critical review and functional safety of a battery management system for large-scale lithium-ion battery pack technologies," International Journal of Coal Science & Technology, **9**, 2022.05, DOI: 10.1007/s40789-022-00494-0

[15] Infineon, "TLE9012DQU, Li-Ion battery monitoring and balancing IC," Infineon web-site, 2022.02.18

[16] Texas Instruments, "BQ79616-Q1",16-S automotive precision battery monitor, balancer and integrated protector with ASIL-D compliance," Texas Instrumnets web-site, 2022.09, DOI: 10.1109/MSSC.2022.3164853

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 3, 29-39 (2023)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Measurement System for Evaluation of Radar Algorithms using Replication of Vital Sign Micro Movement and Dynamic Clutter

Christoph Domnik[*,1,2], Daniel Erni[2], Christoph Degen[1]

[1]*Faculty of Electrical Engineering and Computer Science, Hochschule Niederrhein - University of Applied Sciences, D-47805 Krefeld, Germany*

[2]*General and Theoretical Electrical Engineering (ATE), Faculty of Engineering, University of Duisburg-Essen, and CENIDE – Center for Nanointegration Duisburg- Essen, D-47048 Duisburg, Germany*

ARTICLE INFO

Article history:
Received: 27 February, 2023
Accepted: 24 April, 2023
Online: 15 May, 2023

Keywords:
FMCW Radar
Vital sign detection
Dynamic clutter
Digital beamforming
Radar measurement system

ABSTRACT

In this paper we present a measurement system that is able to evaluate radar algorithms for vital signs sensing applications. For such medical applications, it is crucial to develop robust and reliable algorithms that are tested in a laboratory environment. The presented measurement system generates reproducible vital sign micro movement and dynamic clutter using loudspeakers to replicate realistic scenarios with two moving objects. It is described how realistic vital sign movement patterns are prepared using signal synthesis or recorded measurements, e.g. from a published dataset. The capability of the system to evaluate radar algorithms is demonstrated by investigating the impact of a beamforming algorithm on dynamic clutter. During the measurements presented in this paper, one loudspeaker replicates different vital sign movement patterns and the other loudspeaker creates dynamic clutter. It is shown that a digital beamforming improves the dynamic clutter rejection and leads to a better quality of the radar phase signal.

1 Introduction

This paper is an extension of work originally presented in the 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [1]. The concept of both publications is to create repeatable and realistic scenarios for testing, comparing and optimizing different post processing algorithms for radar technology used in vital sign sensing scenarios. Those algorithms should be tested in corresponding laboratory environment before used in clinical studies or real world applications.

Research in radar technology for medical applications recently gains increased attention. The ability of radar sensors to measure movement and micro movement like vibration [2] or movement from vital signs [3] [4] [5] leads to many applications. Doppler cardiograms, for example, can be used for measurements that make significant conditions of the heart visible comparable to the electrocardiogram [6]. Another discussed application is respiration movement measurement for motion-adaptive cancer radiotherapy [7].

Medical scenarios have special requirements for the technology used. It is very important to have a robust and reliable system. The system has to be easy in use for medical professionals without creating a huge amount of additional work. This means, the system must be tolerant against inaccurate positioning, random body movement of the target person [8] [9], radar self-motion [10] and clutter. A challenge for many medical scenarios is the clutter of other moving or stationary objects in proximity to the person, whose vital signs are to be measured. FMCW radars can help to distinguish objects at different distances. A limitation of the field of view by using a dielectric lens or smart antennas with beamforming capabilities could reduce this problem further. However, the dielectric lens would require a very accurate positioning of the radar, and problems in measuring the target due to unconscious movements occurred. This can be improved in terms of achieving a more robust condition with digital beamforming in post processing. By applying digital beamforming in post processing, a wide angle range is recorded first and, then, the target angle can be found and tracked afterwards. In this paper, we will address the mentioned effect by introducing a measurement system for evaluating radar post processing algorithms, e.g. beamforming algorithms.

In contrast to the setup presented in our previous paper [1], the setup that will be introduced in the following will use a second loudspeaker to replicate clutter from a second moving object. The

[*]Corresponding Author: Christoph Domnik, christoph.domnik@hs-niederrhein.de

previous paper [1] introduced a setup to generate reproducible movement patterns analog to chest wall micro movement from vital signs. That setup consisted of a single loudspeaker actuated by a function generator. It was shown that the diaphragm deflection measured by radars is linear to the voltage applied to the loudspeaker. Although this means that accurate measurements of the replicated micro movement are possible with the presented setup in [1], the replication of more realistic vital sign micro movement is problematic with that setup, which was addressed by the reviewers and numerous readers at the conference. To meet this request in this present paper, we will describe a new hardware setup using more flexible signal generation.

A loudspeaker for the replication of vital signs is also used in [11]. Another approach for generating movement of vital signs is introduced in [12] and [13]. There, a mechanical chest model for testing a wearable device is used to measure the chest circumference change. The simulation of the chest circumference change is not necessary for testing radar sensors and algorithms because the radar would only measure the movements effective to its range. A simulation approach based on a mathematical model that simulates the chest wall movement is presented in [14]. The chest wall motion simulation is useful to develop new algorithms for medical diagnosis while we focus more on the algorithms of the radar signal processing.

In the following chapters, at first, we will describe the radar signal processing used in this paper. Afterwards, the hardware of the measurement setup will be presented as well as the generation of realistic movement signals. Then, the impact of dynamic clutter will be investigated and a digital beamforming algorithm is tested concerning its rejection of the dynamic clutter. Finally, realistic heartbeat movement from the clinical recorded dataset from Schellenberger [15] will be used.

2 Radar and Signal Processing

A moving target creates a Doppler shift, which can be measured by continuous wave radars [16] as well as by FMCW radars [17]. Using FMCW radar sensors, multiple frequency ramps have to be captured consecutively. After range FFT and target finder algorithm mentioned in section 2.2, the range difference ΔR results in a phase difference $\Delta\Phi$. The phase difference $\Delta\Phi$ can be described as

$$\Delta\Phi = \frac{4\pi}{\lambda} \cdot \Delta R \tag{1}$$

using the wavelength λ [18]. The equation is unambiguous if $-\pi < \Delta\Phi < \pi$. Hence, larger macro movement of an object and a low sample rate of the phase signal can cause problems in measurements. It is also notably that only movement radial to the radar is measured.

2.1 Radar Hardware and Settings

During all measurements, the Radarbook2[1] from INRAS with the 77 GHz RF77II-IFX-TX2RX16_D01 frontend is used. Figure 1 shows a picture of the radar frontend.

[1] https://inras.at/en/radarbook2/

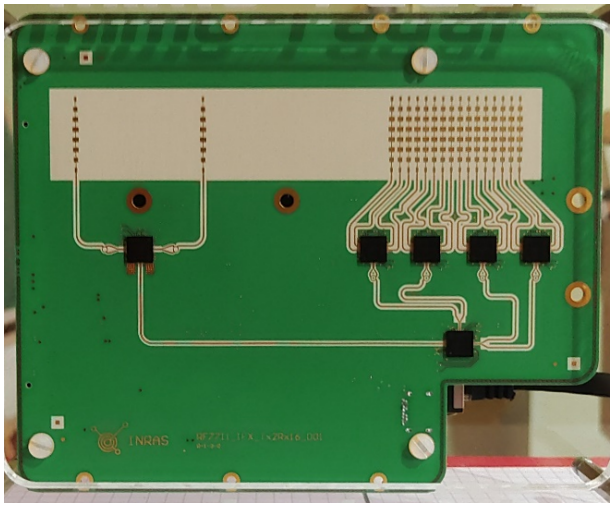


Figure 1: Picture of 77 GHz frontend of Radarbook2.

The Radarbook2 with 77 GHz frontend is configured according to the parameters depicted in Table 1. No further signal processing is done internally by the radar FPGA during the measurements conducted for this article. The raw ADC samples are transferred to the post processing device via LAN interface.

Table 1: Radar parameters and our default settings of Radarbook2 77 GHz. These default values are used, if not otherwise specified.

Radar antenna characteristic 3 dB beamwidth	Horizontal (RX, TX)	76.5°
	Vertical (RX, TX)	12.8°
	No. of used TX channels	1
	No. of used RX channels	16
Radar settings	Mode of operation	FMCW
	EIRP	3.2 dBm
	ADC sample rate f_s	5 MHz
	Frame time T_F	5 ms
	Samples per chirp N	1024
	Chirps per frame	1
Chirp configuration	Start frequency f_{Start}	76 GHz
	Stop frequency f_{Stop}	78 GHz
	Bandwidth B	2 GHz
	Ramp up time $T_C = \frac{N}{f_s}$	204.8 μs
	Slope $S = \frac{B}{T_C}$	9.7656 $\frac{\text{MHz}}{\mu\text{s}}$

2.2 Signal Processing

The key point of the presented measurement system is the evaluation of radar algorithms. This means that the software framework can be adapted to different algorithms used in different scenarios. In this work we use a post processing of the raw FMCW radar data as displayed in Figure 2.

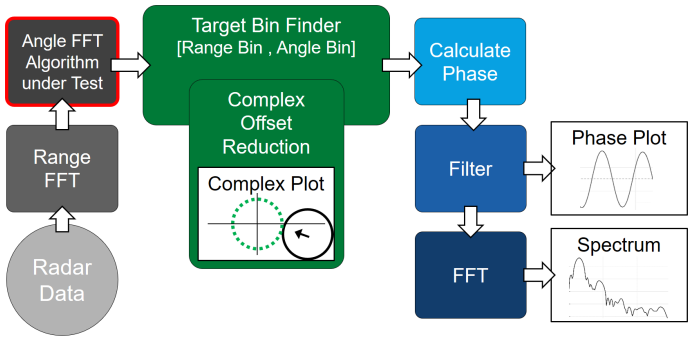


Figure 2: Block diagram of radar signal processing with angle FFT algorithm under test.

The first step in the signal post processing is a range FFT of the radar ADC data. The spectral lines of the range FFT stand for different distances. The range resolution or real bin width Bin_W can be calculated with

$$\text{Range Resolution} = Bin_W = \frac{c_0}{2 \cdot B} \approx 7.5\,cm \tag{2}$$

using the parameters presented in Table 1. Two different objects can only be distinguished if their distance to the radar has at least one bin width difference, otherwise they are measured as a single object. The distance R between radar and target object can be calculated with

$$R = Bin \cdot Bin_W \tag{3}$$

using Bin_W and the index of the spectral line where the target object can be found called Bin.

In this paper we use zero padding to increase the number of spectral lines with a factor of four. So, the bin with zero padding Bin_{ZP} result in $Bin \cdot 4$ and the reading accuracy of the distance will be higher but the range resolution does not change.

The next step, which is highlighted with a red outline in the block diagram in Figure 2, is an optional digital beamforming algorithm. As mentioned in the introduction, digital beamforming is of special interest in this work and will be used to demonstrate the benefits of this measurement system. The used radar has 16 receiving channels, which allows to calculate a digital beamforming post processing algorithm. Further explanation of digital beamforming algorithm and clutter rejection with and without digital beamforming can be found in chapter 5.4.

Necessary in any case is a target bin finder. The easiest way is to find the bin with the maximal magnitude. However, in measurements with static clutter this could lead to a wrong range bin and angle bin. Moreover, the phase signal of the target can be distorted by static clutter. We use the complex offset reduction algorithm (COR) for elimination of static clutter as signal processing step shown in Figure 2.

After the range and angle FFT, the complex signals of a moving target without clutter will be on a circle in the complex plot. The center of the circle is in the point of origin. When the magnitude of the target is changing over time, the complex plot can show a helix but the center has to be in the point of origin. When the center of the complex signal is not in the point of origin, this is because of a second, not moving object in proximity to the target object, where the overall effect is attributed to static clutter. More about static clutter can be found in [17]. Our COR algorithm finds the complex offset and subtracts it from the complex signal. This will be done in an area around the previously found range bin and angle bin.

After COR, the target bin finder determines the correct maximal magnitude again and, then, outputs the target data at correct range and angle bin to the phase calculation. The phase signal is calculated using the arctangent function and a phase unwrapping. After the phase calculation, other signal processing algorithms can be used for performance tests, e.g. in medical applications. An example for a performance test of beamforming algorithms is presented in chapter 5 by calculating the clutter rejection with and without digital beamforming.

2.3 *Measurement of Human Heartbeat and Respiration*

Using the previously described radar sensor and signal post processing, we have done measurements with a sitting person. The person was sitting still on a chair with a distance of 60 cm to the radar. The respiration and heartbeat curves are shown in Figure 3. The respiration measurement was recorded under resting conditions, and the heartbeat measurement was recorded during an apnea after exhalation scenario.

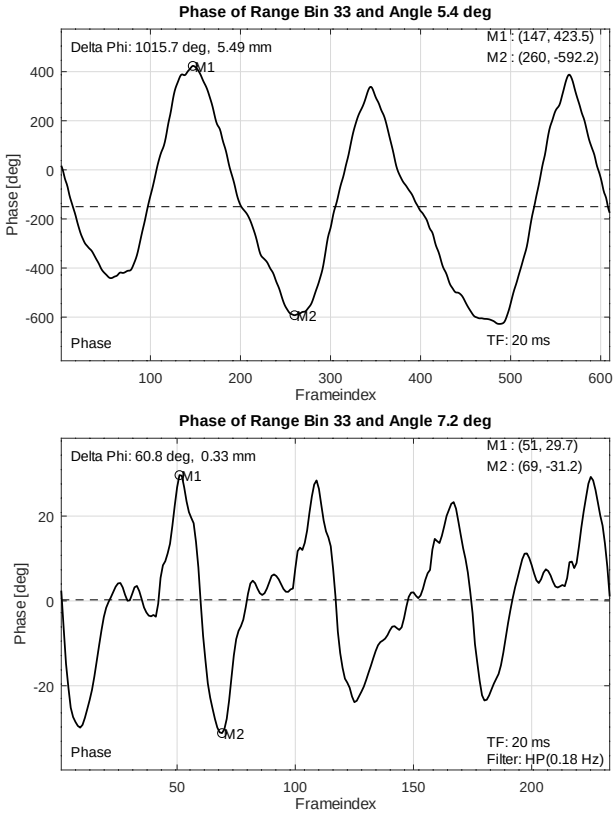


Figure 3: Radar measurements of human respiration (top) and heartbeat (bottom).

The movement amplitude of the respiration measurement is 5.49 mm with a breathing rate of approximately 15 breaths per minute. For the heartbeat measurement, the movement amplitude

is 0.33 mm with a heart rate of approximately 52 beats per minute. The maximal movement that can be expected from the heartbeat is 0.6 mm in average according to different studies mentioned in [19].

3 Hardware Setup

In this chapter, the hardware of the introduced measurement system is described. It is the foundation for the investigations of this work. In our previous work we evaluated a loudspeaker setup and different radar sensors in regard to the ability of replicating and recording micro movement from vital signs. The results showed that "the setup is particularly suitable for the generation of micro movement analog to CWmM[2] of vital signs. The diaphragm deflection is linear to the applied voltage [...]." as outlined in [1]. This result is displayed again in Figure 4. The following sections describe the differences of the hardware used for the present paper.

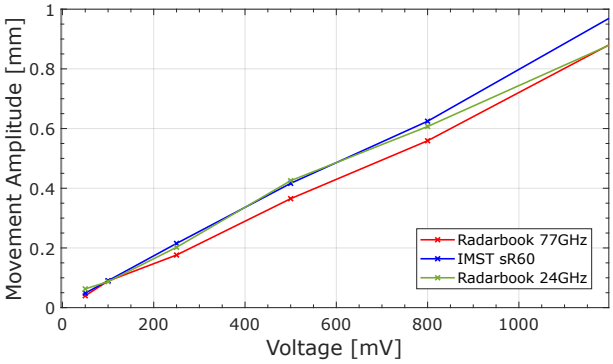


Figure 4: Measured micro movement using three radars at different voltage amplitudes on the loudspeaker. Figure from [1].

3.1 Loudspeaker

As mentioned in the introduction, the evaluation of beamforming algorithms is an important feature of the presented measurement system. In order to realize this, it is necessary to have two loudspeakers that can be positioned independently. In preparations of the measurements, we built two identically constructed casings for the low-frequency loudspeakers W 300 - 8 Ohm from Visaton[3]. The specifications of the loudspeaker include a maximal stroke of ±14 mm and a DC impedance of 6.6 Ω. The outcome is depicted in Figure 5.

We applied copper foil to both loudspeaker diaphragms. The diameter of the copper foil is 22.3 cm, which is much larger than it was in the previous setup. A new feature, however, is the static foil ring around the diaphragm. The foil ring is used to shadow the inside of the loudspeaker that is not covered with the copper foil on the diaphragm. The static clutter that is resulting from the foil ring can be eliminated using the complex offset reduction mentioned in 2.2.

[2]CWmM abbr. of chest wall micro movement
[3]https://www.visaton.de/de/produkte/chassis/tieftoener/w-300-8-ohm

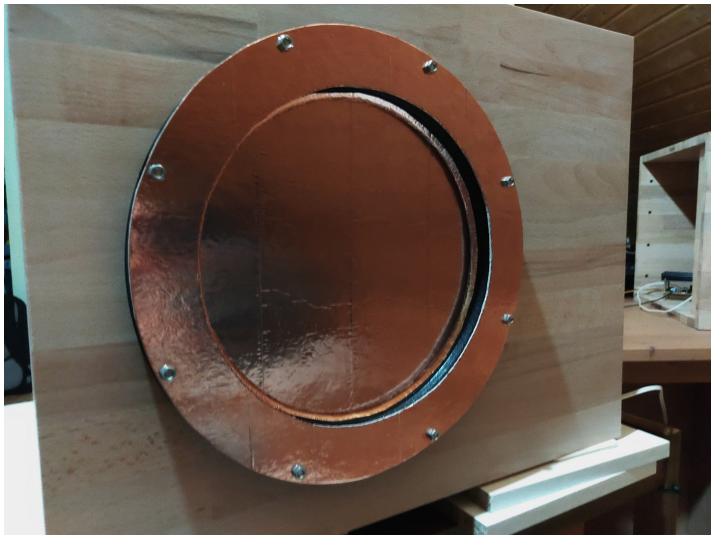


Figure 5: One of the new loudspeakers with copper foil on diaphragm and foil ring.

3.2 Amplifier

The loudspeakers need a suitable signal to generate a movement. In [1] a function generator is used to drive the loudspeakers, but this approach is very limited as mentioned before. A flexible strategy is necessary to reproduce real chest wall micro movement from vital signs. The best results were achieved using a sound card and an amplifier built for this measurement system. The amplifier is based on the TDA 2050 audio amplifier. All circuits that usually block low frequencies are excluded in the amplifier circuit. In Figure 6 you can see an amplifier.

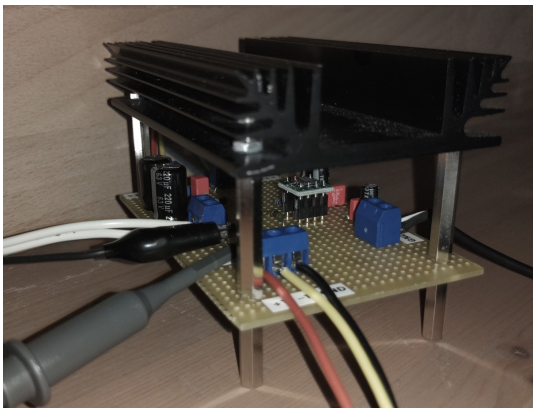


Figure 6: One of the new amplifiers.

The circuit also contains a low pass filter to prevent aliasing during the recording by the radar. The cut off frequency is currently adjusted to 10 Hz. Furthermore, an attenuation is integrated for achieving very small movements with the full bandwidth of the sound card. Different attenuation factors can be used to optimize the system for different micro movements like heart beat and respiration. The movement achieved with the current configuration is up to 10 mm using the full sound card's dynamic range.

4 Signal Preparation for the Measurement System

The hardware described in the previous chapter needs audio signals to generate micro movement. In this chapter, we will explain how audio signals for our measurement system are generated. In our previous work, we took a simpler approach to generate movement with a loudspeaker. It is further described in [1]. There, we applied a waveform generator to generate the signal that moves the loudspeaker. The signal applied in the previous setup was a sine signal for replicating respiration movement and an asymmetric sine signal for replicating heartbeat movement. Unfortunately, this approach is limited in terms of replication of realistic vital sign micro movement or of special conditions of the pulmonary or cardiovascular system like diseases. Therefore, the generation of realistic signals was a key point in the development of the new measurement system.

Because the signals are all played back by a sound card, they have to be saved as regular audio files. For the measurements for this article, the audio files are saved in the uncompressed waveform audio file format (.wav). The signals are saved with a sampling frequency of 44100 Hz and a resolution of at least 16 bits per sample.

4.1 Signal Synthesis

One strategy to generate movement signals is to synthesize them. It is possible to create a signal based on signature points and interpolation. Another approach is to use Fourier-synthesis for periodical signals.

For the signal synthesis we started with the generation of sine signals as fundamental oscillations. Those fundamental sine signals were used as signals during testing and calibration of the hardware setup. Additionally, in some measurements it is advantageous to use sine signals to show an investigated effect more clearly. This is why a sine signal is used for each loudspeaker during the investigation of dynamic clutter and beamforming algorithms that will be described in chapter 5.

By using Fourier-synthesis, periodical signals with higher complexity are created in the following:

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{N} a_k \cdot cos\,(k \cdot \omega \cdot t) + \sum_{k=1}^{N} b_k \cdot sin\,(k \cdot \omega \cdot t) \quad (4)$$

Figure 7 shows a synthesized heartbeat signal using 0.88 Hz as fundamental oscillation and eight harmonics. The Fourier coefficients are a_0=0, a_k=0 and b_k=[1.00, 0.97, 0.35, 0.06, 0.03, 0.03, 0.11, 0.06, 0.05] for k = 1,2, ..., 9. The coefficients b_k are optimized to synthesize a heartbeat movement signal based on the frequency components of the heartbeat movement measurement presented in chapter 2.3. It has much more resemblance to a real heartbeat micro movement than the asymmetric sine signal used with the previous setup in [1].

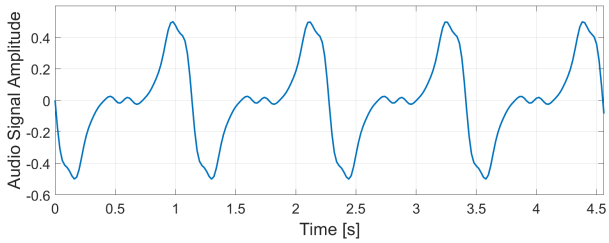


Figure 7: Pattern of heartbeat micro movement created with signal synthesis.

4.2 Vital Sign Micro Movement Patterns

For replication of real respiration and heartbeat movement, it is utterly important to use recordings of previous measurements as data source. Those recordings contain movement patterns of diseases and relevant conditions of the pulmonary or cardiovascular system, which are indispensable in the development of medical analysis and radar algorithms. Each measurement containing a movement pattern of the chest wall could be used independently of the technology that recorded those movement patterns. This gives us the opportunity to use published libraries of micro movement from vital signs. In our work we used the clinical recorded dataset published from Schellenberger in [15]. Figure 8 shows a heartbeat signal extracted from the dataset. The extracted part of the signal shows the heartbeat during apnea after exhalation from the measurement GDN0009_3_Apnea from the clinical recorded dataset.

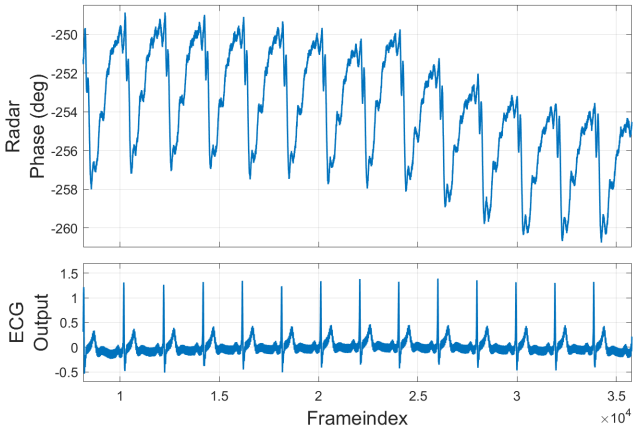


Figure 8: Heartbeat movement GDN0009_3_Apnea data from Schellenberger [15]. Top curve: Radar phase calculated from CW radar IQ values of the dataset. Bottom curve: Reference ECG measurement of the dataset.

The measurements were done using a 24 GHz continuous wave radar. Also signals of synchronized reference sensors like three channel ECG and Continuous Noninvasive Arterial Pressure (CNAP) were recorded during each of their measurements. For more information on how the dataset is recorded please refer to [20] and [21]. The following paragraphs explain how to prepare previously recorded data containing chest wall micro movement to use them as signals for our radar measurement system.

The first step is to cut one part from the recorded measurement with the targeted length and content. If the signal contains decisive changes like the shift from respiration to heartbeat in an apnea after

exhalation scenario, it would contain a DC offset during the heartbeat movement. This DC component will not be included in the played signal. The part of the heartbeat signal in Figure 8 starts with a QRS complex followed by the systole part of the cardiac cycle. The systole containing the contraction of the heart is visible as a falling edge in the radar phase signal of the dataset.

The next step is the scaling of the signal amplitude. For data recorded with radar sensors, the movement amplitude can be calculated using (1). Then the signal has to be scaled to the corresponding audio signal amplitude to generate the calculated movement amplitude. For special investigation or medical scenarios, movement amplitude can be increased or decreased during this preparation step. The movement amplitude of the dataset signal shown in Figure 10 as green curve was increased from 0.15 mm to approximately 0.5 mm.

After scaling the signal, it is necessary to adjust the sampling rate to 44100 Hz. As mentioned in chapter 2 we are recording our measurements with a sampling frequency of 200 Hz, and the measurements from the clinical recorded dataset published in [15] are recorded with a sampling frequency of 2 kHz. That is, oversampling to 44100 Hz is the next preparation step with additional low pass filtering to reduce oversampling artifacts. This filter also prevents aliasing during recording with the radar. Because of the typically configured frame time of 5 ms the highest frequency in the movement signal must be lower than 100 Hz. The signal is now ready to be used in the measurement system to generate realistic vital sign micro movement.

A performance test of our measurement system with play back of the prepared signal and recording with the radar was realized. Figure 9 shows the complex data of the radar recording with complex offset reduction.

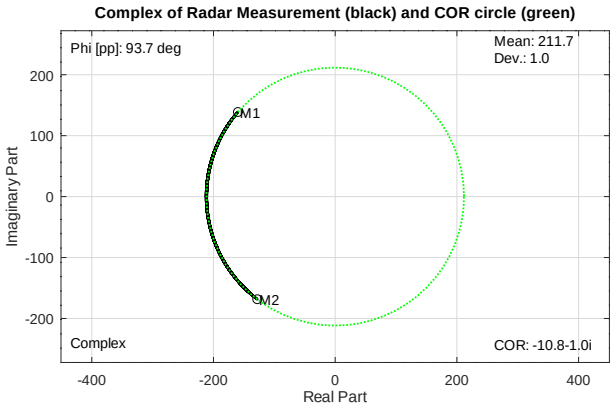


Figure 9: Complex signal of radar measurement (black) of realistic vital sign micro movement using the measurement system and COR circle (green).

Figure 10 shows the prepared signal and the recorded radar phase curve. When comparing both curves you can see, that the measured radar phase fits the prepared signal very good. Only the short peaks are decreased. These small differences between both curves are related to the low pass filter inside the amplifier that controls the loudspeaker mentioned in chapter 3.2. To allow higher frequency components, we have to increase the low pass cutoff frequency.

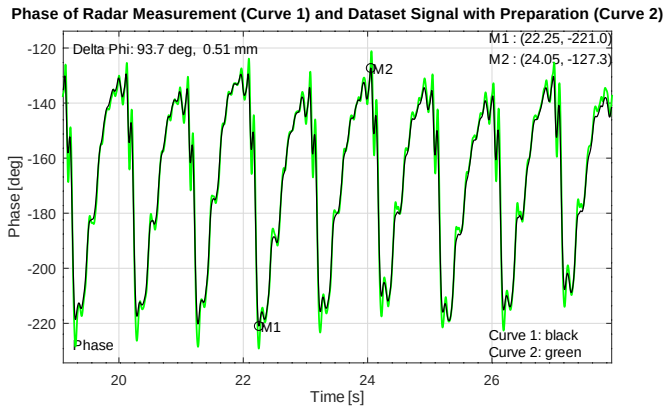


Figure 10: Schellenberger dataset signal after preparation (green) and the related phase curve of the radar measurement using our measurement system (black).

5 Investigation of Dynamic Clutter

All objects in the field of view of the radar sensor have an impact on the measured data. In real measurement scenarios, multiple objects are in the field of view and not all of them are static. Figure 11 displays the complex target signal and phase signal of a measurement with a second moving object close to the target object. A complex offset reduction as mentioned in chapter 2.2 is not able to reduce the clutter that is changing dynamically over time. This is called dynamic clutter. To investigate the effect of dynamic clutter, we created the following scenario with the hardware described in 3.

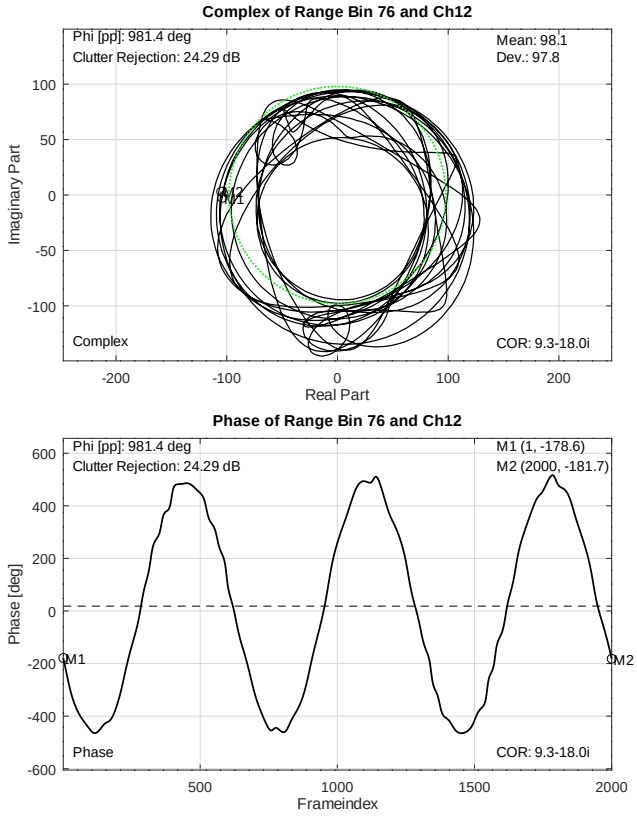


Figure 11: Complex and phase plot from dynamic clutter measurement in range bin$_{ZP}$ 76 without beamforming using the measurement scenario explained below.

5.1 Measurement Setup

To investigate the impact of dynamic clutter, we used two loudspeakers close to each other in our evaluation setup. They are positioned in different distances and different angles in front of the radar. Figure 12 shows a picture of the used measurement setup. The right loudspeaker is the target object, whose movement we want to measure, and is called *target loudspeaker* in the following chapters. It is placed at 150 cm at an angle of -9° relative to the radar position. Angles left to the radar are defined as negative angles. The left loudspeaker is the additional moving object next to the target. It is used to generate the clutter movement and is called *clutter loudspeaker* in the following chapters. The clutter loudspeaker is placed at 130 cm at an angle of -36°.

The measurement setup is chosen to simulate realistic scenarios. The two loudspeakers stand in alike position and distance to each other mimicking a patient lying on a hospital bed and a healthcare professional standing next to the patient during a diagnostic measurement using a radar from above. In this scenario the patient is represented by the target loudspeaker, and the healthcare professional is represented by the clutter loudspeaker. For easier measurement using the loudspeaker setup, the arrangement is shifted to the horizontal plane with the radar standing in front of the loudspeakers instead of being mounted on the ceiling.

Figure 12: Picture of the measurement setup to investigate dynamic clutter with two loudspeakers.

5.2 Movement Signal

Each loudspeaker needs to play a suitable signal so that it is possible to investigate the impact of dynamic clutter. In this paper we present a measurement, in which each loudspeaker plays a sine signal. This ensures that the effect of dynamic clutter is not overlaid with other effects using more complex signals. The oscillation frequencies of the sine signals are chosen to fit usual respiration frequencies. With 0.3 Hz frequency and a movement peak-to-peak amplitude of 5.2 mm the target loudspeaker represents a typical respiration movement for resting humans, while the 0.5 Hz sine with a movement peak-to-peak amplitude of 4.7 mm on the clutter loudspeaker represents a faster respiration.

For the investigation of the dynamic clutter, it is important to compare the signal with dynamic clutter to a reference signal without clutter. Therefore, the clutter movement is not active the whole measurement time. Figure 13 shows the structure of the used audio signal. The full measurement time is 30 s. The signal is designed to have three equal parts. At first, both signals are active and the signal with dynamic clutter can be examined. This part has to be the first part in the signal because the hardware setup including the amplifier with the low pass filter and the loudspeakers need some seconds before the measurement to settle to a stable state for those low frequencies. It is important to consider this settling time while generating the audio signal, but it will not be measured by the radar. In Figure 13 this settling time is not displayed. The middle part of the signal is used to turn off the signal of the clutter loudspeaker. After turning the clutter signal off, the hardware setup also needs a settling time, but it is shorter than the settling time after turning a signal on. The last part, named part C in Figure 13, of the measured signal is used as a reference. Only the 0.3 Hz sine signal on the target loudspeaker is active.

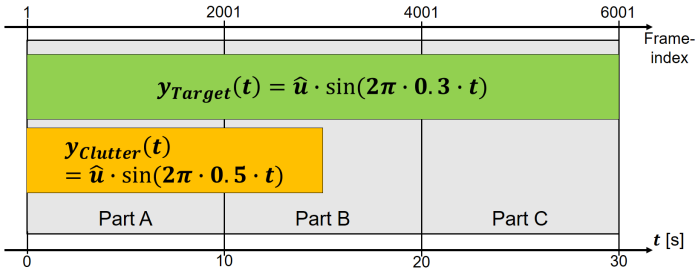


Figure 13: Audio signal during the dynamic clutter measurement.

Each part of the signal is 10 s long and, therefore, three oscillations of the 0.3 Hz target signal and five oscillations of the 0.5 Hz clutter signal fit exactly in each part. This is very important for the subsequent processing of the radar measurement. There, the signal of "Part A" in Figure 13 is compared with the reference signal in "Part C". This is only possible if both signals are in synchronized state. A synchronized state means that the signals are in the exact same state at the beginning of each part. Also the sampling of the radar phase signal has to fit to the synchronization, otherwise a systematic error would occur during the calculations described in the next section. With the radar frame time mentioned in Table 1 of 5 ms, the synchronization points of the audio signal occur at the radar frame indices 1, 2001 and 4001.

5.3 Dynamic Clutter Rejection

The measurements for this investigation are realized with the Radarbook2 using the post processing presented in chapter 2.2. In the dynamic clutter rejection measurements, the COR algorithm is used slightly different than usual to achieve a better performance. For every bin, the COR algorithm finds the complex offset only using the signal "Part C" and uses this complex offset for all three signal parts. The result is that in "Part A" no static clutter is remaining, and the dynamic clutter can be investigated now. The quantification of the impact of dynamic clutter needs an extra step of processing. In (5) the dynamic clutter is defined as additive noise on the reference signal.

$$\Phi_{Clutter} = \Phi - \Phi_{Ref} \tag{5}$$

Equation (5) includes Φ as the signal with dynamic clutter, Φ_{Ref} as reference signal without clutter and $\Phi_{Clutter}$ as the dynamic clutter we want to investigate. Therefore, the dynamic clutter can be calculated as the difference between the measured signals of the different synchronized parts. Φ and Φ_{Ref} are measured radar phase signals that contain the information about the loudspeaker movement as described in (1). After calculating $\Phi_{Clutter}$, the dynamic clutter rejection R_C in dB can be defined as in (6) using the peak to peak value of the reference signal $\Phi_{PP,Ref}$ and the peak to peak value of the clutter curve $\Phi_{PP,Clutter}$.

$$R_C = 20 \cdot log_{10}\left(\frac{\Phi_{PP,Ref}}{\Phi_{PP,Clutter}}\right) \tag{6}$$

In the measurement scenario described in section 5.1, there are two loudspeakers at the distances 130 cm and 150 cm. With the range resolution of the radar mentioned in 2.2, these targets could be measured in range Bin 17 and 20. Therefore, the targets will be measured at Bin_{ZP} 68 and 80 with using zero padding. Nevertheless, it is important to keep in mind that the real range resolution of the radar is lower than the Bin_{ZP} suggests.

The chosen scenario allows to specifically investigate the range bins in between the two moving loudspeakers to indicate the robustness of the evaluated algorithms. The following Table 2 shows the results of the dynamic clutter investigation measurement.

Table 2: Dynamic clutter rejection in different range bins without beamforming.

$\mathbf{Bin_{ZP}}$	$\Phi_{PP,Clutter}$ [°]	**Clutter Rejection [dB]**
80	13	37.36
79	14.4	36.49
78	17.1	34.98
77	24.7	31.76
76	58.5	24.29
75	547.1	4.87

All range bins in which the dynamic clutter rejection in dB is higher than zero are evaluated. The according peak to peak value of the reference signals in all bins is about 958°. For the calculation of the dynamic clutter rejection, the exact value of each range bin was certainly used. The peak to peak value is equivalent to a movement of 5.18 mm of the target loudspeaker diaphragm. This is a typical movement range of the chest wall generated by the respiration, but it is important to note that the dynamic clutter rejection will be lower for signals with lower movement amplitude like the heartbeat movement.

In the range bin corresponding to the target loudspeaker, the dynamic clutter rejection is 37.36 dB, and it drops to 24.29 dB in range bin 76. This means that with a target detection one real range bin too low, the signal quality of the target movement measurement in this detected range bin is greatly reduced. The graphs in Figure 11 clearly visualize the changes due to the dynamic clutter. The problem with visible differences in the signal becomes apparent when considering the visual analysis by healthcare professionals.

5.4 Impact of Beamforming on Dynamic Clutter

In this paper we not only want to investigate the dynamic clutter but also how a digital beamforming can reduce the dynamic clutter for a scenario with multiple radar targets. With the presented measurement system, it is possible to test the performance of different beamforming algorithms concerning their capability to reduce dynamic clutter. In the scope of this article, we will demonstrate this by discussing a beamforming algorithm and its effect on the dynamic clutter.

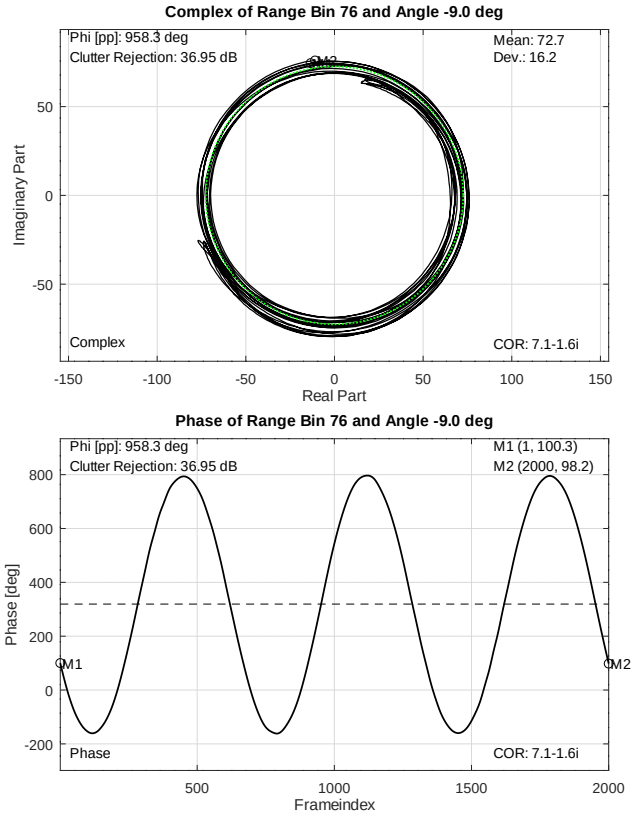


Figure 14: Complex and phase plot from dynamic clutter measurement in range bin_{ZP} 76 with beamforming.

In [22], three different beamforming algorithms are compared, Bartlett, Capon and MUSIC algorithm. Those algorithms are candidates for future studies using our presented measurement system. By now, we implemented a second Fourier transform on all radar RX channels as mentioned in chapter 2.2. As already used in the

range FFT, a zero padding with factor four is used in the angle FFT. Like mentioned before, this does not change the angle resolution of the radar, but adds more points to the spectral curve.

After the angle FFT, the same processing was done on the data as in section 5.3. The angle of the target loudspeaker for all presented results is found by the target bin finder described in chapter 2.2. The results of the processed data are presented in Table 3. Figure 14 shows the radar phase and complex curves with dynamic clutter in range bin 76. In comparison to the curves in Figure 11 this is a great improvement.

Table 3: Dynamic clutter rejection of different range bins with beamforming.

$\mathbf{Bin}_{ZP}$	$\Phi_{PP,Clutter}$ [°]	Clutter Rejection [dB]
80	10.4	39.26
79	10.6	39.15
78	10.9	38.90
77	11.3	38.60
76	13.6	36.95
75	33.2	29.19

Figure 15 condenses all values from the Tables 2 and 3 into two curves. There you can see that the usage of beamforming greatly increases the signal quality in the observed range bins. Even when analyzing the data in range bin 80, where we expect the target signal to be, a slight improvement thanks to digital beamforming is measurable. Further, dealing with movements with smaller micro movement (like the one from the heartbeat), this could be even more important. As mentioned before, the visual representation of the phase signal related to the chest wall micro movement could be important to healthcare professionals in future diagnosis. A comparison of the curves in Figure 11 and 14 distinctly visualizes that the dynamic clutter rejection is more effective when using beamforming. Especially in real scenarios it is important to achieve a high robustness because the perfect adjustment cannot be expected at all times.

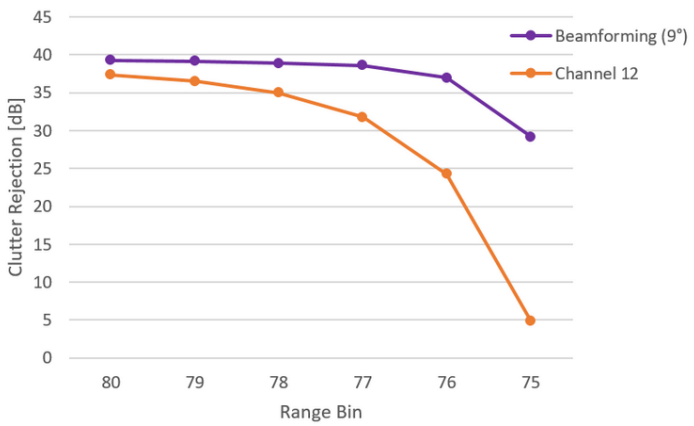


Figure 15: Dynamic clutter rejection with and without beamforming in different range bins.

6 Replication and Recording of realistic Micro Movement

In this chapter, measurements of realistic heartbeat movements are presented. The objective is to survey the effect of dynamic clutter on realistic micro movement with smaller movement amplitude. As already mentioned in 4.2, we used an apnea after exhalation scenario from the GDN0009_3_Apnea measurement of the clinical recorded dataset of Schellenberger [15]. Also, the chapter 4.2 explains the preparation of the dataset data for measurements.

For the measurement presented in this chapter, the prepared heartbeat motion signal is replayed on the target loudspeaker. In addition, a clutter movement is generated by the clutter loudspeaker at the same time. The used clutter signal is a sine with 0.4 Hz frequency and 4.7 mm movement peak-to-peak amplitude.

The recorded radar phase signal is shown in Figure 16. The investigated range bin is 76, which is one real Bin_W apart from the target bin. This constellation was already used in chapter 5. On the recorded data, the signal processing from 2.2 is used without a digital beamforming.

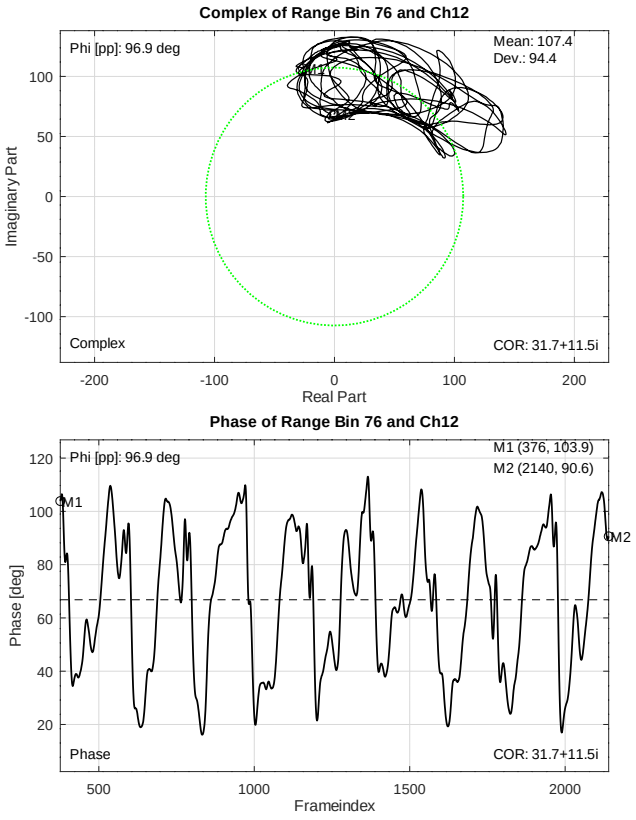


Figure 16: Complex and phase plot of radar measurement using heartbeat movement from dataset without digital beamforming.

In the performance test without any dynamic clutter presented in Figure 10 the heartbeat curve is clearly visible. However, during the measurement with dynamic clutter, the signal is strongly disturbed and the heartbeat movement is hardly visible. To improve the signal quality, the measured data were then processed again, this time with an additional angle FFT as digital beamforming. The results of the

second signal processing using digital beamforming is presented in Figure 17. There, the quality of the curve is significantly higher.

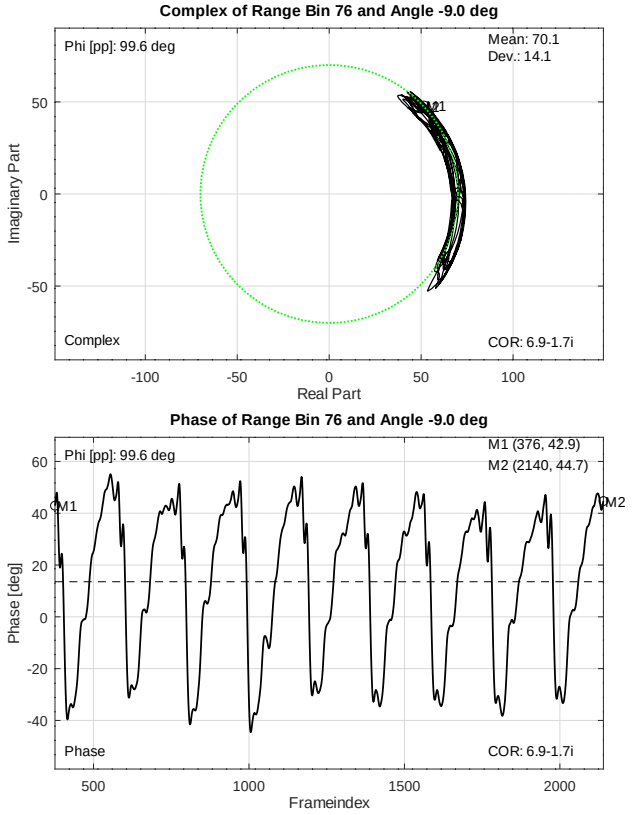


Figure 17: Complex and phase plot of radar measurement using heartbeat movement from dataset with digital beamforming.

The measurements in this chapter confirm once again the relevance of digital beamforming for vital sign sensing scenarios. They are an important supplement to the results of the previous chapter, in which the influence of dynamic clutter was quantitatively evaluated using the clutter rejection. Both measurement methods are steps in the evaluation of different beamforming algorithms.

7 Conclusion

In this paper we presented a measurement system that can be used for evaluating radar algorithms by replication of vital sign micro movement and dynamic clutter. It is an extension of work originally presented in the 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) [1]. In contrast to the previous work, the measurement setup was completely re-designed to be able to replicate realistic micro movement from vital signs and to enable scenarios with two moving objects. Movement data used for the replication was synthesized or taken from recorded movement data. The heartbeat movement from the clinically recorded dataset [15] was successfully reproduced and recorded by the measurement setup.

The ability to evaluate radar algorithms was demonstrated by the investigation of dynamic clutter. It was shown that an angle FFT algorithm increases the dynamic clutter rejection and signal quality in a scenario with dynamic clutter. This increase in signal quality is also clearly visible in the phase signal related to the chest wall micro movement, which can be important to healthcare professionals in future diagnosis.

In future studies, different beamforming algorithms like Capon and MUSIC can be compared using the presented measurement system. It is suitable to evaluate signal processing algorithms for radar applications by replicating real micro movement scenarios. This is crucial in optimizing robust and reliable systems for medical applications. In addition, it is also possible to test radar algorithms implemented in a signal processor inside of the radar sensor at realtime using the measurement system.

References

[1] C. Domnik, M. Meuleners, C. Degen, “Radar Evaluation Setup for the Replication of Chest Wall Movement from Vital Signs,” in 2022 44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2022, doi:10.1109/embc48229.2022.9871203.

[2] V. Chen, F. Li, S.-S. Ho, H. Wechsler, “Micro-doppler effect in radar: phenomenon, model, and simulation study,” IEEE Transactions on Aerospace and Electronic Systems, **42**(1), 2–21, 2006, doi:10.1109/taes.2006.1603402.

[3] Y. Zhang, F. Qi, H. Lv, F. Liang, J. Wang, “Bioradar Technology: Recent Research and Advancements,” IEEE Microwave Magazine, **20**(8), 58–73, 2019, doi:10.1109/mmm.2019.2915491.

[4] J. C. Lin, “Microwave sensing of physiological movement and volume change: A review,” Bioelectromagnetics, **13**(6), 557–565, 1992, doi:10.1002/bem.2250130610.

[5] C. Feng, X. Jiang, M.-G. Jeong, H. Hong, C.-H. Fu, X. Yang, E. Wang, X. Zhu, X. Liu, “Multitarget Vital Signs Measurement With Chest Motion Imaging Based on MIMO Radar,” IEEE Transactions on Microwave Theory and Techniques, **69**(11), 4735–4747, 2021, doi:10.1109/tmtt.2021.3076239.

[6] S. Dong, Y. Zhang, C. Ma, C. Zhu, Z. Gu, Q. Lv, B. Zhang, C. Li, L. Ran, “Doppler Cardiogram: A Remote Detection of Human Heart Activities,” IEEE Transactions on Microwave Theory and Techniques, **68**(3), 1132–1141, 2020, doi:10.1109/tmtt.2019.2948844.

[7] C. Gu, R. Li, H. Zhang, A. Y. C. Fung, C. Torres, S. B. Jiang, C. Li, “Accurate Respiration Measurement Using DC-Coupled Continuous-Wave Radar Sensor for Motion-Adaptive Cancer Radiotherapy,” IEEE Transactions on Biomedical Engineering, **59**(11), 3117–3123, 2012, doi:10.1109/tbme.2012.2206591.

[8] M.-C. Tang, F.-K. Wang, T.-S. Horng, “A single radar-based vital sign monitoring system with resistance to large body motion,” in 2017 IEEE MTT-S International Microwave Symposium (IMS), IEEE, 2017, doi:10.1109/mwsym.2017.8058758.

[9] J.-M. Munoz-Ferreras, Z. Peng, R. Gomez-Garcia, C. Li, “Random body movement mitigation for FMCW-radar-based vital-sign monitoring,” in 2016 IEEE Topical Conference on Biomedical Wireless Technologies, Networks, and Sensing Systems (BioWireleSS), IEEE, 2016, doi:10.1109/biowireless.2016.7445551.

[10] E. Cardillo, C. Li, A. Caddemi, “Vital Sign Detection and Radar Self-Motion Cancellation Through Clutter Identification,” IEEE Transactions on Microwave Theory and Techniques, **69**(3), 1932–1942, 2021, doi:10.1109/tmtt.2021.3049514.

[11] A. Marnach, D. Schmiech, A. R. Diewald, “Verification of Algorithm for an I/Q-Radar System for Breathing Detection in an Incubator,” in 2019 International Conference on Electromagnetics in Advanced Applications (ICEAA), IEEE, 2019, doi:10.1109/iceaa.2019.8879336.

[12] B. Padasdao, E. Shahhaidar, C. Stickley, O. Boric-Lubecke, “Electromagnetic Biosensing of Respiratory Rate,” IEEE Sensors Journal, **13**(11), 4204–4211, 2013, doi:10.1109/jsen.2013.2266253.

[13] B. Padasdao, E. Shahhaidar, O. Boric-Lubecke, "Measuring chest circumference change during respiration with an electromagnetic biosensor," in 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2013, doi:10.1109/embc.2013.6609906.

[14] A. Singh, S. U. Rehman, S. Yongchareon, P. H. J. Chong, "Modelling of Chest Wall Motion for Cardiorespiratory Activity for Radar-Based NCVS Systems," Sensors, **20**(18), 5094, 2020, doi:10.3390/s20185094.

[15] S. Schellenberger, K. Shi, T. Steigleder, A. Malessa, F. Michler, L. Hameyer, N. Neumann, F. Lurz, R. Weigel, C. Ostgathe, A. Koelpin, "A dataset of clinically recorded radar vital signs with synchronised reference sensor signals," 2020, doi:10.6084/M9.FIGSHARE.12186516.V2.

[16] A. Singh, B.-K. Park, O. Boric-Lubecke, I. Mostafanezhad, V. M. Lubecke, "Physiological Doppler Radar Overview," in Doppler Radar Physiological Sensing, 69–94, John Wiley & Sons, Inc, 2016, doi:10.1002/9781119078418.ch4.

[17] J. Liu, Y. Li, C. Li, C. Gu, J.-F. Mao, "Accurate Measurement of Human Vital Signs With Linear FMCW Radars Under Proximity Stationary Clutters," IEEE Transactions on Biomedical Circuits and Systems, **15**(6), 1393–1404, 2021, doi:10.1109/tbcas.2021.3123830.

[18] Y. Wang, A. Ren, M. Zhou, W. Wang, X. Yang, "A Novel Detection and Recognition Method for Continuous Hand Gesture Using FMCW Radar," IEEE Access, **8**, 167264–167275, 2020, doi:10.1109/access.2020.3023187.

[19] A. D. Droitcour, O. Boric-Lubecke, "Physiological Motion and Measurement," in Doppler Radar Physiological Sensing, 39–68, John Wiley & Sons, Inc, 2016, doi:10.1002/9781119078418.ch3.

[20] S. Schellenberger, K. Shi, T. Steigleder, A. Malessa, F. Michler, L. Hameyer, N. Neumann, F. Lurz, R. Weigel, C. Ostgathe, A. Koelpin, "A dataset of clinically recorded radar vital signs with synchronised reference sensor signals," Scientific Data, **7**(1), 2020, doi:10.1038/s41597-020-00629-5.

[21] K. Shi, S. Schellenberger, C. Will, T. Steigleder, F. Michler, J. Fuchs, R. Weigel, C. Ostgathe, A. Koelpin, "A dataset of radar-recorded heart sounds and vital signs including synchronised reference sensor signals," Scientific Data, **7**(1), 2020, doi:10.1038/s41597-020-0390-1.

[22] C. Degen, "On single snapshot direction-of-arrival estimation," in 2017 IEEE International Conference on Wireless for Space and Extreme Environments (WiSEE), IEEE, 2017, doi:10.1109/wisee.2017.8124899.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 2, 30-37 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

A Multiplatform Application for Automatic Recognition of Personality Traits in Learning Environments

Víctor Manuel Bátiz Beltrán*,[1], Ramón Zatarain Cabada[1], María Lucía Barrón Estrada[1], Héctor Manuel Cárdenas López[1], Hugo Jair Escalante[2]

[1]Tecnológico Nacional de México campus Culiacán, Culiacán, Sinaloa, 80220, México

[2]Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla, Puebla, 72840, México

ARTICLE INFO

Article history:
Received: 30 December, 2022
Accepted: 12 February, 2023
Online: 11 March, 2023

Keywords:
Automatic Recognition of Personality
Deep Learning
Web Platform
Standardized Personality Tests
Intelligent Learning Environments

ABSTRACT

The present work shows the development of a data collection platform that allows the researcher to collect new video and voice data sets in Spanish. It also allows the application of a standardized personality test and stores this information to analyze the effectiveness of the automatic personality recognizers concerning the results of a standardized personality test of the same participant. Thus, it has elements to improve the evaluated models. These optimized models can then be integrated into intelligent learning environments to personalize and adapt the content presented to students based on their dominant personality traits. To evaluate the developed platform, an intervention was conducted to apply the standardized personality test and record videos of the participants. The data collected were also used to evaluate three machine learning models for automatic personality recognition.

1. Introduction

This paper is an extension of a discussion paper originally presented in [1]. One of the most widely used and accepted models for determining personality based on written tests are the trait-based models and specifically the Big-Five model. This model is usually represented by the acronym OCEAN where each letter refers to a term that represents each of the five personality traits: Openness to experience, Conscientiousness, Extraversion, Agreeableness and Neuroticism [2].

One of the most relevant efforts regarding the definition of the questions (items, as they are known in the field of psychology) to be used for the Big-Five model, is the one conducted by the International Personality Item Pool (IPIP), which we can consider as a scientific laboratory for the development of advanced measures of personality traits and other individual differences that are in the public domain thanks to its Web site (http://ipip.ori.org/). This site maintains an inventory of thousands of items and hundreds of scales for the measurement of personality traits and is generally based on the studies conducted by Goldberg [3–5].

In recent years, research has been conducted with the aim of implementing automatic personality recognizers through machine learning. These studies have focused mainly on using the Big-Five model to detect apparent personality based on text, voice, or facial features. The main challenge facing these investigations is the difficulty of having a representative dataset, the need to label the images, and the fact that these efforts are typically not in the public domain and therefore it is difficult to reproduce their results [2].

The main contribution of this work is the development of an integrated environment that allows assessing the personality traits of an individual by using a standardized test based on the Big-Five model and allows capturing video interactions in Spanish. Deep learning based automatic recognizers uses these videos and seek to determine the same personality aspects. The above, to be able to evaluate the effectiveness of such automatic recognizers with respect to the standardized test and thus have relevant information to improve the model used by the recognizers that can be integrated into different intelligent learning systems to add new features such as personalized instruction and feedback to students.

This paper is structured in the following order: Section 2 presents related works in the areas of standardized testing and automatic personality recognition; Section 3 presents an analysis of the proposed data collection platform; Section 4 describes the experiments, results, and discussion; and finally, Section 5 presents conclusions and future work.

*Corresponding Author: Víctor Manuel Bátiz Beltrán, Email: victor.bb@culiacan.tecnm.mx

https://dx.doi.org/10.25046/aj080204

2. Related Works

In this section we describe some research works related to the area of standardized tests and automatic personality recognition. These works, although separate efforts, are related to elements of the present research and were considered as the foundation for the development and integration of this project.

2.1. Standardized Personality Tests

Studies have shown that one of the best approaches to personality detection is the Big Five model. Its strength lies in the general acceptance that personality traits, although they may exhibit some changes, remain relatively stable throughout a person's life [6].

In recent years, several studies have been carried out that present adaptations to different languages of the items provided by the IPIP, to evaluate their applicability in different cultures, finding positive results. As an example of the above, we found the study conducted in [7] for the adaptation and contextualization of 100 items of the IPIP repository in the Argentine environment, obtaining satisfactory results in their reliability studies, and on the other hand the adaptation made by the authors in [6] for the application of a reduced version of the IPIP questionnaire in French-speaking participants. This version consisted of 20 items in total, where each of the personality traits was evaluated with 4 items, obtaining as a result the confirmation of the cross-cultural relevance of the personality indicators, of the model of the Big-Five in participants with diverse idiomatic and cultural backgrounds.

2.2. Automatic Recognition of Personality

In the field of automatic recognition, several approaches to apparent personality recognition have been proposed in recent years. Some studies have worked on automatic recognition based on textual information, such as information generated by users on social networks like Facebook, Twitter, and YouTube. Such is the case of the study presented in [8] where diverse approaches such as multivariate regression and univariate approaches such as decision trees and support vector machines are analyzed for automatic recognition.

Other studies have worked on the recognition of apparent personality based on the voice of participants. In [9] the authors propose a system based on a convolutional neural network that evaluates a voice signal and returns values for the five personality traits of the Big-Five model. They conclude that the correlation between different dimensions of a voice signal can help infer personality traits.

Likewise, research has been carried out for the detection of apparent personality based on images extracted from videos of the participants, using various models of neural networks [2]. In the research work conducted in [10], the authors present an apparent personality recognition model based on convolutional neural networks using images extracted from short video clips. They conclude that facial information plays a key role in predicting personality traits.

Renewed interest in the world of artificial intelligence and machine learning, as well as the existence of competencies such as those conducted by ChaLearn Looking at People, have helped the development of various neural network models for apparent personality detection based on first impression [2,11].

3. Data Collection Platform

As we are working with sensitive data from individuals, it is important to emphasize that participants are notified that the data collected from the standardized test and the videos are used internally for the experiments by the team of researchers. Therefore, no information that reveals or compromises their identity is published without prior consent. For this purpose, the platform always requests their registration to have their contact information.

3.1. Architecture

Data collection presents a challenge, as we must establish a system for storing and consulting the information. Nowadays, thanks to the advancement of technology, we can develop environments that make use of the Internet and thus be able to reach more people regardless of their location or the device they use to connect to the Internet. Therefore, we chose to develop a cloud platform that would work on any device and that would allow us to store the information in a repository located on the Internet to facilitate the study of the data.

We chose to use a layered architectural model on a client/server architecture. Three layers are defined: presentation, application logic and data. In Figure 1 we can appreciate the logical view of the platform.

The presentation layer shows users a graphical interface that offers them the option of registering with the system or logging in. The presentation layer uses the application logic layer to execute the operations supported by the system. The application logic layer, in turn, connects to the data layer which contains the database that stores the information regarding user identification, IPIP test and automatic recognizer results and in this layer, we have the file storage whose function is to store video files (including audio) related to the users.

The developed application is a cross-platform system hosted on the Internet cloud, using the free hosting offered by Google Firebase as part of its services.

For the development of the user interface, we decided to use React (https://es.reactjs.org/) and for the logic part of the application and storage of information and files, we opted for the free services of Google Firebase. In addition to availability, another advantage of using these services is the support in privacy and data security offered by Firebase, since it is certified in the main security and privacy standards (https://firebase.google.com/support/privacy).

In Figure 2 we can appreciate a partial view of the application of the standardized personality test in the data collection platform. Answering the test is the first step the user must take after logging into the system. The resulting scores are used as the actual values of the participant's personality traits.

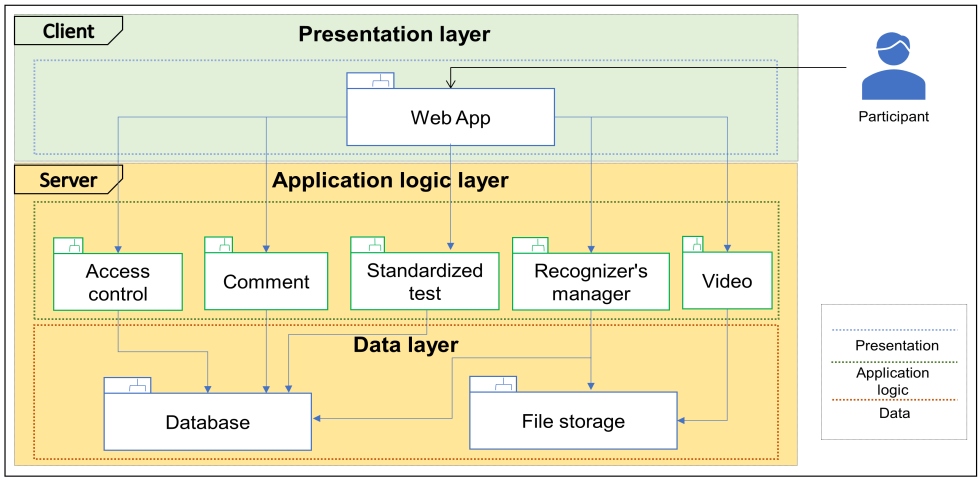


Figure 1: Logical view of the platform.

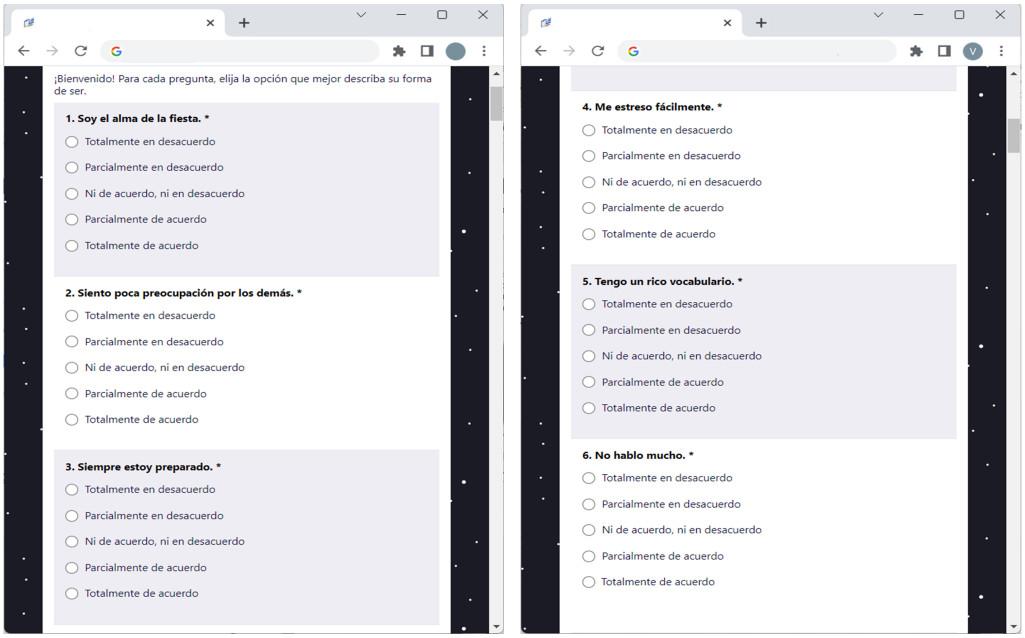


Figure 2: Partial view of standardized personality test.

In Figure 3 we can see the interface where participants record a video talking about a topic of their choice.

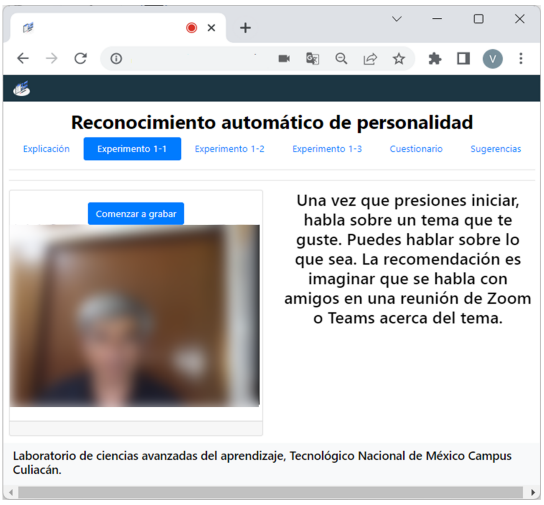


Figure 3: Video Recording.

3.2. Automatic Recognition of Personality

As an example of the use of our platform, we have decided to evaluate three automatic recognition models based on deep learning, using convolutional neural networks (CNN), and Long Short-Term Memory (LSTM) neural networks. These automatic recognizers were trained using ChaLearn's personality dataset which contains 10,000 videos with an approximate duration of 15 seconds each one [12]. The following is a description of the architectures of these automatic recognizers.

3.2.1. Discrete convolutional residual neural network (TNMCUL1)

The architecture used by this evaluated automatic recognizer (see Figure 4) is a discrete convolutional residual neural network (ResNet). We have our input layer of size 500x500x3.

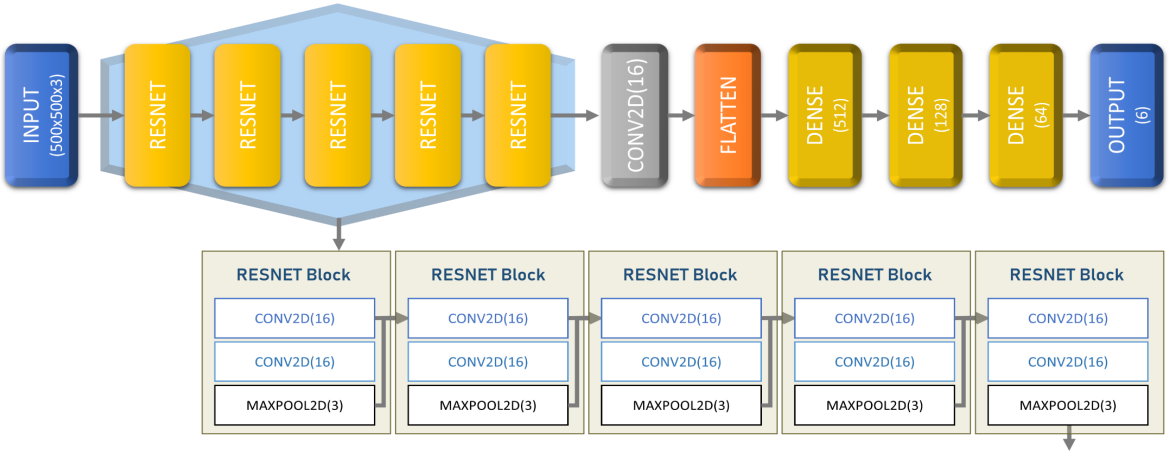


Figure 4: Discrete convolutional neural network topology.

Then we created 5 ResNet modules, where each module contains a two-dimensional convolutional layer (Conv2D) connected to another Conv2D layer, with 16 filters in 3 dimensions in both, and finally connected to a two-dimensional maximum grouping layer (Maxpool2D) with a stride of 3. A concatenation layer was used to add the characteristics of each ResNet block. Then a single Conv2D layer was used with 16 filters in 3 dimensions. We flattened the features vector and connected 4 densely connected layers, each with 512, 128, 64 and 6 neurons respectively. All layers used ReLU activation except the last one that used sigmoid activation for regression. Loss was measured using the mean absolute error (MAE).

3.2.2. Continuous convolutional residual neural network (TNMCUL2)

The second considered architecture is a time-distributed convolutional neural networks. We used sets of 30 images taken from each video creating vectors of dimensions 30x500x500x3. Then, we proceeded to use 5 ResNet modules to process the data from the image vectors and use the attribute vectors created by convolution as input for a 3 layered neural network for feature classification with a 6-neuron output layer for regression.

The full architecture is detailed in this way: first, we used an input layer of size 30x500x500x3. Then, we created a time distributed layer to wrap 5 ResNet modules with the same configuration, as explained on TNMCUL1. Finally, we used a global average pooling 3D layer, a flattening layer for the feature vector, and 4 densely connected layers, with 512, 256, 128 and 6 neurons, respectively. The hidden layers used ReLU activation, and the output layer used sigmoid activation for regression.

This architecture was designed to explore the use of time distributed layers in our study to add another dimension for the previous architecture. We aimed to use the full video image data on a full array, different to our previous architecture that only used individual frames. Figure 5 shows the topology of this neural network.

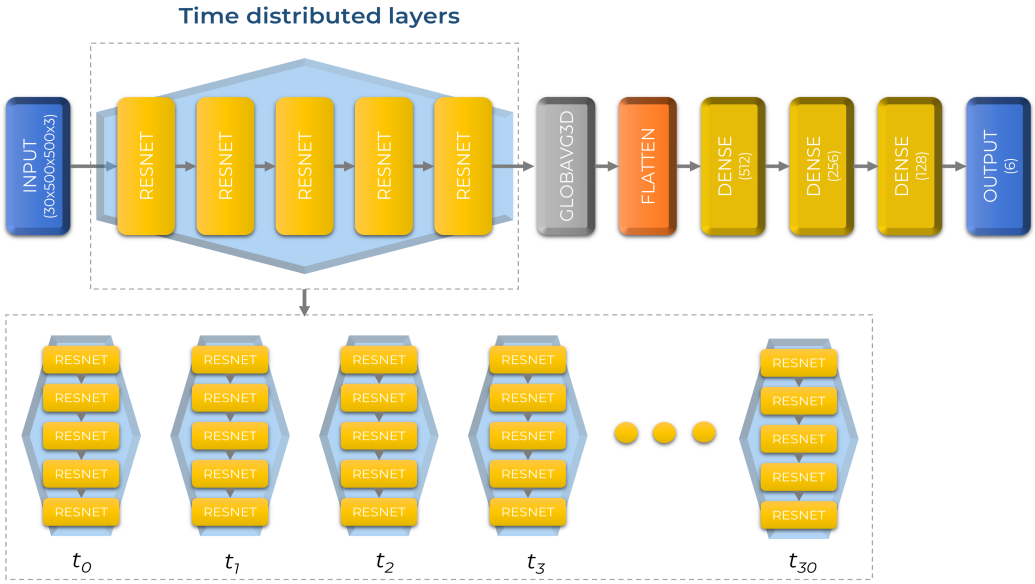


Figure 5: Continuous convolutional residual neural network topology.

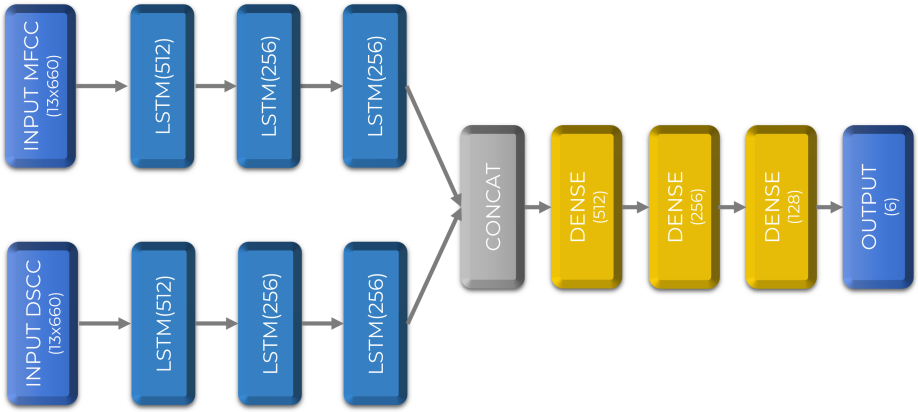


Figure 6: Speech MFCC + DSCC LSTM neural network topology.

3.2.3. Speech MFCC + DSCC LSTM neural network (TNMCUL3)

The third architecture used Long Short-Term Memory (LSTM) neural networks for audio processing. We used an audio file MFCC and DSCC vectors per video, using dual input layers connected to LSTM layers. Then, we used our two feature vectors for feature concatenation and used a 4 layered neural network for feature classification with a 6-neuron output layer for regression.

The full architecture is structured as follows: we used two twin feature extraction LSTM. First, we created input layers for MFCC and DSCC both used as input layers with a dimension of 13x660, connected to 3 LSTM layers of 512, 256 and 256 neurons, respectively. We then used a concatenation layer for the feature fusion and added 4 densely connected layers with 512, 256, 128 and 6 neurons, respectively. The hidden layers used ReLU and the output layer sigmoid activation for the final regression.

This architecture was designed to explore the use of LSTM layers in our study to explore a different modality than the ones used before. We aimed to use only audio data from the videos. Figure 6 shows the topology of this neural network.

Table 1 shows the accuracy results of the three models used (called TNMCUL1, TNMCUL2, and TNMCUL3) and its comparison against state-of-the-art approaches, included in the publications of the best participants in the apparent personality recognition contests based on First Impressions of ChaLearn [11,13]. TNMCUL2 and TNMCUL1 obtained an accuracy of 0.942215 and 0.936158 respectively, slightly surpassing the other models. TNMCUL3 obtained an accuracy of 0.864853, below the rest of the models.

Table 1: Comparison between our models and other state-of-the-art approaches (prepared with our own data and results in [13]).

Name	Technique	Accuracy
TNMCUL2	**CNN Continuos**	**0.942215**
TNMCUL1	**CNN Discrete**	**0.936158**
NJU-LAMDA	Deep Multi-Modal Regression	0.912968
evolgen	Multi-modal LSTM Neural Network with Randomized Training	0.912063
DCC	Multi-modal Deep ResNet 2D kernels	0.910933
Ucas	AlexNET, VGG, ResNet with HOG3D, LBP-TOP	0.909824
BU-NKU	Deep feature extraction with regularized regression and feature level fusion	0.909387
Pandora	Multi-modal deep feature extraction single frame and late fusion	0.906275
Pilab	Speech features 1000 forest random trees regression	0.893602
Kaizoku	Multi-modal parallel CNN	0.882571
TNMCUL3	**LSTM**	**0.864853**

3.3. Standardized Personality Test

For the standardized test we used a 50-item IPIP representation of the markers mentioned by Goldberg for the factorial structure of the Big-Five model [3]. Each of the five personality traits is evaluated by means of 10 items, which in turn are rated by the participant on a 5-element Likert scale (strongly disagree; partially disagree; neither in agreement, nor in disagreement; partially agree and fully agree) based on participant level of agreement or disagreement with respect to each statement displayed. Each option has a value of 1 to 5 points, so 50 is the maximum score per trait. In the end, we convert the score obtained to a value between 0 and 1. This information is stored in the cloud repository and registered to which user it belongs. These values are used to compare them against the results of automatic recognizers.

3.4. Workflow for Data Collection

Figure 7 shows the workflow used for data collection: as a first step, the participant must register on the platform and log in. Once inside the platform, the participant must answer the standardized 50-item IPIP test. Next, the participant must record videos with an approximate duration of one minute each.

The platform stores the recorded videos in our cloud repository. Subsequently, the collected videos are processed and evaluated using the automatic recognizers and the generated information is stored in the cloud repository, linking the corresponding data to each user.

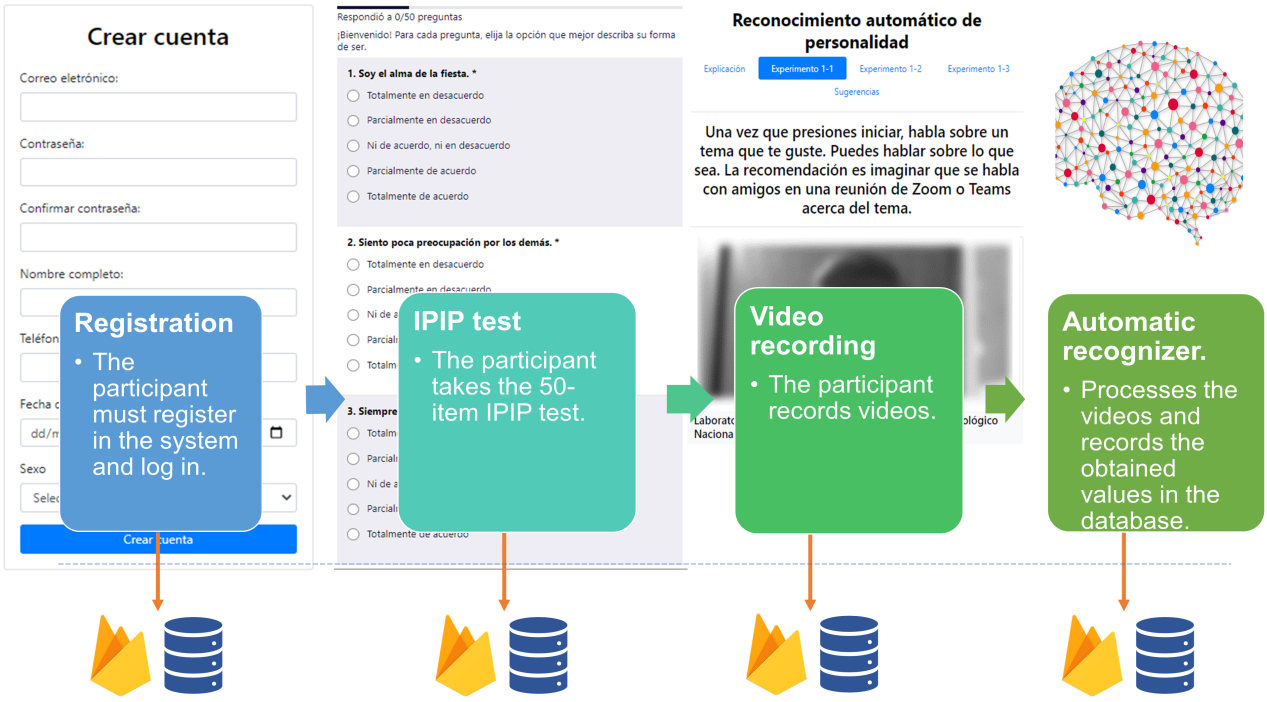


Figure 7: Workflow for data collection.

4. Results and Discussion

In this section we present the initial experiment, the tests, and the results obtained.

4.1. Data Collection Details

Each participant was invited to answer the IPIP test and then record, for one minute, a video where they were asked to speak freely about any topic. Each video is used to extract the images and audios that feed the automatic recognizers and the results are stored directly in the database.

Table 2: Descriptive Statistics of the IPIP Tests.

Trait	Participants	Mean	Standard deviation
Openness	32	0.7344	0.1450
Conscientiousness	32	0.6844	0.1568
Extraversion	32	0.5969	0.1750
Agreeableness	32	0.7906	0.1376
Neuroticism	32	0.5656	0.2598

4.2. Intervention Results

Thirty-two individuals participated in the intervention with the IPIP test, of whom 15 were male and 17 were female. All participants ranged from 23 to 44 years of age. Table 2 shows the descriptive statistics of the data collected for each of the personality traits. It can be observed that the personality traits with the highest mean value were agreeableness with a mean of 0.7906 and openness with a mean of 0.7344. Both traits also presented the least variation with standard deviations of 0.1376 and 0.1450, respectively. The factor with the lowest mean value was neuroticism.

For the evaluation of the selected automatic recognizers of apparent personality, 84 videos were collected. The videos were the product of the intervention of 21 participants (11 of the original participants did not record a video), of which 13 are male and 8 are female. The age range of the participants is between 23 and 40 years old. Table 3 shows the mean absolute error (MAE) values obtained by comparing each value of the personality traits predicted by the automatic apparent personality recognizers against the corresponding value for the participant based on the IPIP test.

Analyzing the results, it was possible to detect that the mean absolute error (MAE) was lower in the extraversion factor and higher in agreeableness. However, in all personality traits the value is too high, so it is not possible to consider that the automatic recognition models evaluated have made an adequate prediction. An interesting aspect is that TNMCUL3 scores better in 4 of the 5 personality traits. TNMCUL2 scores better in Extraversion.

Table 3: Mean Absolute Error (MAE) of each Personality Trait.

Model	Videos	Technique	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
TNMCUL1	84	CNN Discrete	0.2683	0.2684	0.1941	0.3698	0.2806
TNMCUL2	84	CNN Continuous	0.2483	0.2200	0.1786	0.2200	0.2427
TNMCUL3	84	LSTM	0.2262	0.2150	0.2087	0.2150	0.2426

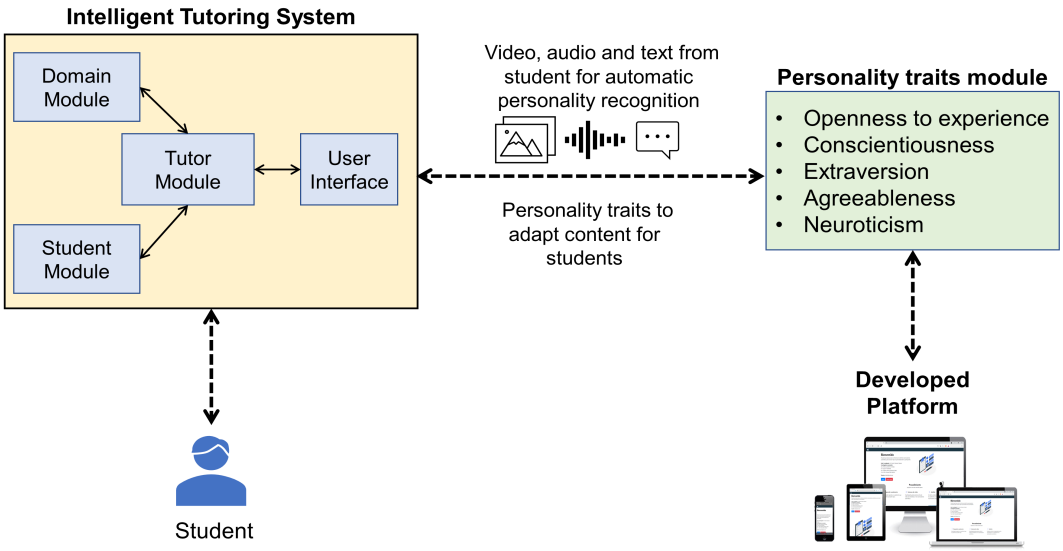


Figure 8: Platform integration with Learning and Tutoring Systems.

4.3. Personality Recognition for Intelligent Learning Environments

Our proposal is to use the information on personality traits and videos collected with the help of the developed platform to evaluate and optimize automatic personality recognition models that can be integrated into intelligent learning environments. The use of an automatic personality recognition model in an intelligent learning environment or tutoring system would allow exploring the idea of presenting adaptive content in real time to the learner based on their dominant personality traits with the goal of achieving the greatest possible impact on learners during their cognitive process.

In Figure 8 we show the proposal to combine an intelligent tutoring system with a personality traits module that makes use of the bank of automatic personality recognizers optimized with our platform.

The learning or tutoring systems communicate with the personality traits module and send it video, image, audio, or text information of the learner which will be used as input to the automatic recognizers. The personality traits module returns as output the presence or absence of the student's Big-Five personality traits. This information can be used by the intelligent tutoring system to make decisions about the content presented to the student.

5. Conclusions and Future Work

The developed platform allows quite a simple and applicable data collection through any device with Internet access from any location and supports the immediate availability of the collected data for analysis.

We have added as a secondary contribution, the evaluation of three automatic recognition models to review the functionality of the platform. In this first exercise, we have found that the evaluated recognizers present a gap in the results with respect to the IPIP test.

The construction of a dataset of Spanish language videos and personality test results is also considered a relevant contribution that can serve as a starting point for future studies.

Additionally, we presented a proposal to use our platform for improving automatic recognizers that could be integrated into tools such as intelligent learning environments or tutoring systems to personalize instruction and feedback based on the detected personality of the participant.

As future work, it is proposed to continue the improvement of the assessed recognizers using the collected dataset and the results of the IPIP tests and hyperparameter optimization techniques. It is suggested to contemplate the evaluation of automatic recognizers of apparent personality based on text to corroborate if the results are like those of the standardized test and, failing that, to work on the retraining of these models, taking advantage of the dataset that is being formed with videos in Spanish.

Another approach that can be addressed is the use of classification algorithms to determine the presence or absence of each personality trait.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors want to thank their institutions for all the support on this research project.

References

[1] V.M. Bátiz Beltrán, R. Zatarain Cabada, M.L. Barrón Estrada, H.M. Cárdenas López, H.J. Escalante, "A multiplatform application for automatic recognition of personality traits for Learning Environments," in 2022 International Conference on Advanced Learning Technologies (ICALT), 49–50, 2022, doi:10.1109/ICALT55010.2022.00022.

[2] J.C.S. Jacques Junior, Y. Gucluturk, M. Perez, U. Guclu, C. Andujar, X. Baro, H.J. Escalante, I. Guyon, M.A.J. van Gerven, R. van Lier, S. Escalera, "First Impressions: A Survey on Vision-Based Apparent Personality Trait Analysis," IEEE Transactions on Affective Computing, **13**(1), 75–95, 2022, doi:10.1109/TAFFC.2019.2930058.

[3] L.R. Goldberg, "The development of markers for the Big-Five factor structure.," Psychological Assessment, **4**(1), 26–42, 1992, doi:10.1037/1040-3590.4.1.26.

[4] L.R. Goldberg, "A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models," Personality Psychology in Europe, **7**(1), 7–28, 1999.

[5] L.R. Goldberg, J.A. Johnson, H.W. Eber, R. Hogan, M.C. Ashton, C.R. Cloninger, H.G. Gough, "The international personality item pool and the future of public-domain personality measures," Journal of Research in Personality, **40**(1), 84–96, 2006, doi:10.1016/j.jrp.2005.08.007.

[6] O. Laverdière, D. Gamache, A.J.S. Morin, L. Diguer, "French adaptation of the Mini-IPIP: A short measure of the Big Five," European Review of Applied Psychology, **70**(3), 100512, 2020, doi:10.1016/J.ERAP.2019.100512.

[7] M. Gross, M. Cupani, "Adaptation of the 100 IPIP items measuring the big five factors," Revista Mexicana de Psicología, **33**, 17–29, 2016.

[8] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M.-F. Moens, M. de Cock, "Computational personality recognition in social media," User Modeling and User-Adapted Interaction, **26**(2–3), 109–142, 2016, doi:10.1007/s11257-016-9171-0.

[9] J. Yu, K. Markov, A. Karpov, Speaking Style Based Apparent Personality Recognition, 540–548, 2019, doi:10.1007/978-3-030-26061-3_55.

[10] C. Ventura, D. Masip, A. Lapedriza, "Interpreting cnn models for apparent personality trait regression," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 55–63, 2017.

[11] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H.J. Escalante, S. Escalera, ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results, Springer International Publishing, Cham: 400–418, 2016, doi:10.1007/978-3-319-49409-8_32.

[12] H.J. Escalante, H. Kaya, A.A. Salah, S. Escalera, Y. Gucluturk, U. Guclu, X. Baro, I. Guyon, J.C.S.J. Junior, M. Madadi, S. Ayache, E. Viegas, F. Gurpnar, A.S. Wicaksana, C.C.S. Liem, M.A.J. van Gerven, R. van Lier, "Modeling, Recognizing, and Explaining Apparent Personality From Videos," IEEE Transactions on Affective Computing, **13**(2), 894–911, 2022, doi:10.1109/TAFFC.2020.2973984.

[13] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, A. Mittal, Bi-modal First Impressions Recognition Using Temporally Ordered Deep Audio and Stochastic Visual Features, Springer International Publishing, Cham: 337–348, 2016, doi:10.1007/978-3-319-49409-8_27.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 1, 30-37 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Analysis and Trend Estimation of Rainfall and Seasonality Index for Marathwada Region

Himanshu Bana*, Rahul Dev Garg

Geomatics Engineering, Indian Institute of Technolog,y Roorkee-247667, India

ARTICLE INFO

Article history:
Received: 15 September, 2022
Accepted: 23 December, 2022
Online: 24 January, 2023

Keywords:
Seasonality Index
trend estimation
Marathwada
Temporal Analysis

ABSTRACT

Droughts are undesirable and highly unwanted form of disasters. It is essential to analyse the cause of such extreme events and act accordingly to pave the way for a sustainable future. The present research work conducts a seasonality and trend analysis of rainfall over the eight districts of Marathwada region. The study is carried out for the last 39 years ranging from 1980 to 2018. The rainfall data pertaining to pre-monsoon season, monsoon season (Kharif), and annual average have been analysed. The trend has been estimated using Sen's slope estimation process along with Mann-Kendal test. It was determined that the all the eight districts of the region show a negative trend in the annual rainfall received. Nanded district showed the largest negative trend in the annual rainfall. Out of eight districts seven districts of the region show a decline in rainfall during the monsoon season. The district of Nanded showed largest decline in the rainfall received during monsoon season. The research work presents the discussion on possible causes of such trends estimated. The research creates a robust foundation of advanced computation techniques for prediction of droughts.

1. Introduction

Melting of glaciers, frequent droughts, and increase in regional temperature are some of the climatic changes that are expected to affect the agricultural scenario of the world [1]. Due to these adverse climate changes, it has been predicted by the intergovernmental panel on climate change (IPCC) that these unfavourable events may lead to scarcity of drinking water and water resources [2]. As per the panel this scarcity of water resource might led to drop in per capita freshwater availability. The effect of this drop would be visible by 2025.

Many researchers in the past have indicated that the change in climatic conditions will bring both scarcities of precipitation and increased intensity of precipitation [3-5]. The increased intensity of precipitation will result in intense flooding, flash flooding, and higher run-off during the monsoon. As the run-off will increase the lesser precipitation will percolate. The ground which is bound to negatively affect the ground water recharge will subsequently result in falling of water table and lower volume of water available for anthropometric activities [6]. Prediction of scarcity of water along with prediction of increased intensity of precipitation indicates that climate change will affect the precipitation at both local and regional scales.

The research presented in [7] indicated in their research work that in Asia-Pacific region the agricultural activities are highly dependent on ground water and monsoon. Thus, depleting water table and less precipitation will adversely affect the cropping system in the region. As the cropping system will be affected it will lay an effect on yield productivity and the net area sown under the principal crops in the region [8].

IPCC has also predicted the probability that the global surface temperature might increase by 5.8ºC by the end of year 2021 [2]. Many researchers in the past have worked upon the sensitivity of the crops to the surface temperature [9,10]. The study in [11] focused on accessing the probable impact of temperature rise on the production of wheat in India. They determined through mathematical modelling that a 1oC rise in temperature is sufficient to drastically reduce the production of wheat.

Study of rainfall variation in India is of special interest to researchers for a long time [12]. The research in [13] focused on studying the variations of climatic parameters in different regions of India even before the subject of climate change was prominent. The special interest in studying the rainfall variations comes from the fact that Indian agriculture is entirely dependent on rainfall. If the states of Punjab and Haryana is not considered no state in India has a proper network of canals and channels that can supply water for irrigation to the farmers. Due to unavailability of irrigation infrastructure the farmers in India are dependent on monsoon. If the monsoon performs poorly in any year, the production of Kharif crops gets affected drastically. It is due to these monsoon dependent characteristics of Indian agriculture; it is called 'Gamble on Rains'.

*Corresponding Author: Himanshu Bana, himanshu_roorkee1@yahoo.com

The research in [14] focused on determining the rainfall trend in the North Eastern states of India. They focused on determining the trend present in the North Eastern states because these states suffer due to scarce as well as heavy rainfall [15]. The study in [16] concluded that due to inadequacy of the irrigation system a decreased rainfall results in poor agricultural production while an increased rainfall always poses a certain danger of flooding due to Brahmaputra breaking its banks. Researchers in the past also indicated that monsoon presents a decreasing trend in the states of Chhattisgarh, Jharkhand, and Kerala [17]. In recent years, the monsoon in India is weakened by the El Nino Southern Oscillations (ENSO) especially in the year 2009, 2015, and 2017 [18]. The ENSO negatively affects the monsoon over India. This negative effect causes less than normal rainfall during phases of El Nino.

The Marathwada region of Maharashtra is a drought prone area [19]. Latur and Osmanabad districts of the region are some of the worst affected regions of the country [14,20]. In 2016, numerous full capacity trains only with water wagons were ferried to Latur to meet the water scarcity of the district [10]. The year of 2016 was not the first time a Latur district from Marathwada region, has suffered from severe water shortages in. Latur has faced droughts in 1980s as well in 1990 [21]. However, the event of scarcity of water in April 2016 was an extreme event. The time of the study is selected from 1980 because the region has shown more susceptibility for getting affected by the drought since then. The research work thus, tries to determine and evaluate the trend present in the rainfall received by the districts of Marathwada region through a time-series analysis for the years 1980-2018.

2. Research Method

The region selected for the study is Marathwada as shown in Figure 1. The Marathwada region is a group of districts located in south western region of state of Maharashtra. The region is comprised of districts namely Beed, Latur, Parbhani, Hingoli, Jalna, Aurangabad, Osmananbad, and Nanded. The region lies near to the northern ranges of western Ghant. The location of the study region has been shown in the map. The region was earlier known for its sugarcane and cotton production. However, the region has started to witness increased frequency of below normal rainfall during monsoon which has reduced the sugarcane cultivation in the region. Further, the area becomes area of interest for the study because the district of Latur recently faced one of the worst water crises in history of Independent India [20].

3. Data

The present study is based on data recorded at 8 stations in the Marathwada region. The period of the data is from 1980 to 2018 (Last 39 years). The data was procured from the India Meteorological Department (IMD). The rainfall data used in this research work was recorded in the stations in the form of direct observation and was subjected to standard normal homogeneity test for homogenization. The data contained no missing values. The trend in the rainfall in the selected districts was estimated on an annual, pre-monsoon, and Kharif season (main sowing season in India that begins in June. The season begins with advent of South West Monsoon making a landfall at Kerala). The pre monsoon months were selected as March, April, and May. The Kharif season was selected as June-July-August, and September.

4. Analysis

For the study seasonality index (SI), standard deviation (SD), coefficient of variance (CV), Sen Slope and Mann-Kendall test were utilized. The seasonality index helps in determination of contrast in the rainfall regime. This process is done by utilization of rainfall monthly distribution data. In other words, the seasonality index helps in identification of monthly rainfall variability [22]. The seasonality is a function of mean of monthly rainfall and mean annual rainfall. The seasonality index is computed as:

$$SI=\frac{1}{A}\sum_{n}^{12}\left|X_n-\frac{A}{12}\right| \quad (1)$$

where, X_n is the rainfall calculated for the nth month. A is the total annual rainfall. Theoretical variations in the seasonality index can be from 0 to 1.83. If all the months in a year records equal rainfall, than the SI becomes zero. If all the rainfall occurs in one month than the SI becomes 1.83 [23]. The SI also suggests changes in rainfall pattern [24]. The rainfall regimes associated with the different values of the seasonality index has been shown in the table 1.

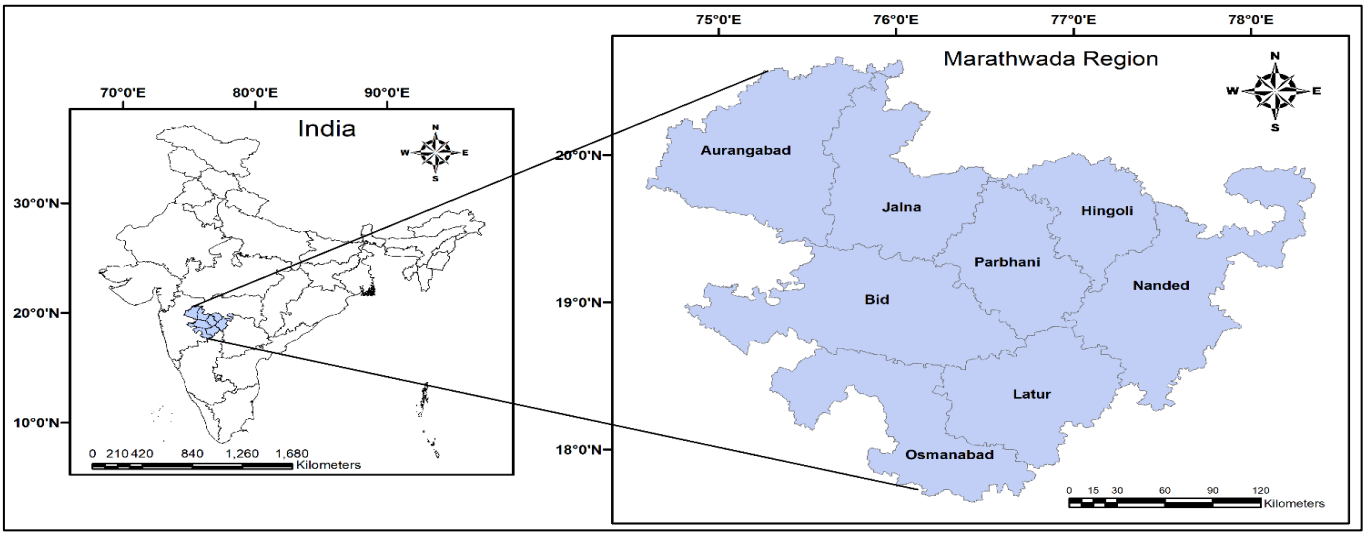


Figure 1: Study Area: Marathwada region

Table 1: Rainfall regimes and associated SI (Walsh & Lawer, 1981)

Regimes	Seasonality Index (SI)
Very Equable	Less than or equals 0.19
Equable, but with a definite wetter season	0.20-0.39
Rather Seasonal with short drier season	0.40-0.59
Seasonal	0.60-0.79
Marked Seasonal with long drier season	0.80-0.99
Mostly rain in 3 months or less	1.00-1.19
Extreme, almost rain in 1 to 3 months	Greater than or equals 1.20

Along with the identification of rainfall regime, SI indicates towards soil and vegetation characteristics along with hydrological stress in the region. Rainfall trend analysis can be performed by using different available parametric and non-parametric methods [25]. These analyses are in general meant to analyse the trend present in long term dataset. However, these techniques are also used to determine trend in short-term data series [26]. These short-term data series can be truncated to ten data points. The restriction associated with the use of parametric test in the trend determination is that the data points in the time series should follow a distribution. Non-parametric tests do not pose such restrictions and are minimally affected by any outliers present in the dataset. In the present work the trend present in the rainfall data was analysed using Mann-Kendal test and the Sen-slope estimates. The slope estimates were determined using Sen-slope estimation process. The Mann-Kendall tests are widely utilized for determination of trends which are monotonous in nature in the non-cyclic data sets [27]. The wide popularity in determining the trend present in the rainfall data using Mann-Kendall test is due to because Mann-Kendal test does not require a particular distribution. Another important characteristic of Mann-Kendal test is that it is least sensitive towards in homogeneity present in the time series [28]. The time series is thus assumed to obey the following model.

$$xi = f(ti) + et \quad (2)$$

where, f(t)is the monotonic decreasing or increasing function of time; the residual is represented by et. It is also assumed further that the variance of the distribution is constant in time. Further, it is also assumed that the autocorrelation in the data set is zero [29].

The estimate of slope which is denoted by Q is calculated by the following process [30].

Firstly, the slope between all pairs of data values is determined. The procedure for the same is as follows.

$$Q_i = \frac{x_j - x_k}{j-k} \text{ fori=1,2,3, … k} \quad (3)$$

where, x_j and x_k are the data values at time j and k (j>k).

In time series that contains n values the N would be determined using

$$N = n(n-1)/2 \quad (4)$$

Where, N is the no of iterations and T is total observations.

The N values are arranged in ascending order before the application of T in the equation presented below. Estimator of Sen's slope is representative of the median of these N values of Q_i

The Q is thus given by,

$$\text{If N is odd, } Q = T_{(N+1)/2} \quad (5)$$

If N is even, $Q = ½ (T_{N/2} + T_{N+2/2})$

Normal distribution is applied for determination of two-sided confidence interval related to the estimation of the slope. In the dataset of the time series the downward or the decreasing trend is indicated by the negative value of Q, while an upward or increasing trend is determined by positive value of Q. The Mann-Kendal test null hypothesis assures that the data series had normal distribution. A significance level of 0.001 has been used for the testing-module. These characteristics correlate to the fact that a probability of 0.01% exists for the randomness in the time series data. Similarly, if the significance level is kept as 0.05 then it is assumed that there exists a 5% of probability that the values of the time series are from random distribution. The Mann-Kendall test statistics S is determined using the formula [31].

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sgn(x_j - x_k) \quad (6)$$

$$sgn(x_j - x_k) = \begin{cases} 1 \; if \; x_j - x_k > 0 \\ 0 \; if \; x_j - x_k = 0 \\ -1 \; if \; x_j - x_k < 0 \end{cases} \quad (7)$$

The number of data points in the time series under consideration is denoted by n and x_k, x_j are the rainfall sum in the year j and k respectively. Here, the j^{th} value is greater than the k^{th} value. The present work fixes the significance level as 0.05 for the test. Since the number of data points would be greater than10 the distribution of S was approximated using a normal distribution. The estimation of trend line was done using linear regression method [32]. In this research work slopes that are of significant nature at significance level of 5% has been marked with*. Slopes that are not of significant nature are not marked with any sign in the result tables.

5. Results

The results of the present study are as follows:

5.1. Seasonality Index

The results of the seasonality index (SI) have been presented in the table 2. The SI reveals that Beed, Latur and Osmanabad face frequent long drier season (SI ranging between 0.8 - 0.99, are marked by red colour cells). In these districts the long drier season occurred for 11, 13 and 9 times respectively in 39 years. Parbhani, Hingoli, Jalna, Aurangabad, and Nanded also face long drier season but in these districts the occurrence of long drier season is less as compared to the rest of three districts. The district of Nanded faced least number of long drier seasons in last 39 years. The SI in the Nanded district for most of the years is greater than 1 which indicates that the district receives rains in

almost three months or less. The district of Latur faced two consecutive long drier seasons in last 39 years. The first occurred from 1985 to 1987 and the second occurred from 2013 to 2015. Further, the district of Latur faced long drier season in alternate years from 1987 to 1992 and from 2002 to 2006.

Table 2: Seasonality Index for the last 39 years for Districts of Marathwada Region

Year	Beed	Parbhani	Hingoli	Latur	Jalna	Aurangabad	Osmanabad	Nanded
1980	1.21	1.07	1.3	1.19	1.25	1.24	1.25	1.24
1981	0.85	1.15	1.03	1.02	1.24	0.9	0.94	1.16
1982	1.05	1.06	0.94	1.04	0.9	0.92	0.95	1.05
1983	1.17	1.04	1.19	1.15	0.92	1.24	1.12	1.17
1984	1.11	1.04	1.11	1.07	1.24	1.1	1.11	1.09
1985	1.09	1.09	1.07	0.97	1.1	1.15	1.05	1.03
1986	1.1	1.28	1.14	0.87	1.15	1.22	1.11	1.07
1987	0.99	1.18	1.12	0.92	1.22	1.05	0.94	1.03
1988	1.26	0.98	1.24	1.27	1.05	1.09	1.19	1.16
1989	1.14	1.39	1.2	1.11	1.09	1.2	1.08	1.22
1990	0.98	1.06	1.03	0.97	1.2	0.95	0.92	0.97
1991	1.27	0.95	1.38	1.12	0.95	1.33	1.09	1.25
1992	1.14	1.03	1.11	0.97	1.33	1.15	1.01	1.13
1993	0.95	0.86	1	1.04	1.15	1.01	1	1.05
1994	0.88	1.24	1.16	1.05	1.01	0.92	1.02	1.11
1995	0.93	0.78	0.94	0.92	0.92	1.1	0.9	1.02
1996	1.17	1.08	1.2	1.14	1.1	1.12	1.1	1.16
1997	0.75	1.13	0.91	0.71	1.12	0.87	0.76	0.78
1998	1.1	1.31	1.04	1.06	0.87	1.07	1.12	1.08
1999	1.05	1.2	1.14	0.95	1.07	1.09	1	1.13
2000	1.2	1.17	1.37	1.13	1.09	1.12	1.11	1.2
2001	1.2	1.23	1.23	1.14	1.12	1.13	1.17	1.2
2002	1.14	1	1.26	0.94	1.13	1.17	1	1.25
2003	1.06	1.13	1.25	1.18	1.17	1.04	1.01	1.18
2004	0.89	1.13	1.07	0.89	1.04	1.12	0.83	1.01
2005	1.16	1.31	1.12	1.09	1.12	1.2	1.16	1.11
2006	0.95	1.21	1.17	0.96	1.2	1.11	1	1.19
2007	1.28	0.91	1.31	1.26	1.11	1.22	1.15	1.23
2008	1.05	1.14	1.24	1.13	1.22	1.24	1.14	1.22
2009	0.84	1.23	0.89	1.02	1.24	1.05	0.92	1.05
2010	1.05	1.17	1.12	1	1.05	0.94	1.01	1.07
2011	1.18	1.02	1.28	1.3	0.94	1.25	1.15	1.26
2012	1.15	0.98	1.17	1.16	1.25	1.17	1.13	1.15
2013	1.01	0.99	1.07	0.96	1.17	1.09	1.01	1.01
2014	0.9	1.13	1.1	0.85	1.09	0.91	0.79	1.08
2015	0.87	1.16	1.01	0.81	0.91	1.05	0.82	0.87
2016	1.12	1.27	1.12	1.07	1.05	1.2	1.09	1.09
2017	1.14	1.2	1.15	1.1	1.2	1.15	1.1	1.1
2018	1.18	1.11	1.33	1.05	1.15	1.23	0.95	1.19

5.2 Rainfall Trend

The trend in the annual rainfall for the various district of Marathwada regions has been shown in the Table 3.

Table 3: Annual Rainfall Trend Analysis for the last 39 Years (1980-2018)

District	Mean (mm)	Standard Deviation (mm)	Sen's Slope (mm/year)	Mann Kendall (mm/year)
Beed	745.33	187.45	-1.962	-0.75
Parbhani	872.20	246.12	-6.090	-1.62
Hingoli	903.80	250.53	-4.771	-1.28
Latur	814.83	210.11	-0.460	-0.12
Jalna	747.47	167.87	-1.703	-0.53
Aurangabad	680.276	156.01	-1.787	-0.70
Osmanabad	737.724	193.05	0.021	0.00
Nanded	985.91	311.88	-5.60*	-1.65

*Statistically significant at 5%, Significance tested using MK Test.

The mean rainfall in the Beed district was observed as 745.33 mm from 1980 to 2018. The annual rainfall in Beed district years shows a negative trend for the last 39 in the annual rainfall. The Sen's slope determined for the annual rainfall received by the district in the last 39 years is -1.962 mm/Year. The mean rainfall in the Parbhani district was observed as 872.20 mm from 1980 to 2018. The annual rainfall of Parbhani for the

last 39 years shows a negative trend. The Sen's slope determined for the annual rainfall received by the district in the last 39 years is -6.090 mm/Year. The mean rainfall in the country's one of most severely and frequently drought affected district Latur was observed as 814.83 mm from 1980 to 2018. The annual rainfall trend in Latur district shows a negative trend for the last 39 years. The Sen's slope determined for the annual rainfall received by the district in the last 39 years is -4.771 mm/Year. The mean rainfall in the Nanded district was observed as 985.91 mm from 1980 to 2018. Out of the 8 selected districts. Nanded received highest mean annual rainfall during the last 39 years. The annual rainfall in Nanded district for the last 39 years shows a negative trend. The Sen's slope determined for the annual rainfall received by the district in the last 39 years is -5.60 mm/Year which was statistically significant negative trend.

The pre-monsoon season holds special importance in the Marathwada region known for its water intensive crops such as sugarcane [33]. Industries in and around the Marathwada region are also known to use water intensively. These industries include the sugar mills and the cotton dying industries. Real-estate activities are also on the rise in the region [34]. Traditional construction approach adopted in the region requires excessive use of water for curing of concrete and the wall plaster. The demand for the sugarcane increases in summer season due to fresh juice stalls and sugar mills boost their production to stock the sugar for the upcoming festive seasons [35].

Farmers use excessive water in the fields during pre-monsoon season in order to increase the brix content of the crop and inter nodal gap in sugarcane [4]. Farmers in the region typically use drenching method of irrigation which also accounts for loss of precious water reserves [36]. Such anthropogenic activities require intense water and may pose a certain threat to the water availability in this already drought susceptible region of the Maharashtra state. From table 4, it is evident that Pre-monsoon rainfall trend is positive for all the 8 districts under consideration. The districts of Jalna and Aurangabad show a positive trend in the pre-monsoon rainfall, but the trend determined is minuscule. Largest positive trend in the pre-monsoon rainfall was observed for the Osmanabad district. The increasing positive trend is beneficial for the water intensive crops sown in the district.

Table 4: Pre-monsoon Rainfall Trend Analysis of last 39 Years (1980-2018)

District	Mean (mm)	Standard Deviation (mm)	Sen's Slope (mm/yr)	Mann Kendall (mm/yr)
Beed	31.86	29.81	0.295	0.69
Parbhani	27.67	28.30	0.192	0.77
Hingoli	22.02	24.33	0.108	0.31
Latur	43.82	33.28	0.393	0.95
Jalna	20.76	24.23	0.017	0.08
Aurangabad	16.55	23.58	0.049	0.55
Osmanabad	36.92	29.72	0.485	1.63
Nanded	29.82	31.34	0.254	1.27

However, the positive trend determined for the pre-monsoon rainfall in the districts was not statistically significant.

The monsoon season is the time to sow the Kharif crops. Kharif crops are known to be water intensive. Farmers in the region have a strong affinity for the sowing water intensive crops in the region. Scarcity of water in the germination and early development stage results in osmotic stress [37]. Such stress exploits the growth of the crop. Farmers are known to sow crops like cotton and groundnut in the early monsoon season. Water intensive crops like sugarcane are sown in the middle of the monsoon season so that the crop can be harvested by March to May [38]. Therefore, good monsoon is essential for the Marathwada region from agricultural perspective. Table 5 depicts that the highest mean rainfall in the monsoon season was received by Nanded district. The district of Aurangabad and Osmanabad received mean rainfall of 557.71 mm and 571.12 mm, respectively. Latur which is one of the drought susceptible districts of the country received a mean rainfall of 656.82 mm in the monsoon season. Sen's slope estimate shows the negative trend in the rainfall received by the districts during monsoon.

The district of Beed shows a negative trend of -0.956 mm/Year. The district of Parbhani shows a negative trend of -2.809 mm/Year. The district of Nanded shows the highest negative trend of -3.996 mm/Year followed by the district of Hingoli which shows a negative trend of -3.154 mm/Year. Only the district of Latur shows a positive trend of 0.458 mm/Year for the rainfall received in monsoon season. The negative trend in seven out of eight districts of region is bad from the perspective of sugarcane producers of the region. Figure 2 shows the trend obtained for the rainfall received by the districts in the monsoon season.

Table 5: Kharif Rainfall Trend Analysis of last 39 Years (1980-2018)

District	Mean (mm)	Standard Deviation (mm)	Sen's Slope (mm/yr)	Mann Kendall (mm/yr)
Beed	609.66	179.318	-0.956	-0.34
Parbhani	744.89	229.24	-2.809	-0.77
Hingoli	792.54	227.45	-3.154	-0.90
Latur	656.82	191.40	0.458	0.19
Jalna	620.59	152.30	-1.385	-0.60
Aurangabad	557.71	131.31	-1.669	-0.68
Osmanabad	571.12	167.68	-0.266	-0.05
Nanded	836.07	274.27	-3.996	-1.23

From the analysis of annual and monsoon rainfall received by the districts it is evident that the scarcity of the rainfall in the region is on the rise. Seven districts of region showed a negative trend in the annual rainfall received while seven out of eight districts of the region showed a negative trend in the rainfall received in the monsoon season.

6. Discussion

The Thar Desert and adjoining areas of central and northern subcontinents heat up during the summers. This creates a void. To fill up the void the air from the Indian ocean rush into the mainland. The air is laden with moisture picked up from the ocean surface. The Himalayas regulates the air-flow and prevent its influx into the central Asia. As the wind rises, precipitation occurs and India receives rainfall [39]. This rainfall season is also known as the Southwest (SW) monsoon. Marathwada region of Maharashtra state receives the SW monsoon. The period of SW monsoon starts from June and

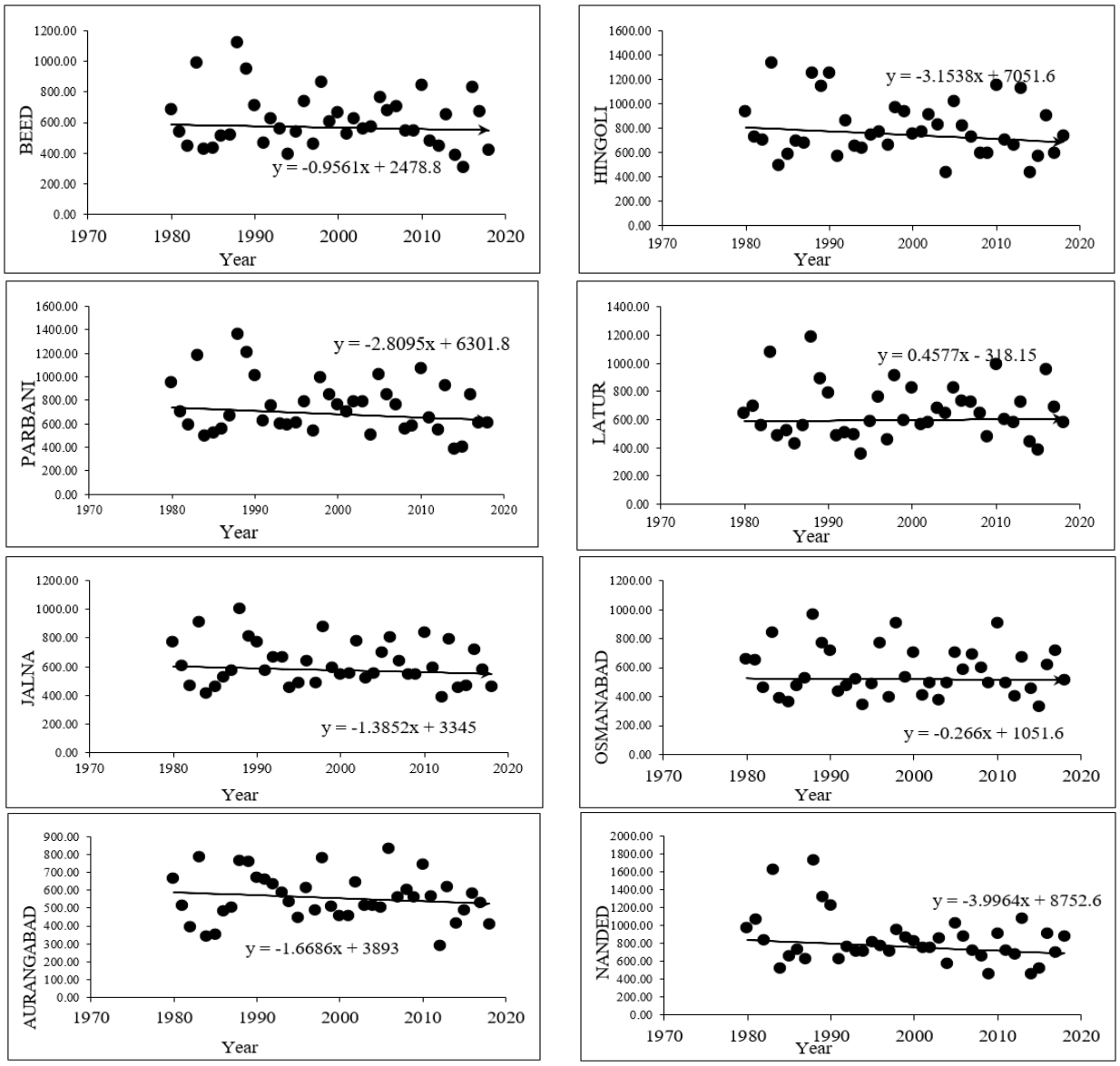


Figure 2: Monsoon rainfall trend from 1980 to 2018 for districts of Marathwada

ends in early to mid of October. The SW monsoon is considered as the principal rainy season in India. Nearly, the whole country receives rainfall during this period. The Southwest monsoon accounts for nearly 75% of rainfall in the country and thus agrarian activities are dependent on it [40].

The SW monsoon is important from India's agricultural perspective. India does not possess any significant irrigation network. Only the states of Punjab and Haryana have proper irrigation infrastructure in place [41,42] Agriculture of rest of the country either survives on monsoon or depends upon the ground water source [43]. It is due to this high dependency on the monsoon, Indian agriculture is often referred as gamble on rains [44].

The characteristics of Indian agriculture is such that the farmers' despite being heavily dependent on rains sow water intensive crops on large scale in the Kharif season (monsoon). Such crops are Paddy and sugarcane. Paddy is prominently grown in the central and eastern regions of India such as Chhattisgarh and West Bengal while Sugarcane is a prominent Kharif crop of Maharashtra which belongs to western region of India. These crops are sensitive to climate change as it causes rise in temperature, and water scarcity [6]. These effects along with ENSO bring in uncertainty over the amount of rainfall [45]. Thus, climate change poses a certain threat to the Indian agriculture.

From the analysis, it is evident that amount of rainfall in the districts of Marathwada is decreasing. Declining trend was determined for the annual rainfall in all the districts. Farmers in the region are known to produce sugarcane. The sugarcane demands for extensive irrigation for increased brix content, grass weight, and larger node to node distance. With the decreasing rainfall in the region, the farmers are facing a loss in the production. The average productivity of sugarcane crop in the Marathwada region is 50 tonnes per acre while the average productivity of the crop in the state of Maharashtra is 80 tonnes per acre [46]. Thus, it can be concluded that decline in the rainfall in monsoon season (Kharif) is projecting its effect on the productivity of the crop in the region. Absence of irrigation infrastructure in the region results in utilization of ground water by the farmers for the irrigation purposes. This activity further

adds up to the woes of the farmers itself. With decline in rainfall the percolation of water during the rainy season also declines which restricts the recharge of the water table [47].

Utilization of ground water in such cases only degrades the water table. Farmers who are not sowing sugarcane are also facing the effects of decreased rainfall. In recent years farmers of the Latur district had to re-sow their crops because of the long drier season [48].

It is being observed that along with annual rainfall the monsoon rainfall is also depicting a negative trend. The trends are huge for districts like Hingoli, Parbhani, and Nanded. Latur and Osmanabad are the districts that are already receiving less amount of rainfall. In such cases, when the amount of rainfall received is declining the farmers should shift from the sugarcane crop to less water intensive crops. Agriculture activity of such crops is an anthropogenic activity that is adding to water crisis of the region. The seasonality index indicates that Latur frequently faces long drier seasons.

The inferior quality of soil in the Udgir, Ausa and Ahmedpur taluka of the Latur district becomes hard during the long drier seasons which along with steep terrain of the region restrict the percolation of water during the rainy season thus further restricting the ground water recharge. Agriculture is not the only anthropogenic activity that are creating water crisis in the region. Jhum style or the shifting style of agriculture is also prominent in the region. Illegal encroachment of forest land is common in the region. The land is cleared by burning the vegetation present [1]. The burning of vegetation leads compressed temperature difference between the land and the Indian Ocean [49]. This reduction in the temperature between land and sea restricts the draft of air from the ocean that further decreases the rainfall amount. Sugarcane crop not sold to the sugar mills is crushed down in makeshift factories to produce Jaggery and country liquor.

The bagasse is used as a bio-fuel for production of heat needed for making jaggery and liquor. The burning of such bio-fuels is a prominent activity in the region [50]. Although bagasse is low in sulphur content but burning it on a large scale releases ample amount of sulphur dioxide, greenhouse gases and nitrogen oxides into the atmosphere [51]. These emissions further reduce the difference between land and ocean temperature and thus acts as a weakening force for the monsoon system. The warming of Indian Ocean is also leading to rainfall woes in India. The warming of Indian Ocean is leading to decrease the difference between ocean and land temperature. This further reduces the rainfall received by the region. Warming of Indian ocean also results in the occurrence of extreme events [52].

The extreme rainfall event in the districts of Latur and Nanded has been credited to the warming of Indian Ocean [53]. Occurrence of such extreme rainfall events in the backdrop of reducing monsoon might lead to sequence of catastrophic events such as loss of livestock, poverty, and agrarian crisis.

7. Conclusion and Future Scope

Drought is an extreme event which has exponentially grown in numbers affecting countless lives and resources' availability. The prominent challenge is experienced by Suryaputra countries falling under International Solar Alliance like India, Brazil, Australia and South Africa. This creates an alarming issue for a necessity of studies that can assist in eradicating such treacherous events. The region of Marathwada has been a pivotal region prone to such events. All the districts of the Marathwada region of the Maharashtra state are witness a decline in the annual rainfall. The decline in the annual rainfall was largest for the Nanded and Parbhani district. The decline of the annual rainfall in the Nanded district was found to be statistically significant. Out of eight districts seven districts of the region have witnessed a decline in the monsoon rainfall over the last 39 years.

The seasonality index calculated indicates that Latur, Beed, and Osmanabad are the drier districts of the region which receives rainfall in more than 3 months. The negative rainfall trend observed might pose a threat to the highly monsoon dependent agriculture of the region. The farmers of the region therefore should migrate from sowing water intensive crops to less water intensive crops such as Sorghum and pearl millet. Sugarcane can also be replaced with less water intensive mandarin which is a less water intensive cash crop. In cases where the farmers are not able to shift to other crops irrigation management system such as irrigation through drip irrigation and rain pipes should be implemented. If negative trend in the rainfall is not effectively managed through change and control of anthropogenic activities the region of Marathwada might enter advanced phases of agrarian crisis which might also lead to the collapse of the agriculture system of the districts comprising it.

The study presented rainfall and seasonality trends which portrayed the topographical and climatic conditions of the districts of Marathwada region. Based on the analysed dataset of 39 years, it can very well be inferred that a spatial time-series analysis yields fruitful information regarding the aspects, characteristics and trend analysis which acts as dominant prerequisites for advanced computation techniques to be deployed for analysis and prediction of droughts in the upcoming years.

References

[1] K. Gabhiye, C. Mandal, Agro-Ecological Zones, their Soil Resource and Cropping Systems. Nagpur: National Bureau of Soil Survey and Land use Planning, 2000.

[2] IPCC, The physical science basis. The contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. New York: Cambridge University Press, 2007.

[3] M. Dore, "Climate Change and Changes in global precipitation" Environmental International, **31**, 1167-1181, 2005.

[4] S. Manwar, P. Vadiya, "Characterization and classification of sugarcane growin soils of Latur district", Annals of Plant and Soil Research, **17**(5), 373-377, 2015.

[5] A, Saini, N. Sahu, P. Kumar, S. Nayak, W. Duan, R. Avtar, S. Behera, "Advanced Rainfall Trend Analysis of 117 Years over West Coast Plain and Hill Agro-Climatic Region of India", Atmosphere, **11**(11), 1225, 2020, 10.3390/atmos11111225.

[6] D.Y. Gumel, A.M. Abdullah, A.M. Sood, R.E. Elhadi, M.A. Jamalani, K.A.A.B. Youssef, "Assessing Paddy Rice Yield Sensitivity to Temperature and Rainfall Variability in Peninsular Malaysia Using DSSAT Model". International Journal of Applied Environmental Sciences, **12**(8), 1521-1545, 2017.

[7] M. Parry, O. Canziani, J. Palutikof, P.V.D. Linden, C. Hanson, Aia Climate Change 2007: Impacts, adaptation and vulnerability. In : Fourth Assessment report of the intergovernmental panel on climae change. Cambridge: Cambridge University Press, 2007.

[8] A. Saini, N. Sahu, W. Duan, M. Kumar, R. Avtar, M. Mishra, S. Behera, "Unraveling Intricacies of Monsoon Attributes in Homogenous Monsoon

Regions of India", Frontiers in Earth Science, 10, 1–17, 2022, 10.3389/feart.2022. 794634.
[9] M.A. Semenov, "Impacts of Climate Change on Wheat in England and Wales", Journal of the Royal Society Interface, 6, 2008, 343-350. doi:10.1098/rsif.2008.0285
[10] S. Osmani, P. Patil, "Drought response and relief by Jaldoot Express: A case study of Latur drought", Zenith IJMR, 9(6), 224-236, 2019.
[11] A. Kumar, A. Singh, "Climate Change and its Impact on Wheat Production and Mitigation through Agroforestry Technologies", International Journal of Environmental Sceinces, 5(1), 73-90, 2014.
[12] O. Dhar, B. Parthasarathy, "Trend analysis of annual Indian rainfall", Hydrological Science, 26, 257-260, 1975.
[13] K. Krishnamurthi, Y. Ramanathan, "Sensitivity of the monsoon onset to differential heating", Journal of atmospherc science, 39, 1290-1306, 1982.
[14] D. Duhan, A. Pandey, "Statistical analysis of long term spatial and temporal trends of precipitaton during 1901-2002 at Madhya Pradesh", Atmospheric Research, 122, 136-149, 2013.
[15] V. Kumar, S.K. Jain, Y. Singh, "Analysis of long-term rainfall trends in India", Hydrological Sciences Journal, 55(4), 484–496, 2010, doi:10.1080/02626667.2010.481373.
[16] A. Gupta, "Flood and Floodplain management in North East India: An Ecological Perspective", 1st International Conference on Hydrology and Water Resources in Asia Pacific Region, Kyoto: Hydrology and Water Resources, 1-10, 2003.
[17] S. Swain, M.K. Verma, M. Verma, "Analysis of Change in Annual Rainfall for Raipur district", IJERT, 3(20), 1-10, 2015.
[18] I. Roy, R.G. Tedeschi, M. Collins, "ENSO teleconnections to the Indian summer monsoon under changing climate", International journal of climatology, 39(6), 3031-3042, 2019.
[19] A. Kulkarni, S. Gadgil, S. Patwardhan, "Monsoon variability, the 2015 Marathwada drought and rainfed agriculture". Current Science, 111(7), 1182-1193, 2016.
[20] SANDRP, Latur Drinking Water Crisis highlights absence of Water Allocation Policy and Management, Retrieved from South Asia Network on Dams, Rivers and People, 2016 https://sandrp.in/2016/04/20/latur-drinking-water-crisis-highlights-absence-of-water-allocation-policy-and-management/
[21] D. Kolekar, V. Vanama, Satellite based Drought Assessment Over Latur, India Using Soil Moisture Derived From SMOS. ISPRS TC V Mid-term Symposium, Geospatial Technology – Pixel to People. Dehradun: ISPRA, 2018, doi:10.5194/isprs-archives-XLII-5-421-2018.
[22] E. Kanellopoulou, "Spatial distribution of rainfall seasonality in Greece", Weather, 57, 215-219, 2002.
[23] S. Ingle, S. Patil, N. Mahale, Y. Mahajan, "Analyzing rainfall seasonality and sNorth Maharashtra", Environmental Earth Sciences, 77, 651-662, 2018.
[24] R. Walsh, D. Lawer, "Rainfall seasonality: description, spatial patterns and change through time", Weather, 36, 201-208, 1981.
[25] R. Yadav, S. Tripathi, G. Pranuthi, S. Dubey, "Trend analysis by Mann-Kendall test for precipitation and temperature for thirteen districts of Uttarakhand", Journal of agrometeorology, 16(2), 164-171, 2014.
[26] I. Ahmad, D. Tang, T, Wang, M, Wang, B, Wagan, "Precipitation Trends over Time Using Mann-Kendall and Spearman's rho Tests in Swat River Basin", Advances in meteorology, 2015, 1-15, 2015, doi:10.1155/2015/431860
[27] H. Tabari, S. Marofi, M. Ahmadi, "Long-term variations of water quality parameters in the Maroon river", Environmental monitoring assess, 177, 273-287, 2011.
[28] N. Karmeshu, Trend Detection in Annual Temperature & Precipitation using the Mann Kendall Test – A Case Study to Assess Climate Change on Select States in the Northeastern United States. Pennsylvania, 2012.
[29] S. Yue, P. Pilon, B. Phinney, G. Cavadias, "The influence of autocorrelation on the ability to detect trend in hydrological series", Hydrological processes, 16(9), 1807-1829, 2002.
[30] S. Shahid, "Trends in the extreme rainfall events in Bangladesh", Theoretical Application of Climatology, 104, 489-499, 2011.
[31] R. Gilbert, Statistical methods for environmental pollution monitoring. New York: Van Nostrand Reinhold, 1987.
[32] F. Wang, W. Shao, H. Yu, G. Wang, X. He, D. Zhang, G. Kan, "Re-evaluation of the Power of the Mann-Kendall Test for Detecting Monotonic Trends in Hydrometeorological Time Series", Frontiers in Earth Science, 8, 1-12, 2020, doi:10.3389/feart.2020.00014
[33] GWP, Droughts and Sugar Industry in Maharashtra – Are We Learning from History? New Delh: Global Water Partnership, 2016.
[34] S. Sandbhor, "Analysis of Behaviour of Real Estate Rates in India-A Case Study of Pune City", International Journal of Economics and Management Engineering, 7(8), 2465-2570, 2013.
[35] R. Singh, "Sugarcane marketing systems in India", Sugar Technology, 13(4), 1-10, 2011.
[36] M. Sabesh, M. Ramesh, H. Prakash, G. Bhaskaran, "Is there any shift in cropping pattern in Maharashtra after the introduction of Bt Cotton", Indian society for cotton improvement, 6(1), 63-70, 2014.
[37] C. Pote, A. Kale, "Effect of Osmotic Stress on Sugarcane (Saccharum officinarum L.) Growth and Physiology", International Journal of Current Microbiology and Applied Sciences, 8(12), 1472-1481, 2019.
[38] R. Garkar, Sugarcane Breeding.Central Sugarcane Research Station, Padegaon, 2017.
[39] M. Rajeevan, D. Pai, R. Kumar, B. Lal, "New statistical models for long-range forecasting of southwest monsoon rainfall over India", Climate Dynamics, 2-17, 2007, doi:10.1007/s00382-006-0197-6
[40] S. Sasane, "Impact of south west monsoon on crop yield: a statistical analysis", International Interdisciplinary Seminar on Geographical and Historical Perspective of Global Problems, 1-10, 2017.
[41] J. Skutsch, J. Rydzewski, Review of research and development needs in irrigation and drainage, Romw: FAO, 2001.
[42] R. Jain, P. Kishore, D. Singh,. (2019). "Irrigation in India: Status, challenges and options", Journal of soil and water conservation, 18(4), 2455-2459, 2019.
[43] K. Kumar, "Climate impacts on Indian agriculture", International journal of climatology, 24(11), 1375-1393, 2004.
[44] S. Gadgil, "The Indian Monsoon". Resonance, 11(8), 8-15, 2006.
[45] K. Tamaddun, "Effects of ENSO on Temperature Precipitation and Potential Evapotranspiration of North India's Monsoon: An Analysis of Trend and Entropy", Water, 11(2), 1-21, 2019, doi:10.3390/w11020189
[46] P. Upreti, A. Singh, "An Economic Analysis of Sugarcane Cultivation and its Productivity in Major Sugar Producing States of Uttar Pradesh and Maharashtra", Economic Affairs, 62(4), 711-718, 2017.
[47] A, Dias, R, Dhawde, N, Surve, A, Weinberg, T, Birdi, N, Mistry, "Impact of climate changes on water availability and quality in the state of maharashtra in western India", Asian Jr. of Microbiol. Biotech. Env. Sc, 17(4), 1071-1081, 2015.
[48] N. Jamwal, Maharashtra Farmers Fear Loss of Kharif Harvest, Blame Met Department. The Wire, 2017.
[49] B. Singh, O. Singh, "Study of Impacts of Global Warming on Climate Change: Rise in Sea Level and Disaster Frequency", Global warming Impacts and Future Perspective, 1-10, 2012, doi:10.5772/50464.
[50] S. Kulkarni, "Development of efficient furnace for jaggery making", International Journal of Recent Scientific Research, 9(5), 26563-226565, 2018.
[51] J. Halofsky, B. Harvey, "Changing wildfire, changing forests: the effects of climate change on fire regimes and vegetation in the Pacific Northwest, USA", Hydrobiologia, 16(4), 1-10, 2020.
[52] M. Roxy, C. Gnanaseelan, Indian Ocean Warming. In Assessment of Climate Change over the Indian Region, 191-206. Springer publications, Singapore, 2020.
[53] A. Yaduvanshi, A. Kulkarni, "Observed changes in extreme rain indices in semiarid and humid regions of Godavari basin, India: risks and opportunities", Natural Hazards, 103, 685-711, 2020.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 36-40 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Fuzzy MPPT for PV System Based on Custom Defuzzification

Abdelmadjid Allaoui*, Mohamed Nacer Tandjoui, Chellali Benachaiba

Electrical Engineering Department, Faculty of Science and Technology, Tahri Mohamed University, Bechar, 08000, Algeria

ARTICLE INFO

Article history:
Received: 20 January, 2023
Accepted: 07 May, 2023
Online: 25 July, 2023

Keywords:
PV system
Fuzzy MPPT
Defuzzification

ABSTRACT

Due to the variations in weather conditions, photovoltaic systems adopt a technique based on maximum power point tracking to extract the maximal power of the solar module. In the literature, there are many different methods classical and intelligent of maximum power point tracking (MPPT). But, due to the semiconductor effect, the current-voltage characteristics of the solar module is nonlinear. This affects its efficiency and make its control not easy. In this contribution, we present a new fuzzy PV MPPT based on custom defuzzification. The obtained power using the proposed fuzzy PV MPPT based on custom defuzzification is significant compared to Pertub & observe and fuzzy PV MPPT in term of performances indices such as: Rise time and overshoot.

1. Introduction

Generally, energy is an important development factor in any economy. Also, energy consumption is a progress indicator. The energy crisis due to the drop of conventional energy sources and the rise in CO_2 emission and environmental pollution has imposed the search for other solutions which are renewables and cleans. As renewable source, photovoltaic power is a very powerful and promising energy potential. The solar energy is converted into electrical energy by solar PV panel. Each type of PV panel has its own specific characteristic according to local conditions such as irradiation, and temperature and this makes the tracking of maximum power point (MPP) a complicated problem. To remedy this problem, many MPPTs algorithms have been presented [1-6]. Conventional MPPTs have proven to be less efficient because of the functioning principle of photovoltaic system which depends on weather.

This, made extracting the maximum power point a difficult task. But, with the development of semiconductor switches which work with high switching frequencies; new MPPT controllers have been developed. Among the intelligent MPPTs, there is the logic-based MPPT which has interested several researchers. In the literature, there are several publications on this fuzzy MPPT controller with more approaches [7-13]. In this context, we will present an intelligent MPPT based on fuzzy logic but which is different with a custom defuzzification function because most of the papers dedicated to this field use default and predefined defuzzification functions.

2. Modeling of photovoltaic system

Photovoltaic system works on the photovoltaic effect to convert solar radiation into direct current. When the sun attacks the panel, the energy is absorbed by a semiconductor. This energy will release the electron-hole pairs from their binding state to supply the load of the photovoltaic system. The output power of the photovoltaic panel depends on environmental variables such as irradiation and temperature. Therefore, to operate the PV system at its maximum power point; the MPPT mechanism is very important and useful. Many MPPT mechanisms have been introduced in the literature by many researchers since the year 1960. Some well-known conventional MPPT methods are incremental conductance, perturbation and observation and constant voltage.

However, this type of method presents classical and limited algorithms. Their implementation requires a good, accurate sensor to measure either voltage or current. Recently, the MPPT based on artificial intelligence is widely used in PV system. These intelligent MPPTs are dynamic with high efficiency and have made the PV system interesting and competitive. Figure 1 shows the block diagram of the proposed standalone PV system. The system consists of a PV array, a MPPT controller combined to a DC- DC converter and a load.

The irradiation (G) and temperature (T) are in charge of the working point of PV system at the maximum power point (MPP) [14]. The cell current, I, which represents the mathematical model of the PV cell can be express as:

*Corresponding Author Abdelmadjid Allaoui, allaoui.abdelmadjid1978@gmail.com

https://dx.doi.org/10.25046/aj080405

$$I = I_{ph} - I_0\left(e^{\left(\frac{q(V+IR_s)}{A.K_cT}\right)} - 1\right)\frac{V+IR_s}{R_{sh}} \quad (1)$$

where, I_{ph} is light-generated cell current (A), I_0 is cell reverse saturation current (A), q is electronic charge, A is ideality factor, K_c is Boltzmann's constant, and T is cell temperature (K).

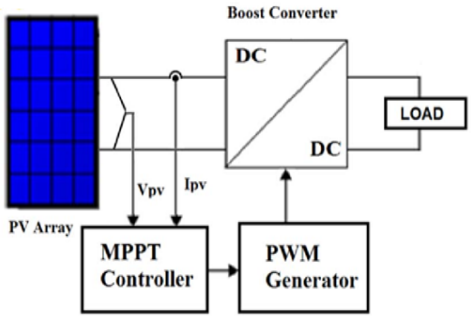


Figure 1: Block diagram of the PV system

3. Perturb & Observe MPPT

Perturb & Observe algorithm is a conventional method. It is used in photovoltaic systems because of its simple implementation. Also, it needs a few measured parameters. It is based on the measure of the PV current and voltage. From these values the power is calculated at each time to find out the maximum power point (MPP).

The principle of this algorithm is based on the operating voltage of the PV module which is perturbed by a small increment and the change of power is observed. If the change of power is positive, then it is supposed that it has moved the operating point closer to the MPP. So, the voltage disturbed in the same track should move the operating point toward the MPP. If the change of power is negative, the operating point has moved away from the MPP. In this case, the direction of perturbation should be reversed.

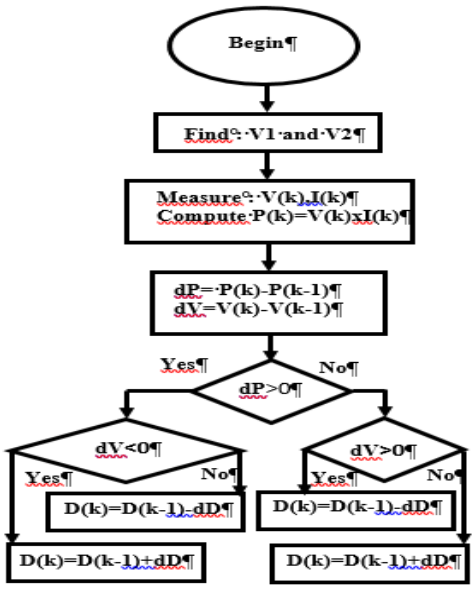


Figure 2: Flowchart of P&O method

4. Fuzzy MPPT based custom defuzzification

The MPPT allows the PV system to work at the maximum despite the variation of its parameters, irradiation, temperature and load. Conventional MPPT methods are limited, however, MPPT based on fuzzy logic fuzzy offers the advantage of being robust, efficient and works to the PPM. The implement of the fuzzy MPPT has three steps: the fuzzification, inference engine and defuzzification (Figure 3).

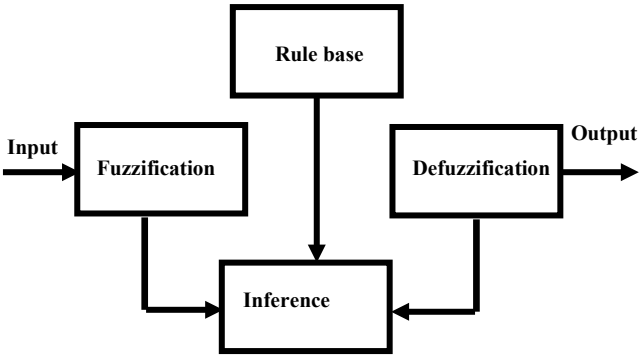


Figure 3: Fuzzy MPPT block diagram

4.1 Fuzzification method

Fuzzification is a method by which sharp values are blurred. To do this, the linguistic variables and the membership functions that will be implemented to model the system must be defined. The principle of fuzzification consists in the decomposition of the universe of discourse of linguistic input and output variables into a number of membership functions.

4.2 Inference engine

Inference is a method by which new information is deduced from the information of premises. Inference in fuzzy logic control systems is a method by which the result of each rule is deduced from the results of each activated rule.

In the literature, dedicated to fuzzy logic, different methods that can be applied to establish an inference engine. The most popular are Mamdani and Takagi-Sugeno-Kang. The Mamdani inference was developed by Ebrahim H. Mamdani in 1975. It was used to modify the behavior of a steam engine. Mamdani's inference was inspired from Lofti Zadeh's paper describing fuzzy sets for systems. For Mamdani max-min inference, the minimal operation is used at implication stage, while the max-min operator is employed to the premises.

4.3 Defuzzification

$$z_{COA} = \frac{\int_z \mu_A(z)zdz}{\int_z \mu_A(z)dz} \quad (2)$$

where, z_{COA} is the control output.

The fuzzy MPPT has two inputs: error (err) and the variation of the error (derr) which are defined by the equations (3) and (4).

$$err = \frac{P_{pv}(\mathrm{k}) - P_{pv}(\mathrm{k}-1)}{V_{pv}(\mathrm{k}) - V_{pv}(\mathrm{k}-1)} \quad (3)$$

$$derr(k) = err(k) - err(k-1) \quad (4)$$

where, $P_{pv}(k)$ and $V_{pv}(k)$ are respectively the output power and instantaneous voltage of the photovoltaic source.

The inference rules are used to evaluate the linguistic values of the activated rules according to their membership degrees. For Mamdani inference, the inference of each activated rule gives a surface. The aggregation of these surfaces gives a final surface which by the defuzzification generates the value of the modulator. The defuzzification allows the conversion of membership degree in crisp value.

Each linguistic variables of the input and the output of the proposed fuzzy MPPT based on custom defuzzification has five membership functions. The used linguistic values are: NB (Negative Big), NS (Negative Small), ZE (Zero), PS (Positive Small) and PB (Positive Big).

The majority of researchs uses Matlab's predefined defuzzification functions for fuzzy MPPT controller. Through this contribution, we will propose a custom function of defuzzification which will boost the action of the fuzzy MPPT controller.

$$Z_{Custom} = \frac{\sum_{i=0}^{n} k\ ymf_i\ x_i}{\sum_{i=0}^{n} ymf_i^{\ 2}} \quad (4)$$

where, k is a gain that regulates the rise time.

5. Simulation and discussion

The proposed fuzzy PV MPPT based on custom defuzzification is compared to P&O MPPT. The simulation system is presented below.

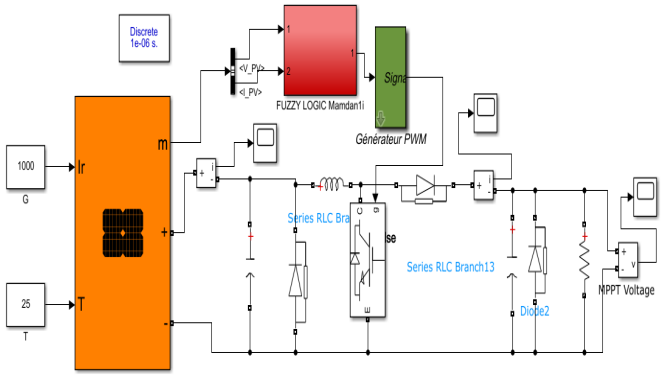


Figure 4: Simulation system under Matlab Simulink

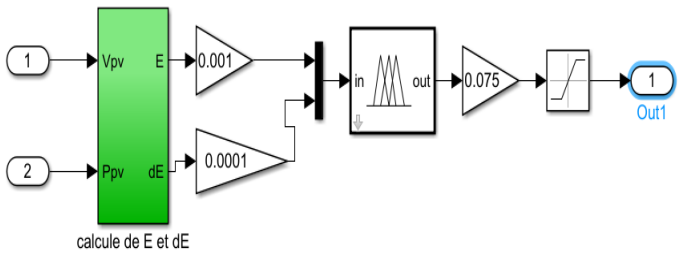


Figure 5: Fuzzy MPPT controller design

Figure 5 shows the design of fuzzy PV MPPT controller which has two inputs: Error and change of error according to "(2)" and "(3)". Table 1 presents the electrical characteristic values of the PV module.

Table 1: Electrical characteristics of the PV module

Electrical characteristics	values
Maximum power (W)	85.383
Cells per module (Ncell)	36
Open circuit voltage Voc (V)	22
Short circuit current Isc (A)	5.2
Voltage at maximum power point Vmp (V)	17.9
Current at maximum power point Imp (A)	4.77

Figures 6-9 show simulation steps of fuzzy controller for PV MPPT with custom defuzzification function. The inputs to the fuzzy controller are error and error variation of the PV system. The error is the ratio between the variation of the power on the variation of the PV voltage. The output of the fuzzy controller represents the modulator which will modulate the carrier of the PWM generator to produce the PWM signal which will drive the switch of the boost converter according to the inferences of fuzzy controller for PV MPPT.

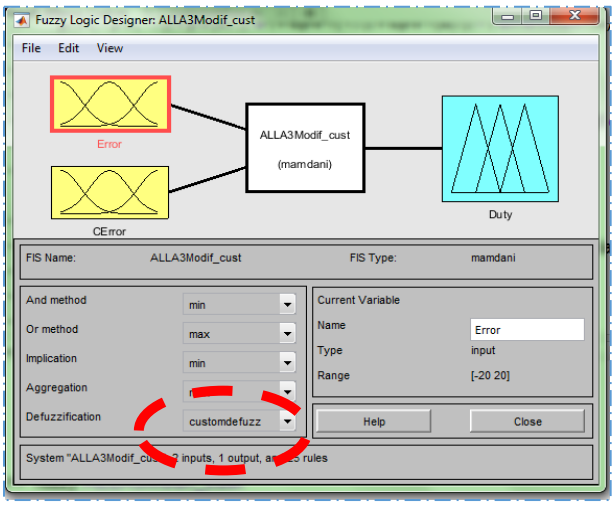


Figure 6: Fuzzy logic designer

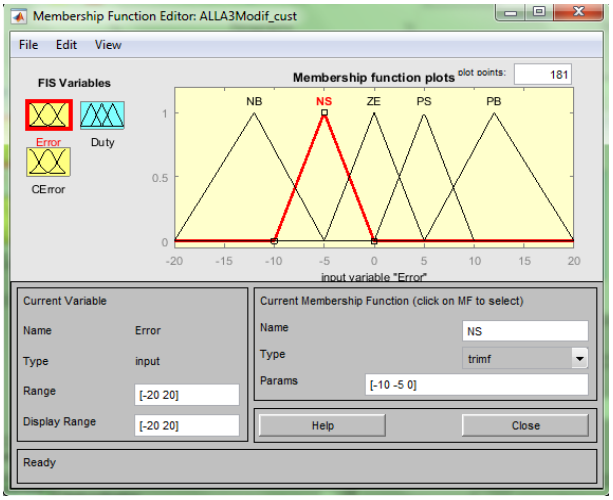


Figure 7: Error membership functions

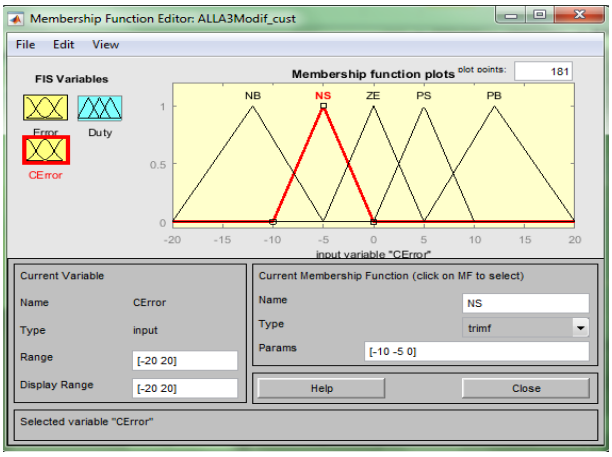


Figure 8: Change of error membership functions

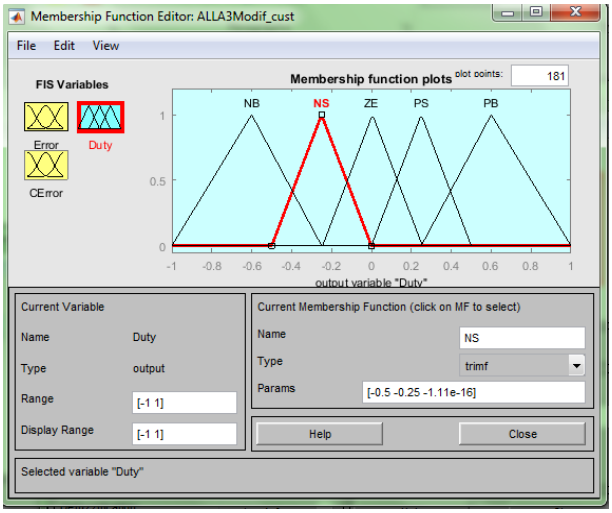


Figure 9: Output membership functions

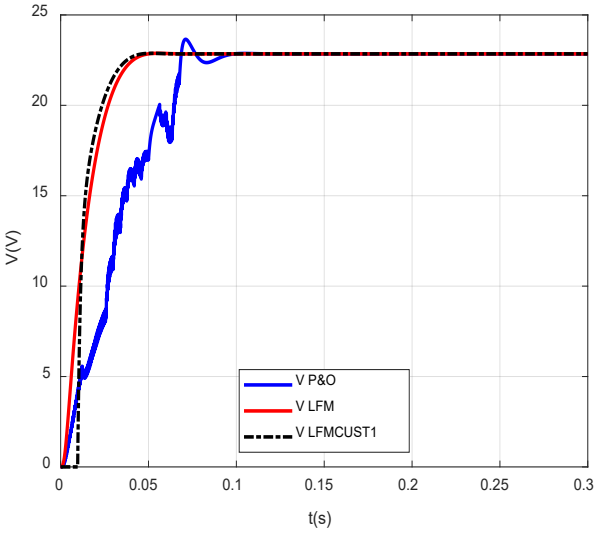


Figure 10: Load voltages

Figures 10-15 present the results obtained from the simulated system which are the voltage, the current and the power of the load. Each figure presents the comparison of the three PV MPPTs which are P&O, fuzzy with Mamdani inference and centroid defuzzification and the response of the proposed fuzzy PV MPPT with Mamdani inference and custom defuzzification. The contribution of fuzzy PV MPPT is interesting. The response of the proposed fuzzy PV MPPT with Mamdani inference and custom defuzzification is better than the other MPPTs. Table 2 shows the load power performance of the three MPPTs, and the response of the proposed fuzzy MPPT for PV system with Mamdani inference and custom defuzzification presents less rise time and overshoot than other MPPTs.

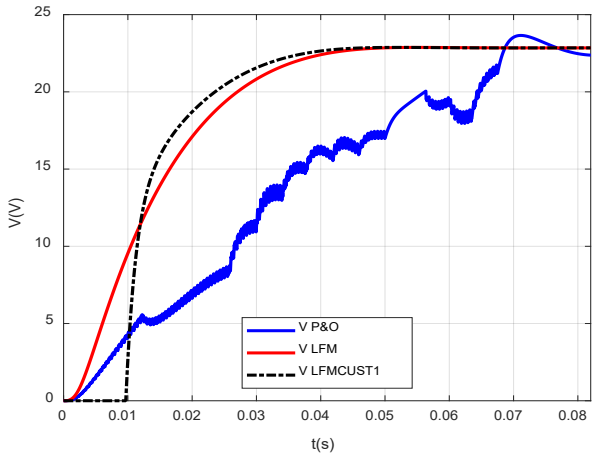


Figure 11: Zoom of load voltages

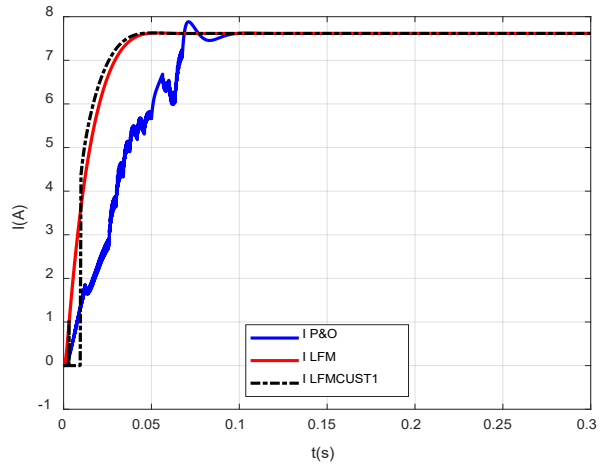


Figure 12: Load currents

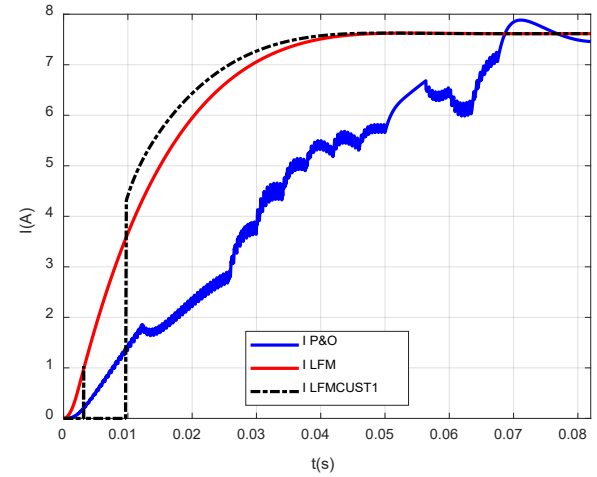


Figure 13: Zoom of load currents

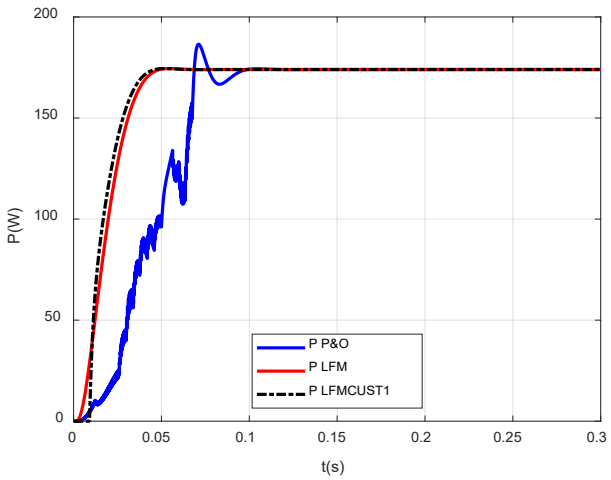


Figure 14: Load powers

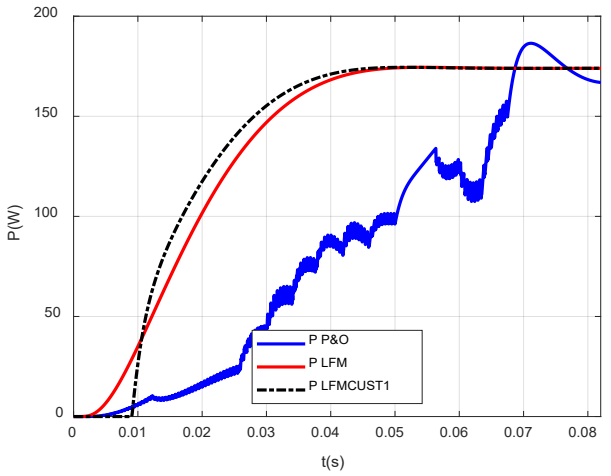


Figure 15: Zoom of load powers

Table 2: Performance of load power responses

Performance indices	P&O	Fuzzy MPPT with Centroid defuzz.	Fuzzy MPPT with Custom defuzz.
RiseTime	0.0470	0.0262	0.0209
SettlingTime	0.0902	0.0422	0.0393
SettlingMin	149.7038	156.5957	156.5957
SettlingMax	186.4765	174.4732	174.4730
Overshoot	7.1669	0.2762	0.2760
Peak	186.4765	174.4732	174.4730
PeakTime	0.0712	0.0545	0.0516

6. Conclusion

Conventional MPPTs have proven to be less efficient because of the functioning principle of photovoltaic system which depends on weather conditions. The principle of fuzzy logic based on the degree of membership has made it possible to implement a more efficient and more robust MPPT controller than conventional ones.

In order to make the MPPT controller faster and at the same time enrich the literature in this domain, we have proposed a new fuzzy PV MPPT controller based on custom defuzzification function. The results obtained show the action of defuzzification is important in the system response.

The proposed fuzzy PV MPPT controller based on custom defuzzification has confirmed this action and has given better results than both MPPT controllers P&O and Fuzzy MPPT for PV system according to of the performance indices of the responses.

References

[1] J. M. Riquelme-Dominguez, S. Martinez, "Systematic Evaluation of Photovoltaic MPPT Algorithms Using State-Space Models Under Different Dynamic Test Procedures," IEEE POWER & ENERGY SOCIETY SECTION, **10**, 45772–45783, 2022, doi: 10.1109/ACCESS.2022.3170714.

[2] X. Li, Q. Wang, H. Wen, and W. Xiao, "Comprehensive studies on operational principles for maximum power point tracking in photovoltaic systems," IEEE Access, **7**, 121407–121420, 2019, doi: 10.1109/ACCESS.2019.2937100.

[3] R. Dutta, R. P. Gupta , "Performance analysis of MPPT based PV system: A case study," 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies, Patna, India, doi: 10.1109/ICEFEET51821.2022.9847729.

[4] M. Etezadinejad, B. Asaei, S. Farhangi, A. Anvari-Moghaddam, "An Improved and Fast MPPT Algorithm for PV Systems Under Partially Shaded Conditions," IEEE Transactions on Sustainable Energy, **13**(2), 732–742, 2022, doi:10.1109/TSTE.2021.3130827.

[5] X. Li, H. Wen, Y. Hu, L. Jiang, "A novel beta parameter based fuzzy-logic controller for photovoltaic MPPT application',Renewable Energy, **130**, 416-427, 2019, doi:10.1016/j.renene.2018.06.071.

[6] J. S. Ko, J. H. Huh, J. C. Kim, "Overview of maximum power point tracking methods for PV system in micro grid," Electronics, **9**(5), 816, 1–22, 2020, doi.org/10.3390/electronics9050816.

[7] R.B. Bollipo, S. Mikkili, P. K. Bonthagorla, "Hybrid, optimal, intelligent and classical PV MPPT techniques: a review," CSEE Journal of Power and Energy Systems, **7**(1), 9–33, 2021, doi: 10.17775/CSEEJPES.2019.02720.

[8] Tao Hai, Jincheng Zhoua, Kengo Muranak, "An efficient fuzzy-logic based MPPT controller for grid-connected PV systems by farmland fertility optimization algorithm," Optik, **267**, 2022, doi.org/10.1016/j.ijleo.2022.169636

[9] P. Boonraksa, T. Chaisa-Ard, S. Sommat, P. Pimpru, T. Booraksa, B. Marungsri, "Design and Simulation of Fuzzy logic controller based MPPT of PV module using Matlab Simulink," International Electrical Engineering Congress, 2022, doi: 10.1109/iEECON53204.2022.9741641.

[10] F. Mehazzem, M. André, R. Calif, "Efficient Output Photovoltaic Power Prediction Based on MPPT Fuzzy Logic Technique and Solar Spatio-Temporal Forecasting Approach in a Tropical Insular Region," Energies, **15**, 1–21, 2022, doi.org/10.3390/en15228671.

[11] C. R. Algarín, J. T. Giraldo, O. R. Álvarez, "Fuzzy Logic Based MPPT Controller for a PV System," Energies, **10**(12), 2–18, 2017, doi: 10.3390/en10122036.

[12] G.F.T. Kebir, C. Larbes, A. Ilinca, T. Obeidi, S T. Kebir, "Study of the Intelligent Behavior of a Maximum Photovoltaic Energy Tracking Fuzzy Controller," Energies, **11**, 2–20, 2018, doi: 10.3390/en11123263.

[13] Tehzeeb-ul Hassan, R. Abbassi, H. Jerbi, K. Mehmood, M.F. Tahir, K. M. Cheema, R. M. Elavarasan, F. Ali, I. A. Khan, "A Novel Algorithm for MPPT of an Isolated PV System Using Push Pull Converter with Fuzzy Logic Controller," Energies, **13**, 2–21, 2020, doi: 10.3390/en13154007.

[14] M. Seyedmahmoudian, R. Rahmani, S. Mekhilef, A.M.T. Oo, A. Stojcevski, T. K. Soon, A. S. Ghandhari, "Simulation and Hardware Implementation of New Maximum Power Point Tracking Technique for Partially Shaded PV System Using Hybrid DEPSO Method," IEEE Trans. Sustain. Energy, **6**(3), 850–862, 2015, doi:10.1109/TSTE.2015.2413359.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 3, 40-47 (2023)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Indoor Positioning: Comparing Different Techniques and Dealing with a user Authentication use Case

Joaquín Pérez Balbela[*,1], Aruna Prem Bianzino[2]

[1] *Universidad Internacional de La Rioja - UNIR, Uruguay*

[2] *Fundación Tecnológica Advantx - FUNDITEC, Madrid, Spain*

ARTICLE INFO

Article history:
Received: 28 February, 2023
Accepted: 26 April, 2023
Online: 15 May, 2023

Keywords:
Indoor Positioning
Taxonomy
Technique Comparison
Real Scenario

ABSTRACT

Indoor positioning is a complex issue with many, heterogeneous application cases, each one presenting different requirements and environments. In such a complex ecosystem, an agile[1] taxonomy is needed to be able to select a proper solution for a given scenario, as well as practical recommendations for the most used solutions. Besides providing these tools, we analyze a real-world scenario and its requirements, selecting a practical solution and evaluating it together with its implications and consequences, providing a reference guideline for practical applications of indoor positioning.

1 Introduction

The location of objects and people in spaces where satellite technologies fail, or lack precision (e.g., inside complex buildings and underground locations) is a complex issue, while it represents a needed feature with many application cases, including access control, offering of personalized services, navigation in complex structures, crowd control, etc. User location may also be used as an authentication element in the case of critical infrastructure control, such as power plants.

Due to the strong need and the heterogeneity of the use cases, many different solutions have been proposed and developed. Each different context (e.g., commercial building, military facility, critical infrastructure, etc.) presents different scenarios (e.g., available network and infrastructure, barriers, variability and people presence, etc.) and different needs (e.g., cost limits, precision needs, security, privacy, etc.). Similarly, different solutions present different requirements, technologies, costs, and precision. Due to the complexity of the resulting ecosystem, it is important to properly understand the needs of the scenario and the implications of the available solutions. In this paper, we will contrast different techniques for indoor positioning, focusing on different key aspects with the aim of producing guidelines for the technology selection in common scenarios. Finally, we will evaluate different solutions in a real environment and will consider a practical case of location as an authentication element in the context of critical infrastructure management and monitoring in energy production and usage.

The rest of this paper is organized as follows: in Section II we introduce a taxonomy for Indoor positioning solutions, analyzing the different relevant aspects, their solution space, and practical implications. In Section III we analyze different available techniques for indoor positioning, placing them in the proposed taxonomy, analyzing their practical implication, and providing recommendations for the scenarios and environments in which its deployment could be more suitable. In Section IV we present some real-world examples of commercial applications integrating Indoor Positioning, together with the corresponding main characteristics, limitations, and requirements. This analysis includes examples covering applications in the Sport, Logistics, and Food industries. In Section V we evaluate a real-world use case, i.e., the use of indoor location as an authentication factor in critical infrastructure (i.e., energy production management and monitoring), detailing its requirements, selecting a practical solution and testing it in a real, relevant environment, and analyzing its use accounting for the relevant evaluation criteria, i.e., precision, security, confidentiality, integrity, and availability. Finally, in Section V we draw the conclusion of our research work and describe possible future extensions.

[*]Corresponding Author: Joaquín Pérez Balbela, Email: joaquinperezbalbela@gmail.com

2 Solution Taxonomy

Different attempts are available in the literature presenting a classification of positioning techniques. Among the most recent ones, we would like to highlight [1] and [2]. Still, these works perform a much wider analysis and include dimensions that may not be relevant to the scenario analysis and technology selection, such as the purpose (i.e., healthcare, retail, etc.), resulting in a less agile overview and lacking practical recommendations for real-world scenarios. This is why we present here a minimal, yet relevant taxonomy, allowing to order the most used and relevant solutions for indoor positioning.

As introduced, many different techniques are available for indoor positioning, targeting different scenarios and contexts, using different technologies and different base solutions. In this section, we will describe the resulting solution space and define how we limited our scope.

First of all, in the remaining of this paper, we will refer to the object or person to be located as the "target device" (TD), as we will consider solutions where the location is calculated for a device, co-located with the target object or person.

For what concerns the taxonomy dimensions, in the first place, the scope of the techniques may range from (i) knowing the location of a TD in space from an external system (i.e., **location**), (ii) allowing a TD to know their location in space (i.e., **positioning**), (iii) following the location of a TD over time (i.e., **tracking**), (iv) to computing an (optimal) path from the current location of a TD to a specified destination (i.e., **navigation**). This dimension actually depends on the problem to be solved and most solutions may be used for any scope. Eventual differentiation points in this dimension may be represented by the frequency at which the location is estimated, e.g., if a solution estimated the TD's location once every minute, it will result in a poor user experience if used in a navigation context.

Secondly, different information may be used as input for the solution, including input from different sensors of the TD and/or from the environment. The solution may use one or more of the following:

- **Device motion**, including acceleration and changes in the acceleration, tilting angle and its change rate, proximity to specific or generic nearby objects, etc.
- **Environment sensing**, including light intensity, environmental noise, magnetic field, atmospheric pressure, temperature, environment recognition through the camera, etc.
- **Audio** production and/or sensing of specifically generated signals.
- Communication with an **existing network infrastructure**, including Bluetooth beacons, cellular towers, satellite signals, WiFi, etc.
- Detection of specific **tags**, including visual tags, RFID tags, etc.

Different information may derive from different sensors, whose availability should be checked with the target device set, which may result in different power consumption and costs, may require different infrastructure (e.g., a specific wireless network in the target area), may result in different precision and availability depending on the target scenario, and, finally, may impact in different ways the user's privacy.

Finally, how this information is used is another big differentiation point among different solutions for indoor positioning. In particular, the input information may be analyzed:

- as a **single value** (e.g., an intensity measure to calculate the distance from a reference point),
- as a **value variation** (e.g., acceleration to estimate a path of the TD into an area), or
- against **reference values** (e.g., pre-mapping of values in the target area and contrasting the measured value against the map).

Solutions belonging to different categories result in different compatibility (e.g., need for specific sensors in the TD, or specific infrastructure in the area), different costs (sensors, infrastructure), and energy consumption and workload for the TD (different sensors, infrastructure, sampling frequency, tracking/navigation, vs. location/positioning), different needs for information exchange between the infrastructure and the TD, resulting in different privacy level for the device user, and, certainly, different precision levels. As such, the selection of the solution to be used in a specific context should be carefully evaluated, taking into account all the relevant aspects.

3 Main Techniques Considered

In this work, we considered techniques not needing an existing infrastructure (i.e., Magnetic Map), or based on the communication with an existing network infrastructure and using standard sensors on the TD, specifically considering WiFi and Bluetooth as network infrastructures, due to their high availability and compatibility with existing mobile devices. In particular, we detected four main solution classes for solutions leveraging on existing WiFi infrastructure, on the basis of the used information: Radio Signal Strength (RSS), Angle of Arrival (AoA), Time of Arrival (ToA), and Fingerprinting. For each considered solution type, recommendations on real-world scenarios in which they may be used are included. A summary of the analyzed solutions is reported in Table I.

The different techniques described can be combined with each other to obtain hybrid solutions (e.g., [3] combining RSS and Fingerprinting, [4], combining Bluetooth Beacons and Fingerprinting, or [5], combining RSS and AoA). This generally results in a higher accuracy compared to the use of solutions using a single technique [6], and/or higher system availability, as if one of the combined solutions alone would not be available in certain areas/settings, the others may be. On the other hand, solution hybridization generally results in higher costs as more infrastructure and/or sensors are needed, as well as more computational power. Still, each individual solution results in a baseline for the combination, considering costs and performance.

3.1 Radio Signal Strength (RSS)

The solutions belonging to this classification use the signal strength of the WiFi network as measuring input. This may either be directly measured or as a Signal-to-noise ratio. This allows for estimating the distance from the WiFi transmitter, whose location is known. Using at least 3 different transmitters, a triangulation is possible and the location may be estimated, like in [7], [8], or [9]. The human body itself alters the received signal strength, as such, the body position with respect to the device should be taken into account.

In general, this class of solutions presents a significant variation in the measured distance between the TD and the signal emission points, due to the TD motion pattern, user positioning with respect to the device, and the presence of eventual obstacles (e.g., metal elements in the building structure). On the other hand, this class of solutions is based on commonly available infrastructure, HW available on any device, and uses standard functions. This class of solutions is recommended for usage in areas clear from obstacles and with a line of sight to the access points.

3.2 Angle of Arrival (AoA)

The solutions belonging to this class use the angle at which the WiFi signal is received as input. In this case, a multidirectional antenna emits broadcast signals with a given frequency, while directional antennas emit specific signals at different given frequencies. The time interval between the reception of the multidirectional and the directional signals allows for calculating the distance between the TD and the emitting antenna. This method, originally used to locate flying devices, has been adapted to indoor positioning, providing higher precision [10], [11], [12].

Similarly to what happens in the RSS case, the presence of obstacles between the TD and the emitting antenna may distort the power and direction of the signal. In general, this class of solutions presents higher precision when the TD is far from the base station, especially in spaces with few or no obstacles (e.g., hangars, open spaces, etc.). This class of solutions is recommended for usage in areas clear from obstacles and with a high distance between the TD and the access points.

3.3 Time of Arrival (ToA)

The solutions belonging to this class use the time difference between the signal emission from the base station and its reception at the TD. This input is then used to estimate the distance between the two elements. Also in this class of solution, the location may be estimated using a triangulation (i.e., at least 3 different transmitters) [13], [14]. In this case, the estimation is affected by the signal attenuation introduced by obstacles, which is reduced using higher frequency signals (e.g., UltraWideband - UWB), resulting in higher precision. Still, the precision of this class of solutions is drastically reduced by obstacles between the transmitter and the TD, even using UWB signals [15]. Another solutions that is ToA-based is LiDAR (Light Detection and Ranging) which is used to measure distance using light or laser beams.

Overall, this class of solutions accounts for higher precision, but requires specific hardware (e.g., UWB transmitters and receivers) and higher area coverage, resulting in higher costs. For the usage of this class of solutions in areas with a high number of obstacles, a higher number of transmitters should be used.

3.4 Fingerprinting

The solutions belonging to this class require a preliminary mapping of the signal fingerprint in the target area (e.g., [16] where the WiFi signals are mapped, or [17] where mapping of WiFi as well as of other signals are considered, or [18], where AoA samplings are mapped). The solution then measures the current signals in the TD and maps them to the most similar fingerprint to estimate the TD location.

On the one hand, this class of solutions does not require special hardware, but it is instead based on the present infrastructure (e.g., WiFi network) and on sensors commonly present on mobile devices (e.g., WiFi antenna), using measurements already commonly performed by mobile devices (e.g., WiFi signal strength). On the other hand, any change in the area setting (e.g., replacing/moving an access point, introduction/elimination repositioning of an interfering element, etc.) requires a new mapping. Fingerprinting introduces ambiguity points, where the fingerprint may be similar and therefore it may not be possible to estimate an accurate location in all cases [6]. Finally, the precision of this class of solutions depends on the number of wireless networks present in the area: the higher, the better. This class of solutions is recommended for usage in areas whose configuration and status are stable in time.

3.5 Bluetooth Beacons

Other kinds of wireless networks may be used to communicate with the TD and estimate its location. Solutions belonging to this class use Bluetooth beacons, with a known location, transmitting any payload to know which devices are present in their coverage area. Intersecting the areas to which a device belongs, it is possible to estimate its location. RSS solutions based on Bluetooth are also possible but result in higher energy consumption and lower precision [19], [20], [21].

The energy consumption of this kind of solution is really low, especially if Bluetooth low energy (BLE) or similar standards are used. At the same time, the payload transmitted by the beacons may be used to transmit useful information (push notifications, localized advertisement or service description, etc.), even if, it should be said, the infrastructure does not provide further services as the internet connectivity provided by the WiFi infrastructure of the solution classes analyzed up to now. On the other hand, specific hardware is required (Bluetooth beacons), in a number (and cost) proportional to the required precision and area to be covered. For the usage of this class of solutions in areas with a high number of obstacles, a higher number of beacons should be used.

3.6 Magnetic Map

Similarly to the solutions explored in section III-D, a mapping of the magnetic field present in the different points of the target area may be performed [22], [23], [24]. This kind of solution requires specialized hardware (i.e., at least 3 magnetic field sensors, specifically aligned and placed on the TD). On the other hand, this type of

solution results in higher precision than other methods, although it faces the same drawback as other fingerprinting methods: changes in the environment require the area to be mapped again since its magnetic map may have changed. We have not found evidence of this method being widely implemented, mainly due to its specific hardware requirements.

3.7 Summary

A summary of the analyzed solutions is reported in Table I, together with their classification on the basis of the proposed taxonomy. All these solutions may be used for any *scope*, depending on the application processing the info they return. As we can see, different solutions result in different tradeoffs, especially regarding costs and precision ("Avg. Error" in Table I). Note that, for each technique, a specific implementation and environment should be taken into account to determine its precision. For each technique, in Table I, the reference implementation is reported in the first column. This should be carefully evaluated, together with the other relevant parameters (i.e., confidentiality, availability, security, integrity), when selecting a practical solution for a specific application scenario. An example of this process will be provided in Section V.

Table 1: Solution summary

Name	Info	Usage	Cost	Avg. Error
RSS [7]	Existing Net.	Single value	Low	2.9-16.3m
AoA [10]	Existing Net.	Single value	Low	2.1m
ToA [15]	Existing Net.	Single value	High	2m
Fingerp. [16]	Existing Net.	Ref. values	Mid	0.6-1.3m
Beacons [19]	Existing Net.	Single value	Mid	9.7m
Magn. Map [22]	Environment	Ref. values	High	0.1m-0.16m

4 Using Indoor Positioning: Real World Examples

In this section, we provide some examples of Indoor Positioning as used in real-world commercial applications, together with the corresponding main system characteristics, limitations, and requirements. We will analyze a few examples through Sport, Logistics, and Food industries.

4.1 Using Indoor Positioning in Sports

Some sports have benefited from using indoor positioning techniques, notable examples include tennis, football, and handball.

In tennis, precise ball tracking is required to determine whether it is in or out, and many commercial systems are based on optical cameras that track the ball's position and gather the necessary information for the Hawk-Eye to process. The optical Hawk-Eye system has an estimated error range of 3.6mm at impact [25].

Validations of UWB technology have been executed to compare the accuracy of the optical Hawk-Eye system versus the UWB-based one. In this study, an optical system was paired with a UWB-based Local Positioning System (LPS), and a tennis match was tracked using both systems. Depending on the considered parameters, the mean error range between both systems was in the range of 13.1 to 17.8cm [26]. However, taking the cost of deploying a Hawk-Eye system into consideration, it becomes a competitive alternative. The cost of a professional Hawk-Eye system using 10 high-speed cameras starts at between 60000 to 70000 US dollars per Tennis court [27], while other UWB-based systems' cost can start at approximately half of this value [28].

The last FIFA World Cup 2022 in Qatar used a UWB-based live ball tracking system developed by Kinexon. The official ball contained an internal sensor weighing 7 grams transmitting in the 500Hz band which, according to Kinexon, tracked the ball's movement 100 times per second with centimeter accuracy when paired with their real-time locating system (RTLS). Additional metrics can be derived from this data such as ball possession and speed, among others [29].

A study was conducted to assess Kinexon's solution and compare its error rate to other systems. In this study, the system was used to track both players and the ball, and it was measured in a specially designed circuit on the Fraunhofer L.I.N.K test circuit and in a football game, to gather knowledge regarding how the system would react in different situations. The image of the circuit is included below, and it includes linear sprints, curved sprints, agility tests, and direction changes, among other tests [30].

This system can also be used in handball, although its accuracy is approximately 18,9% lower compared to football due to different sport kinematics [30]. Standard deployment of this system may cost between 20000 to 30000 US dollars per field [28].

As a brief summary, although UWB-based LPS may have a larger error rate compared to visual or infrared systems which are considered the *gold standard* when taking cost into account, those solutions become an interesting choice.

4.2 Using Indoor Positioning in the storage and distribution industry

An important part of shipping orders is the picking phase, in which the desired items are located in the various warehouses and included in the customer's order. Depending on the number of items to be picked, this phase may be a time-consuming one, and, as warehouses grow in size, finding the necessary items to complete orders may be a challenging task.

Indoor positioning solutions can be used to guide the worker or smart robot as it navigates the store searching for items (using an optimized route) or to generate real-time product location maps, for example.

Amazon is one of the biggest worldwide retailers and has a large number of warehouses around the world. Those warehouses are run using a technique called "chaotic storage", in which the items are placed in random locations trying to use the spare space as efficiently as possible, using no predefined zones or locations for those items. This translates into items being tracked solely by software, without relying on the worker or any other method. In Amazon's case, this storage method generates a large amount of raw data that can be used to improve internal processes, such as tracking stock or sending items first with a closer expiration date [31].

4.3 Using Indoor Positioning in the food industry

Restaurants have benefited from using indoor positioning systems to enhance their efficiency, reduce wait times and offer a better customer experience. As an example, fast food restaurants use these LPS to track food delivery after the order is placed through kiosks. A marker is assigned to the order and allows the restaurant's personnel to determine where the customer is seated in order to deliver the order.

For example, one such platform is the LPS deployed at select McDonald's locations, which is based on Bluetooth Low Energy (BLE) technology developed by Acrelec and Radius Technology Inc. These devices are located on plastic table markers, powered by a button cell Lithium battery and their size is approximately two regular coins.

According to the manufacturer, the device's main microprocessor is an ARM-based Cortex M4 SOC (Nordic Semiconductor nRF52832) and includes different sensors such as an accelerometer and fall detection. This SOC states it supports different protocols such as Apple's iBeacon, AltBeacon, and Google's Eddystone, allowing for device interoperability, and has a maximum transmission distance of 100 meters [32]. This SOC is also very popular among enthusiasts, mainly due to its features, cost, and compatibility with Arduino development boards, even though it does not include integrated Wi-Fi functionality.

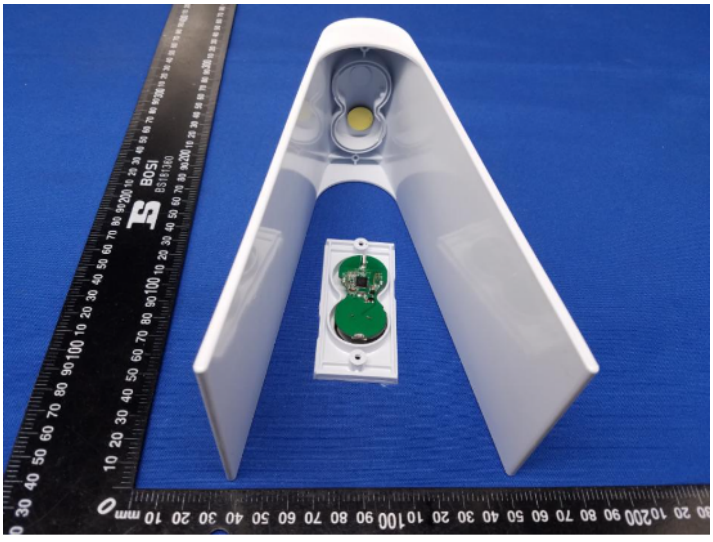


Figure 1: View of the plastic table marker and the internal BLE beacon [32].

Acrelec's platform allows floor maps to be included in their platform, showing the location of the marker on the restaurant's floor plan to help the staff locate the customer in a timely manner.

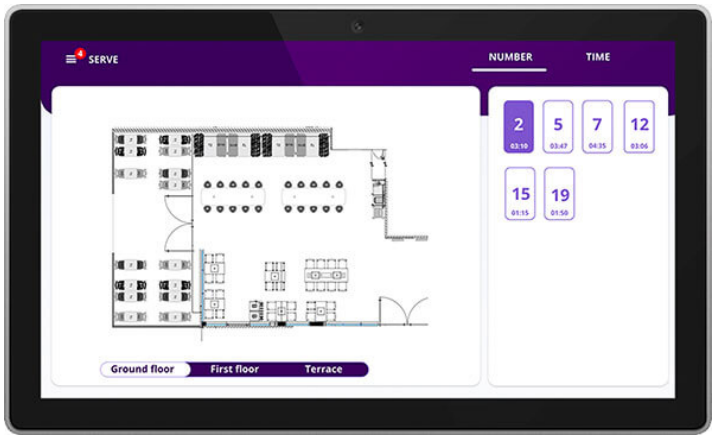


Figure 2: View of Acrelec's Table Service platform [33].

5 The User Authentication Case

More and more users consume resources and services from mobile devices. Authenticating these users is of paramount importance to regulate the access to services, resources, and sensitive data, and to limit it to only authorized users. User authentication is the process of establishing with reasonable accuracy whether users are who they claim to be. The process is usually based on some credentials, whose ownership and verification guarantee to check the user identity, and that may fall into one of the following categories: (i) something that the user knows (e.g., a secret keyword), (ii) something that the user owns (e.g., a physical object like a smart card), (iii) something that the user is (e.g., any measurable physical feature, univocally identifying the user, like fingerprints or iris identification, for instance), (iv) something that the user does (e.g., motion patterns, signature, etc.), or (v) somewhere that the user is (e.g., being in a specific location). This information (identifier), may be combined including information belonging to different categories, to improve security (i.e., Multi-factor authentication, like when a user is asked for a password - something that the user knows - and to enter a code received at their mobile phone - something that the user owns). In particular, in the case of power-plant control and monitoring, as in the Robinson Project, due to the sensitivity of the context and to the possible consequences of identity theft, we proposed an identification mechanism including positioning among the used identification elements: the user must be physically present where they are accessing the service (Power plant monitoring and control), in order for them to be allowed to access the service itself. As physical access is subject to other external security measures, this solution guarantees a very strong identification.

5.1 Target Requirements

In order to select a specific solution for the considered context, we must keep into account the different analyzed parameters. In the considered scenario, location is the only relevant scope of the solution, while the other leading factors are the technology availability and compatibility, the solution availability, precision (Maximum error lower than the distance between the control panel and the access to the closest room, 2.5m in the example taken into account. A differ-

ent threshold or a confidentiality interval may be set depending on the environment), and data security and privacy (extremely relevant in an authentication context). On the other hand, the environment is not subject to frequent changes or variable interference (e.g., crowd presence). As such, the natural choice is to use a Fingerprinting solution based on WiFi.

(a) (b) (c) (d)

Figure 3: Devices used for the solution evaluation: (a) Mikrotik access point, (b) laptop, (c) Galaxy S8 (TD), and (d) Galaxy S21.

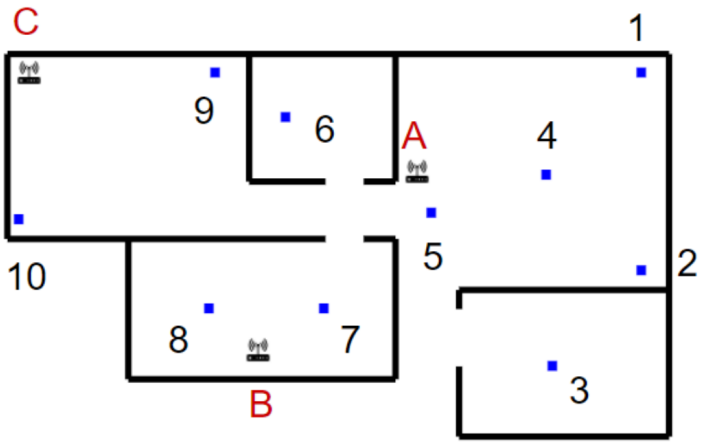


Figure 4: Evaluation environment, including equipment placement (A, B, and C) and measurement points (1 to 10).

5.2 Evaluation Scenario

We used a real environment and deployed the Anyplace tool [34] to evaluate the selected solution. Anyplace relies on fingerprinting, as well as RSS, and combines several methods to obtain a lower error margin. Networking equipment was installed in this environment, using a Mikrotik hap AC^2 access point broadcasting on both 2.4 and 5GHz bands, a laptop equipped with a Killer AC1550i WiFi card, and a Galaxy S21 mobile phone acting as a hotspot, both broadcasting on the 2.4GHz band. A Galaxy S8 mobile phone acted as TD. The used devices are depicted in Figure 1. The software solution used in this scenario was Anyplace, running on the University of Cyprus' public servers. A single-floor, multi-room space divided with masonry walls was used as an evaluation environment. The equipment was placed as shown in Figure 2 to cover an area of about $52m^2$.

Measurements were taken on the points highlighted in Figure 2. At each point, we took four consecutive readings and averaged the result to smoothen possible interference or disruption, measuring the error as the linear distance between the detected and real location. This measurement-taking procedure was repeated on three occasions: with a single access point (i.e., the Mikrotik access point - A), with two access points (i.e., including also the laptop - B), and with three access points (i.e., including also the Galaxy S21 hotspot - C).

5.3 Results

As introduced above, in order to evaluate the solution, we take into account different aspects: the solution precision in estimating the TD location, and the solution security, confidentiality, integrity, and availability.

Evaluating the solution **precision**, the measurement error for the different configurations is reported in Table II for the different measurement points, on average, and as an improvement with respect to the base configuration with a single access point (i.e., "Difference" in Table II).

Table 2: Measurements results: error in the location estimation for the different measurement points, in meters. The difference percentage is calculated from reference measurement (1 AP).

Point	1 AP	2 APs	3 APs
1	0.91	0.83	0.81
2	0.85	0.79	0.78
3	1.09	0.93	0.90
4	0.68	0.64	0.63
5	0.48	0.47	0.44
6	1.15	1.11	1.11
7	1.03	0.83	0.74
8	1.29	0.74	0.69
9	2.88	2.45	2.38
10	3.28	3.4	2.11
Average	1.36	1.22	1.06
Difference	Reference	10.63%	22.36%

As expected, as the number of reachable access points increases, the location estimation becomes more accurate, since the TD can use more anchors to perform distance calculations. As such, by increasing the number of available access points, the precision may be tuned to meet the system requirements and to reduce the probability of false room location to an amount irrelevant even for critical scenarios such as energy-production management and monitoring.

In order to evaluate the measurement error against the set threshold, a similar measurement campaign should be carried out in the target scenario (i.e., the power plant control room), which was not accessible to the authors, but this analysis may be considered a guideline for the one targeting a specific deployment.

Regarding **confidentiality**, Anyplace uses fully-encrypted communications using Transport Layer Security (TLS) between its API, internal components, and clients. Furthermore, devices are identified by their system certificates. This provides an additional layer of protection and makes Man-InThe-Middle (MITM) attacks more difficult, mitigating the risk of modifying the encrypted traffic. Finally, the building may be configured as private, protecting its information through a random string identifier (Universal Unique Identifier - UUID).

Considering the **availability** of this solution, a mirror service may be added to provide redundancy and/or to scale the solution. In this evaluation, Anyplace public productive environment was used, but it can also be downloaded and deployed in a local environment to add eventual mirror services and offer this way a configurable level of availability, and/or load balancing. Additionally, its components can be deployed in containers, which can provide means for automatic autoscaling in cloud environments.

Analyzing the solution's **security**, it supports the integration of a Web Application Firewall (WAF). The WAF is not integrated into the public instancing of Anyplace, but it may be integrated into the case of a local deployment. The WAF integration allows checking the network traffic directed toward the application (i.e., identifying potential attacks), and acting on it if needed. Still, Anyplace does not allow to set an expiration time for active sessions, neither in the case of inactivity, opening the possibility for a malicious third party to use an open session from an unattended device, contrary to the recommendations from the Open Web Application Security Project (OWASP) [35]. As the location is not the only authentication factor in the target solution, this flaw is considered minor.

Finally, we will analyze the solution **integrity**, i.e., the data exchanged among the different solution components is guaranteed to not be altered during transmission. The communication among system components is encrypted using the HTTPS protocol using TLS, and they require a valid certificate, signed by a Certification Authority (CA), protecting data transmission from MITM attacks and other transmission alterations. Finally, on every single node, redundancy may be set allowing for a configurable fault tolerance threshold.

6 Conclusions

In this paper, we analyzed the general problem of indoor positioning, reviewing the different characteristics offered by the different available solutions. We proposed a simple, yet agile and effective taxonomy to sort the solution space, a needed task when selecting the optimal solution for a specific use case. Then we analyzed different available solutions, offering practical recommendations for the scenarios in which they may be used, as well as the requirements they may or may not meet. This represents an addition to previous analysis for similar solution sets. As a following step, we present different real-world applications integrating indoor positioning. We analyze their main characteristics, requirements, and limitations. Finally, we evaluated a real scenario, i.e., location as an authentication factor for a management and monitoring system for energy production. For the analyzed scenario, we select and evaluate a solution in a real test environment. The selected solution was able to meet the precision, security, confidentiality, integrity, and availability requirements of the target system.

As a future work, it would be interesting to evaluate different real case scenarios, possibly involving different application domains. For each scenario, the corresponding system requirements must be evaluated, and a set of guidelines for solutions to be selected would be generated as output, together with guidelines for the configuration to be used. Furthermore, we would compare the performance of indoor positioning solutions across different types of environments, in order to be able to provide more specific guidelines for the technology selection and configuration.

Acknowledgment

This work was funded by EU Horizon 2020 research and innovation programme, Robinson Project, grant agreement N° 957752.

References

[1] F. Zafari, G. Athanasios, K. L. Kin, "A survey of indoor localization systems and technologies," IEEE Communications Surveys & Tutorials, **21**, 2568–2599, 2019, doi:10.1109/comst.2019.2911558.

[2] K. Nguyen, Z. Luo, G. Li, C. Watkins, "A review of smartphones-based indoor positioning: Challenges and applications," IET Cyber-Systems and Robotics, **3**, 1–30, 2021, doi:10.1049/csy2.12004.

[3] S. Li, R. Rashidzadeh, "Hybrid indoor location positioning system," IET Wireless Sensor Systems, **9**, 257–264, 2019.

[4] V. Nair, C. Tsangouri, B. Xiao, G. Olmschenk, W. Seiple, Z. Zhu, "A hybrid indoor positioning system for blind and visually impaired using Bluetooth and Google tango," Journal on technology and persons with disabilities, **6**, 2018, doi:10.1049/iet-wss.2018.5237.

[5] A. Catovic, S. Zafer, "The Cramer-Rao bounds of hybrid TOA/RSS and TDOA/RSS location estimation schemes," IEEE Communications Letters, **8**, 626–628, 2004, doi:10.1109/lcomm.2004.835319.

[6] 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 Surveys & Tutorials, **19**, 1327–1346, 2017, doi:10.1109/comst.2016.2632427.

[7] P. Bahl, V. Padmanabhan, "RADAR: An In-Building RF-based User Location and Tracking System," in Proceedings IEEE Infocom 2000 Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communication Societies, 775–784, 2000, doi:10.1109/infcom.2000.832252.

[8] W. Chen, K. Kao, Y. Chang, C. Chang, "An RSSI-based distributed real-time indoor positioning framework," in 2018 IEEE International Conference on Applied System Invention (ICASI), 1288–1291, 2018, doi:10.1109/icasi.2018.8394528.

[9] J. Yang, C. Yingying, "Indoor localization using improved rss-based lateration methods," in GLOBECOM 2009-2009 IEEE Global Telecommunications Conference, 1–6, 2009, doi:10.1109/glocom.2009.5425237.

[10] D. Niculescu, B. Nath, "VOR Base Stations for Indoor 802.11 Positioning," in MobiCom '04: Proceedings of the 10th annual international conference on Mobile computing and networking, 58–69, 2004, doi:10.1145/1023720.1023727.

[11] M. Kotaru, K. Joshi, D. Bharadia, S. Katti, "Spotfi: Decimeter level localization using wifi," in Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication, 269–282, 2015, doi:10.1145/2785956.2787487.

[12] C. Lim, P. N. Boon, D. Duan, "Robust methods for AOA geo-location in a real-time indoor WiFi system," Journal of Location Based Services, **2**, 112–121, 2008, doi:10.1080/17489720802415189.

[13] A. Fedotov, V. Badenko, V. Kuptsov, S. Ivanov, I. Struchkov, "Location measurement of an object using radio networks for Industry 4.0 applications," in E3S Web of Conferences, EDP Sciences, 2021, doi:10.1051/e3sconf/202126405060.

[14] J. Shen, A. F. Molisch, J. Salmi, "Accurate passive location estimation using TOA measurements," IEEE Transactions on Wireless Communications, **11**, 2182–2192, 2012, doi:10.1109/twc.2012.040412.110697.

[15] G. Hu, P. Feldhaus, Y. Feng, S. Wang, J. Zheng, H. Duan, J. Gu, "Accuracy Improvement of Indoor Real-Time Location Tracking Algorithm for Smart Supermarket Based on Ultra-Wideband," International Journal of Pattern Recognition, **33**, 1–27, 2019, doi:10.1142/s0218001420580045.

[16] M. Youssef, A. Agrawala, "The Horus WLAN location determination system," in Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services - MobiSys, 205–218, 2005, doi:10.1145/1067170.1067193.

[17] X. Zhu, W. Qu, T. Qiu, L. Zhao, M. Atiquzzaman, D. Wu, "Indoor intelligent fingerprint-based localization: Principles, approaches and challenges," IEEE Communications Surveys & Tutorials, **22**, 2634–2657, 2020, doi: 10.1109/comst.2020.3014304.

[18] L. Chen, I. Ahriz, D. L. Ruyet, "AoA-aware probabilistic indoor location fingerprinting using channel state information," IEEE Internet of Things Journal, **107**, 10868–10883, 2020, doi:10.1109/jiot.2020.2990314.

[19] F. Zafari, I. Papapanagiotou, K. Christidis, "Micro-location for Internet of Things equipped Smart Buildings," IEEE Internet of Things Journal, **3**, 96–112, 2016, doi:10.1109/jiot.2015.2442956.

[20] L. Bai, F. Ciravegna, R. Bond, M. Mulvenna, "A low cost indoor positioning system using bluetooth low energy," IEEE Access, **8**, 136858–136871, 2020, doi:10.1109/access.2020.3012342.

[21] S. S. Chawathe, "Beacon placement for indoor localization using bluetooth," in 2008 11th International IEEE Conference on Intelligent Transportation Systems, 980–985, 2008, doi:10.1109/itsc.2008.4732690.

[22] H. Kim, W. Seo, K. Baek, "Indoor Positioning System Using Magnetic Field Map Navigation and an Encoder System," Sensors, **17**, 2017, doi: 10.3390/s17030651.

[23] D. Almeida, E. Pedrosa, F. Curado, "Magnetic mapping for robot navigation in indoor environments," in 2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 1–8, 2021, doi:10.1109/ipin51156.2021.9662528.

[24] M. Frassl, M. Angermann, M. Lichtenstern, P. Robertson, B. J. Julian, M. Doniec, "Magnetic maps of indoor environments for precise localization of legged and non-legged locomotion," in 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, 913–920, 2013, doi: 10.1109/iros.2013.6696459.

[25] F. Rioult, S. Mecheri, B. Mantel, F. Kauffmann, N. Benguigui, "What Can Hawk-Eye Data Reveal about Serve Performance in Tennis?" in MLSA15 - Machine Learning and Data Mining for Sports Analytics workshop (ECML / PKDD 2015), Porto, Portugal, 36–45, 2015.

[26] A. Umek, A. Kos, "Validation of UWB positioning systems for player tracking in tennis," Personal and Ubiquitous Computing, **26**, 1023–1033, 2022, doi:10.1007/s00779-020-01486-0.

[27] K. Wong, "Low-cost Tennis Line Call System with Four Webcams," Department of Applied Physics, Stanford University, 2016.

[28] Compare Sport Tech, "Compare LPS," https://www.comparesportstech.com/compare-lps-tracking-systems, retrieved on February 14th, 2023.

[29] KINEXON, "Everything You Need To Know About Ball Tracking," https://kinexon.com/blog/everything-you-need-to-know-about-ball-tracking/, retrieved on February 14th, 2023.

[30] P. Blauberger, R. Marzilger, M. Lames, "Validation of Player and Ball Tracking with a Local Positioning System," Sensors, **21**, 2021, doi:10.3390/s21041465.

[31] A. Delfanti, The Warehouse, Pluto Press, 2021, doi:10.2307/j.ctv2114fnm.

[32] Federal Communications Commission, "OET Authorization Search Results," https://gov.fccid.io/2ABYU-RBT003, retrieved on February 14th, 2023.

[33] ACRELEC, "Table service," https://acrelec.com/table-service/, retrieved on February 14th, 2023.

[34] K. Georgiou, T. Constambeys, C. Laoudias, L. Petrou, G. Chatzimilioudis, D. Zeinalipour-Yazti, "Anyplace: A Crowdsourced Indoor Information Service," in Proceedings of the 16th IEEE International Conference on Mobile Data Management (MDM '15), 291–294, 2015, doi:10.1109/mdm.2015.80.

[35] OWASP, "Session Management - OWASP Cheat Sheet Series," https://cheatsheetseries.owasp.org/cheatsheets/Session_Management_Cheat_Sheet.html, retrieved on February 14th, 2023.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 41-59 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ
ISSN: 2415-6698

A Circuit Designer's Perspective to MOSFET Behaviour: Common Questions and Practical Insights

Ralf Sommer*,1,2, Carsten Thomas Gatermann[3], Felix Vierling[1]

[1]*Technische Universität Ilmenau, Electrical Engineering and Information Technology, Electronic Circuits and Systems Group, Ilmenau, 98693, Germany*

[3]*IMMS GmbH, Ilmenau, 98693, Germany*

[2]*Technische Universität Ilmenau, Electrical Engineering and Information Technology, Power Systems Group, Ilmenau, 98693, Germany*

ARTICLE INFO

Article history:
Received: 28 February, 2023
Accepted: 14 May, 2023
Online: 25 July, 2023

Keywords:
MOSFET
Equations
Teaching
Symbolic Circuit Analysis

ABSTRACT

Metal Oxide Semiconductor Field-Effect Transistors are commonly taught in courses for electrical engineers as they are the most common components within integrated circuits. However, despite numerous papers and books on MOSFETs, students still struggle with understanding their behaviour, particularly in the saturation region. This paper presents an expanded explanation of MOSFET behaviour, with a consistent and causal derivation of Level 1 MOSFET behaviour from a few equations, aimed at students without an extensive technological background. The paper provides illustrative explanations to help them understand MOSFET behaviour and addresses common students' questions, such as why the current is limited by charge carriers in the semiconductor substrate and why characteristic curves do not follow a parabolic curve in saturation. In addition to providing a comprehensive introduction to MOSFET behaviour from a circuit designer's perspective, this paper also offers valuable insights into interpreting AC parameters in modern MOSFET models. These parameters are often key to understanding and solving circuit problems related to small signal behaviour and frequency response, as demonstrated through various industrial application examples. These examples highlight how to bridge modern MOS models, such as the BSIM model, with MOS-Level 2 modelling, which is easily interpreted by users. By presenting these real-world examples, analysed by a symbolic analysis tool incorporating the BSIM to Level 2 AC model, this paper provides a practical and accessible approach to teaching MOSFETs and their applications in industry.

1. Introduction

This paper is an extension of the contribution presented at the International Conference on Synthesis, Modelling, Analysis and Simulation Methods, and Applications to Circuit Design (SMACD) conference [1]. The extension includes a comprehensive introduction and technological considerations before deriving the current equations consistently, making them easily understandable for undergraduate students and addressing frequently asked questions. The motivation for this paper were the discussions with students in a Metal Oxide Semiconductor Field-Effect Transistors (MOSFET) fundamentals course in 2019 – it became clear that they often struggle with understanding the behaviour of MOSFETs in the saturation region. Specifically, they frequently ask for a simple explanation of why the current is limited by charge carriers in the semiconductor substrate and why the characteristic curves do not follow a parabolic course in saturation. To aid in their understanding, many analogies have been used, such as "students as electrons" rushing in and out of a lecture hall through doors or buses transporting people from one location to another. However, these analogies have failed to adequately explain the behaviour, and often, lecturers resort to citing additional effects that are not described or are parasitic in nature. In fact, a simple explanation of the Gate-channel capacitance fixing the number of charge carriers by $Q = CV$ and the inability of charge carriers to switch polarity for not further following the parabola can logically answer both questions.

*Corresponding Author: Ralf Sommer, ralf.sommer@imms.de

Surprisingly, students told that even lectures held by technologists failed to provide satisfactory answers to these questions.

The industrial application examples have also been expanded, demonstrating how to bridge the gap between modern MOS models such as the Berkeley Short-channel Insulated Gate Field-Effect Transistor Model (BSIM) and MOS-Level 2 modelling that is easily interpretable for the user. This approach in combination with symbolic circuit analysis based on computer algebra [2] can solve even industrial-related circuit problems related to small-signal and frequency response behaviour.

Field-Effect Transistors (FETs) are one of today's fundamental electronic devices whose basic idea is to change the conductivity of a system by the influence of an electric field. The fundamental operating principle of a surface field-effect transistor was first proposed by Lilienfeld ([3], [4] and [5]) and Heil [6] in the early 1930s. Subsequently, Shockley and Pearson [7] studied this idea in the late 1940s. After the first device-grade Si-SiO_2 system was realized by Ligenza and Spitzer in 1960 by thermal oxidation [8], Atalla proposed the basic MOSFET structure based on the Si-SiO_2 system [9]. As a result, the first MOSFET was reported by Kahng and Atalla in 1960 [10]. A complete breakdown of the historical development of the MOSFET can be found in [11] and [12]. For the technology, application and device physics, reference can be made to [13] - [16].

The MOSFETs are themselves subdivided into a wide range of sub-categories. In the following, only the most widespread semiconductor technology for integrated circuits, the MOS technology (based on silicon), will be discussed. Hierarchically, the MOSFET is found as a subgroup of the MISFET, which in turn can be classified under the IG-FET. The abbreviations stand for:

- MOS ... Metal / Oxide / semiconductor,
- MIS ... Metal / insulator / semiconductor and
- IG ... insulated Gate.

In general, FETs are used with a MIS-structure. Replacing the insulator by silicon dioxide (SiO_2) one obtains the MOS-structure with SiO_2 being the Oxide. This type of insulation is preferred because oxides are characterized by the fact that they have a high dielectric strength, which should ideally be infinite in an insulator and also because they increase the Gate capacitance (high relative permittivity κ or ε_r), the importance of which will be discussed later in this paper, without increasing the leakage currents. The insulated Gate significantly reduces the power consumption of the FET (only leakage current flows into the Gate). In addition to oxides, the Gate insulation can also be technologically realized using other materials such as silicon nitride, polymers or a combination of different materials, as in the Metal Nitride Oxide semiconductor FET (MNOSFET). Therefore, MOSFETs are to be considered as a subgroup of MISFETs.

MOSFETs can be used in a wide range of applications due to their favourable characteristics. These include applications such as: analogue switching, high impedance amplifiers, microwave amplifiers and digital circuits (complementary MOS, abbrv. CMOS). Among the most appealing features of a MOSFET are significantly higher input impedances (compared to a bipolar junction Transistor, abbrv. BJT), negative temperature coefficient at high current levels, and the absence of forward-biased pn-junctions. The high input impedances are a result of the Gate insulation, which make them particularly suitable for standard microwave systems. Because of the negative temperature coefficient, the current drops as the temperature rises, resulting in a uniform temperature distribution across the device and preventing thermal runaway and breakdown. A MOSFET is thus thermally stable even in a parallel connection. Due to the lack of forward-biased pn-junctions, minority charge carrier storages cannot form. Between two switching states, unlike BJTs, MOSFETs do not have to compensate for stored charges resulting from diffusion tails, the cause of which is the average lifetime of minority carriers, allowing them to achieve much higher large-signal switching speeds. In addition, MOSFETs can be operated as a constant current source due to their nearly constant current-voltage characteristics in the saturation region. This property is particularly exploited in integrated circuit (IC) technology in the form of diode-connected MOSFETs, e. g. for operating point adjustment or current mirrors.

In the following, this work focuses only on the conventional Bulk (long-channel) MOSFET, i. e., no Silicon on insulator (SOI) MOSFETs or MOSFETs with multiple Gate electrodes are considered.

2. Ease of Use

First, the general structure of a MOSFET must be clarified. For this purpose, Figure 1 shows the typical structure of an n-channel MOSFET, which has three or four terminals: Gate (G), Source (S), Drain (D) and Bulk (B). The controllable current flow occurs between the Source and Drain by manipulating the channel by varying the Gate-Source voltage, which allows the current through the transistor to be selectively influenced according to the current-voltage characteristic. The individual terminals are contacted with regions or semiconductive layers that are diffused into the substrate. Starting from the bottom terminal, the MOSFET Bulk consists of a p-type (PMOSFET: n-type) silicon single crystal, and Source and Drain are formed by two n^+ regions (PMOSFET: p^+). The latter represent reservoirs for the charge carriers that would be minority charge carriers in the substrate. For this reason, like a BJT, a MOSFET can also be modelled by two antiserially connected diodes.

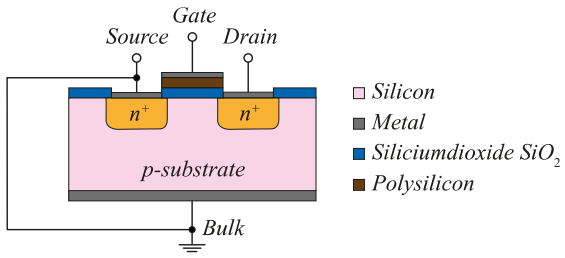


Figure 1: Conventional Bulk NMOSFET – General Structure

In contrast to the BJT, the Gate (analogue BJT: base) of a MOSFET is deliberately so wide that the space charge zones of the Source-substrate- and substrate-Drain- pn-junctions do not overlap. Thus, there is no transistor effect as with a BJT. For a voltage $V_{GS} < V_{th}$, one of the two pn-junctions (Source-substrate or substrate-Drain) on the Source-Drain path is always reverse biased, i. e., one of the two diodes is reverse biased regardless of the polarity of V_{DS}.

The "+" sign of the n^+ respectively p^+-areas means that these are heavily doped with impurity dopants. In most cases, the Gate in NMOSFETs is n^+-type polysilicon. As the name implies, it is not a Si-single-crystal but rather an imperfect crystal with a large number of grain boundaries. The Gate insulator is typically made of a thin SiO_2 layer. In stand-alone MOSFETs the Bulk is internally connected to Source to generate a common reference potential. Current can flow between Source and Drain, when a voltage with the correct polarity (i. e., + at the Gate pin) is applied between Gate and Source.

3. MOS-Structure

The heart of a MOSFET is formed by the MOS structure, which is obtained by omitting the Source and Drain regions of the conventional Bulk MOSFET. The substrate is often, as within this paper, given by silicon. Silicon is in the 4th main group in the periodic table of the elements proposed by Dmitri Mendeleev. According to [17], a chemical bond is particularly stable when the nearest electron gas configuration is reached (noble gas rule). With an electron configuration of: [Ne] $3s^2$ $3p^2$, silicon is still missing four electrons in the 3p-orbital to the next noble gas: Argon [Ne] $3s^2$ $3p^6$ (octet rule). Consequently, a Si atom in an (infinitely extended, i.e. no edges) silicon crystal forms four covalent bonds to the neighbouring Si atoms. Thus, next noble gas configuration is achieved, and the Si crystal is chemically stable. As mentioned before, the p-Bulk MOS structure, as used in NMOSFETs, consists of a Gate electrode (typically n^+ poly-Si), the Gate-Oxide as insulator (or dielectric as explained later) and a p-type substrate (Si Bulk) with a doping concentration N_A. The subscript "A" stands for "acceptor", since p-type doping means the intentional implantation of atoms with less valence electrons than Si into the Si single crystal of the substrate (e. g. boron, B, 3rd main group in the periodic table). Accordingly, the full four bonds to the neighbouring Si atoms in the crystal cannot be formed. As a consequence, the dopant atom is ionized (atomic body is negatively charged), and it provides a free-moving (diffusion processes in the semiconductor crystal) hole to the semiconductor structure. An explanation of the concept "hole" is given in Section 5.

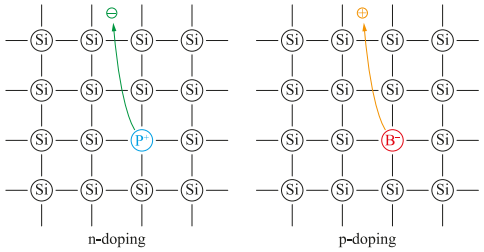


Figure 2: Silicon Crystal Structure with n and p Doping

In the ionized state the dopant atom is called an "acceptor ion" because the underlying atomic body has an affinity for electrons, i. e., the atomic body strives to compensate for the hole and achieve charge neutrality towards the outside. In contrast, an n-Bulk MOS structure of PMOSFETs consists of the Gate (typically p^+ poly-Si), the Gate-Oxide, and the n-type Si substrate with a doping concentration N_D, where the "D" stands for "donor", since atoms with more electrons than Si (like phosphorus, P, 5th main group in the periodic table) are incorporated into the Si single-crystal upon n-doping. With five valence electrons, phosphorus can form five covalent single bonds, unlike silicon. However, since only four neighbours are available for a potential bond in the semiconductor crystal lattice, the pentavalent phosphorus is also ionized (atomic body is positively charged). As a result, the crystal lattice has one free-moving electron (diffusion processes through the crystal lattice) available for charge transport. This is also the reason why these ions are called "donors". Figure 2 is intended to illustrate these relationships.

Because the Bulk contact is assumed to be grounded, the voltage drop across the MOS structure V_{GB} is equal to the applied potential at the Gate V_G, as shown in Figure 3. In the following, electrical potentials shall be labelled with only one letter in the subscript, e.g., V_G for the Gate potential, and voltages (potential differences) with two letters, for example V_{GS} for the Gate-Source voltage. Important for the voltages is the arrangement of the letters. The first letter indicates the start potential, and the second letter indicates the end potential, which makes the voltage arrow V_{XY} go from potential V_X to potential V_Y (according to the way of speaking). Therefore, the following relationship between voltage and potentials applies: $V_{XY} = V_X - V_Y$.

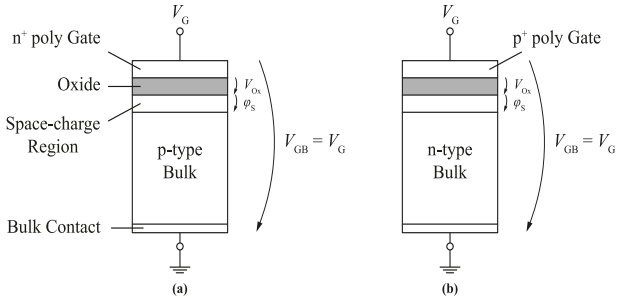


Figure 3: Two-Terminal MOS Structures. (a) p-Bulk MOS. (b) n-Bulk MOS. (adapted from [18]).

To derive the equations and gain a deeper understanding of the operating principle of the MOSFET, the type of transistor used must be specified. In this work, an n-channel transistor is taken as a model, since the NMOS is the standard model in lectures. The derivations and analysis can be performed analogously for a p-type, although some properties must be considered vice versa. This is due to the analogue but inverse MOS structure of the PMOSFET. This means for a PMOSFET, all voltages enter the equations with reversed polarity. In addition, the electrons contribute significantly to the charge transport (current flow) in an NMOSFET and the holes in a PMOSFET.

From a technological point of view, further insight into the physics of the MOS structure can be obtained by means of the energy band diagram. Figure 4 shows the energy band diagrams of the separated components of a NMOS structure under equilibrium conditions, i. e., Gate potential $V_G = 0$. The vacuum energy E_{vac} was chosen as the reference energy level. Each of the three regions has its own electron affinity χ, which is defined by the difference between E_{vac} and the conduction band lower edge E_C, and a work function ϕ, which results from the separation between E_{vac} and the Fermi energy E_F. Due to of the high n-doping of the Gate, the Fermi level is slightly above the conduction band edge E_C. The electron affinity is the same for both the Gate and the Bulk since both are Si, χ_{Si}.

The work function in the Gate ϕ_G corresponds approximately to the electron affinity χ_{Si}, which can be explained by the high doping. In the Bulk, the work function ϕ_B depends strongly on the Bulk doping N_A. The insulator (SiO_2) is characterized by the high band gap E_{Gox} (gap energy E_G = difference between conduction band lower edge E_C and valence band upper edge E_V).

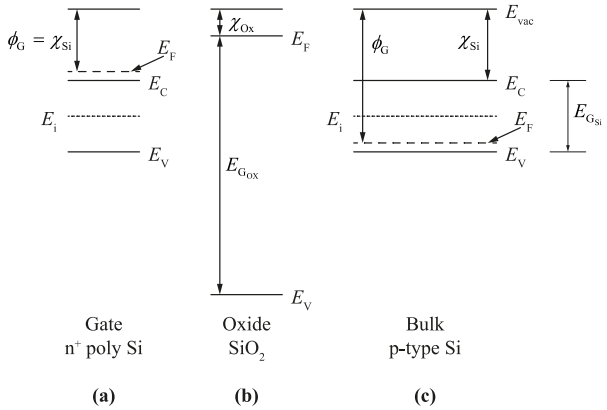


Figure 4: Energy Band Diagrams of the individual (separated) Components of a p-Bulk MOS Structure. (a) Gate (n^+-poly Si). (b) Insulator respectively Oxide (SiO_2). (c) Substrate (p-Bulk). (adapted from [18]).

When the three individual components are brought together (galvanic contact) to form a MOS structure and the equilibrium state is considered with no externally applied voltage ($V_G = 0$), the Fermi level of all three regions must be aligned in the same vertical position (same energy level) throughout the entire structure. To achieve this condition, the band diagram of the Bulk material can be held fixed at the contact points to the Oxide, while the rest of the structure can be pulled down until the Fermi energies E_F in the Bulk material and poly-Si match. This process results in band bending of the conduction and valence bands and of the intrinsic Fermi level E_i. To obtain a correct energy band diagram, the following points should be noted:

- In the substrate, the bands are bent only near the surface, i. e., the Si / SiO_2 interface, while far away from it the band relations remain unchanged. This is also the reason why the contact points to the Oxide should be held fixed during the band displacement.
- The relations between valence and conduction bands remain unchanged compared to the case where all three regions were separated.
- The intensity of the bending of the valence and conduction bands as well as the intrinsic Fermi level E_i are identical.
- Since the doping level of the poly-Si is so high, the bands at the Gate-Oxide interface bend only to a very small extent, so they can be neglected in the following.

The final band diagram of a p-Bulk MOS structure is shown in Figure 5 (a). The reason for the band bending is the work function difference between Gate and substrate (Bulk). To eliminate band bending, i. e., to make the bands flat, a certain voltage must be applied to the Gate. This specific voltage is called flat-band voltage V_{FB}. Note that for a p-Bulk MOS structure, V_{FB} is negative and for n-Bulk structures it is positive.

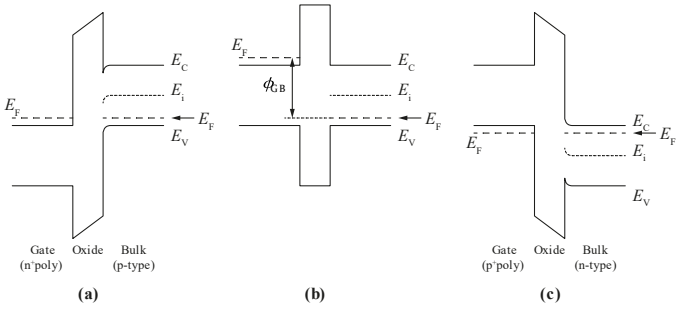


Figure 5: Band Diagrams of different MOS Structures. (a) p-Bulk MOS at $V_G = 0$.(b) p-Bulk MOS at $V_G = V_{FB}$ (c) n-Bulk MOS at $V_G = 0$ (adapted from [18]).

Another important parameter in MOS theory is the surface potential φ_S, which is related to band bending.

$$\varphi_S = \varphi_i(y=0) - \varphi_i(y \to \infty) = -\frac{E_i(y=0) - E_i(y \to \infty)}{e} \quad (1)$$

Here φ_i is the electrostatic potential introduced in Appendix B (cf. (31)) and y is the spatial coordinate in the direction of the perpendicular of the semiconductor surface (i. e. into the depth of the substrate). Therefore, the origin of the y-axis lies in the Oxide / substrate interface and is often referred to as the surface.

4. States of the (N)MOS Structure

To gain a deeper understanding of the operation of an NMOS structure (p-Bulk), the effects of different applied Gate voltages are first investigated. Throughout this analysis, the term "surface" (subscript: "S") is used to reference the interface between the Oxide (insulator) and the substrate (strong band bending). In contrast, the term "Bulk" (subscript: "B") is used to indicate that the analysis is performed far away from the interface, i. e., deep inside the substrate (little to no band bending). Additionally, all the different cases will be considered with a Source-Drain voltage equal to zero, $V_{DS} = 0$. In total, four relevant states can be identified:

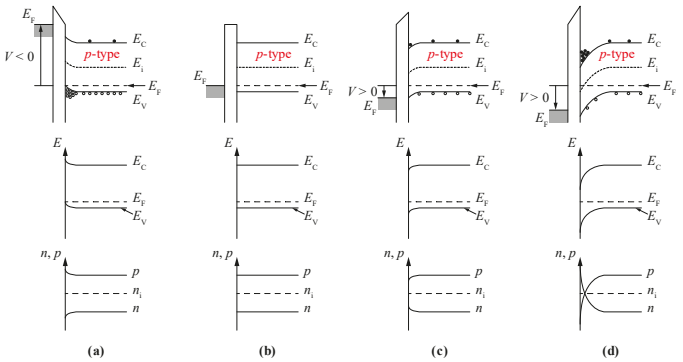


Figure 6: Band Diagram (top), Carrier Concentration (middle, Ordinate has log-scale) of Si Substrate for a p-Bulk MOSFET. (a) Accumulation, $V_G < 0$. (b) Flat-band Case, $V_G = V_{FB}$. (c) Depletion, $V_G > V_{FB}$. (d) Onset of strong Inversion, $V_G > V_{th}$. (modified taken from [18] and [19]).

Note, that from now on only the interior of the substrate is considered, with $x = 0$ being the interface between Oxide / substrate resp. Oxide / Bulk.

(1) $V_G < 0$ resp. $\varphi_S < 0$ (Accumulation case)

A sufficiently large negative voltage at the Gate (in relation to the Source) causes the energy bands to bend upwards near the substrate surface. This leads to a negative surface potential, $\varphi_S < 0$. Now the surface majority carrier (in case of a p-substrate (NMOS): holes) concentration p_S is larger than the Bulk majority carrier concentration p_B, and the surface minority carrier concentration n_S is smaller than its corresponding Bulk value n_B.

Within circuit technology and digital computing technology (CMOS circuit technology) this condition has no further meaning: In analogue circuits, MOSFETs are used primarily as amplifiers, and their behaviour is largely determined by the relationship between the input and output voltages and currents. Hence, most circuits operate in the saturation region, only a few in the linear region, but both require inversion. In digital circuits, MOSFETs are used primarily as switches, and the most important characteristics are their on / off states and the speed with which they can switch between these states. So, the accumulation region has no direct impact on these characteristics, and hence, it is also not relevant to digital circuit design.

(2) $V_{GS} = V_{FB}$ resp. $\varphi_S = 0$ (Flat-band case)

As the name suggests, the energy bands become flat at a less negative voltage, the flat-band voltage V_{FB}, i. e., the band bendings disappear.

(3) $V_{GS} > V_{FB}$ resp. $\varphi_S > 0$ (Depletion case)

When the Gate voltage is more positive than V_{FB}, the energy bands near the surface bend downward and the surface potential becomes positive, $\varphi_S > 0$. Thus, the holes are repelled from the surface (diffusion processes), which leads to the fact that only the uncompensated acceptor atomic hulls (ions) remain (firmly bound to the lattice structure of the substrate) and a space charge region is formed.

(4) $V_{GS} \geq V_{th}$ (Inversion)

Further increase of the Gate voltage enhances the depletion, i. e. n_S increases, p_S decreases and the thickness of the space charge region increases further. Once n_S equals p_S, the type of conductivity near the surface goes inverted. Consequently, in an NMOS with a p-substrate, the depletion region changes from p-conducting to n-conducting. This is called the onset of weak inversion. With the Gate voltage so large that the surface electron concentration is as high as the Bulk hole concentration, $n_S = p_B$, the onset of strong inversion is initiated. Now a channel (for NMOS: n-channel) is formed, which allows an effective current conduction. In other words, with the onset of inversion, the resistance of the channel is reduced. In addition, due to the local charge reversal, the effect of the antiserially connected pn-junctions (diodes; Source-substrate and substrate-Drain) is suppressed.

Note that the threshold voltage V_{th} is one of the most important electrical parameters of MOSFETs. It is defined as the Gate voltage that triggers the transition of the transistor from the off-state to the on-state. Because there is no uniform definition of the transition between the off-state and the on-state, several definitions of the threshold voltage are in use. In the Appendix A one can see the different definitions [18]. The authors restrict themselves within this publication to definition “Constant current”.

As soon as the Gate voltage V_{GS} exceeds V_{th}, a conductive channel is established. If the complete MOSFET, as shown in Figure 1, is now considered, a Drain current I_D can flow through the channel region. Therefore, the state of inversion is also called on-state and all other cases with $V_{GS} < V_{th}$ are called off-state. These terms originate from CMOS circuitry (digital logic).

5. Transition Process in the MOS Structure from the OFF-State to the ON-State

Within the off state, according to the mass action law

$$n \cdot p = n_i^2 , \tag{2}$$

with n being the electron concentration, p the hole concentration and n_i the intrinsic charge carrier concentration, a very temperature- and doping-sensitive charge carrier equilibrium is established, which is shown externally by the electrical charge neutrality. If the voltage between Gate and Bulk or Source (internal connection) is now increased so that the Gate receives a positive voltage, the charge carriers in the p-doped silicon shift. Electrons are increasingly drawn from the substrate below the interface between substrate and insulator (Oxide), which in turn causes the positive “charges” (holes) to be displaced from this region via diffusion processes into deeper substrate layers. It is important here that the electrons from the substrate do not leave it. Nor can they do so since the Gate is insulated by the Oxide. The phenomenon of displacement can be attributed to the effect of the electric field that builds up between the substrate and the Gate metallization. The result is the formation of a Gate-channel capacitance, with the Oxide acting as a dielectric.

Looking at (20), which will be explained in more detail in Section 8, it can be seen that the Drain current I_D, i. e. the current to be controlled, also depends on the Gate capacitance C_{ox}. Accordingly, with an increase in C_{ox}, an increase in I_D can be achieved. To enhance C_{ox} further, the insulator layer thickness was reduced in semiconductor technology. However, this caused several new problems with increasing miniaturization. On the one hand, the maximum field strength before an electric breakdown was reduced and on the other hand, the leakage currents increased drastically due to the tunnel effect. To ensure that C_{ox} can still be increased, and the insulator layer does not become too thin, the trend is toward high-k materials. The artificial word “high-k” is composed of the adjective “high” and the letter “k”. The “k” refers to the Greek letter kappa κ, which is the symbol used in English-speaking countries for relative permittivity. High-k materials (e. g. hafnium dioxide) are mainly used because they have a higher dielectric constant (relative permittivity ε_r), allowing for a thicker dielectric resp. insulator (Oxide) layer without compromising the device’s performance. This helps reduce Gate leakage current, improving the device's power efficiency, and Gate tunnelling current, improving the device's switching speed. In contrast, low-k materials are used to reduce parasitic capacitances in interconnects but are not used for the dielectric layer of very small MOSFETs due to their low dielectric constant. This is the reason why the permittivity of the Gate insulator has been increasingly addressed in recent years.

In the following, the Gate capacitance will be modelled as a plate capacitor with negligible edge effects. This insight gives rise to two explanations for the electron accumulation at the surface:

1) Equivalent charge principle:

Based on this principle, there are always the same number of charges (charges are always quantized) with opposite sign on the two electrodes of a capacitor. Therefore, if positive charges accumulate on the Gate electrode due to the positive (NMOS) Gate potential (conception: "suction" of free electrons from the e. g., highly doped poly-Si), negative charges (electrons) must accumulate at the interface between Oxide and substrate due to the Gate channel capacitance. Otherwise, the electrical charge neutrality of a capacitor cannot be preserved. Accordingly, the number of charge carriers which can accumulate at one electrode is limited by the other electrode (in this case the applied electrode voltage potential) and a proportionality constant which is impressed by the capacitance.

2) Influence:

With the "suction" of the electrons from the Gate electrode, the charge balance (neutrality) is disturbed in such a way that a positive centre of charge is formed. Since equal charges always repel each other (keyword: Coulomb or electrostatic forces), the conceptual "positive" charges (holes) are displaced into deeper substrate layers, leaving the negative charges (electrons). In addition, electrons are drawn out of the substrate under the interface via the same forces.

In an NMOS, the electrons in the p-substrate are the minority charge carriers while the holes are the majority charge carriers. Since the accumulation of electrons near the substrate / Oxide interface now occurs due to the above reasons, the holes within the interface are filled up by neighbouring places of the lattice. The consequence of the enrichment with electrons is noticeable by the depletion layer forming at the substrate / Oxide interface.

If there is a further increase in the Gate-Source voltage, the threshold voltage V_{th} is exceeded, and inversion comes into picture. Under these conditions, so many electrons are locally accumulated that the electrons become majority charge carriers. Thus, more electrons accumulate locally than the doping of the p-substrate can compensate. Because of the locally limited accumulation, the inversion is also locally limited. The so-called charge reversal starts. Incorrectly, it is often also referred to as a doping inversion, but this is wrong in the context of doping, since – in the case of an NMOS – there are no donor ions in the substrate. On the contrary, only acceptor ions can still be found in the p-doped substrate. The charge reversal in an NMOS creates a continuous, low-resistance n-conducting channel, which disables the function of the pn-junction between the n-doped islands and the substrate. Consequently, a current can now flow between Drain and Source: I_D. The channel region is only as large as the inversion zone and isolated from the substrate (channel: n-surplus, substrate: p-surplus) due to further pn-junctions forming (i. e. diodes in blocking direction). It follows that the MOSFET is voltage controlled, which makes it fundamentally different from the BJT, which is current controlled. Now it also becomes apparent where the terms NMOS and PMOS come from. The N and P refer to the conductivity of the forming channel. This is characteristic for the respective type of MOSFET, because an NMOS with an N-channel would immediately compensate all holes or positive charges due to the accumulation of electrons within the channel region and would therefore be non-conductive for these "positive" charge carriers. However, the electrons allow other electrons to pass. The opposite is true for the PMOS. Accordingly, the PMOS also requires a negative voltage at the Gate to form a p-type channel. At this point, a further difference can be identified in comparison to the BJT, since only one charge type contributes to the transport of the current depending on the substrate type. Therefore, the MOSFETs belong to the unipolar transistors and not to the bipolar transistors.

At this point it should be additionally noted that this is only a model, i. e., holes are not actual charge carriers, such as electrons or positrons. Holes are crystal lattice defects, where an electron is missing at the respective lattice place (e. g., consequence of doping with trivalent boron). This defect is interpreted as a conceptual positive "charge". Hence, holes cannot move by themselves. Instead, their mobility results from diffusion effects. For example, an electron from a neighbouring lattice place can fill up the vacancy (hole), creating a new hole at the lattice point where the electron came from. The electron has, so to speak, "jumped" one place further in the lattice. One can imagine this process as "wandering of the holes".

6. Deriving the MOSFET Equations (NMOS)

After explaining the individual states of the MOS structure, analytical equations describing the NMOSFET properties in the on-state (i. e., strong inversion case) are derived below. As mentioned above, the Gate-Source voltage causes to form a capacity between the substrate underneath the Gate and the Gate electrode, whereby – due to the positive potential at the Gate – electrons are attracted towards the Gate. The attracted and accumulated charge carriers create an induced n-layer. This yields a complete n-region from Source to Drain, i. e., a channel which forms kind of a low resistance tube where electrons can go from the Source electrode to Drain electrode. Thereby, the resulting electron flow builds up the desired current I_D through the MOSFET. A further increase of the Gate-substrate voltage attracts more electrons underneath the Gate and a higher current can be achieved. But why is the current limited by the Gate-substrate voltage and the number of electrons attracted from the p-substrate to forming the channel?

Looking at the MOSFET from a circuit-theoretic perspective, the MOSFET can be modelled as a voltage controlled current source (VCCS) as the Gate-substrate voltage resp. the Gate-substrate capacity influences the current between Source and Drain.

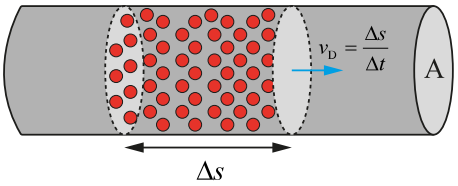


Figure 7: Electrons within a Conductor

The number of charge carriers N being stored in the tube shown in Figure 7 can be calculated using the equality

$$N = n \cdot A \cdot \Delta s = n \cdot V_{\text{Tube}}, \tag{3}$$

where n is the charge carrier density, A the cross-sectional area of the "tube" and Δs describes the way passed by the electrons in each Δt. Multiplying the cross-sectional area A with the way passed Δs gives the volume V_{Tube} in the second part of (3). The current through the MOSFET can be obtained by partially differentiating the charge by the time, i. e.

$$I = \frac{e \cdot n \cdot A \cdot \Delta s}{\Delta t} = e \cdot n \cdot A \cdot v_D , \tag{4}$$

In (4) e describes the elementary charge of an electron with $e = 1.6022 \cdot 10^{-19}$ As and v_D the drift velocity given as

$$v_D = \mu \cdot E , \tag{5}$$

The drift velocity is composed of the mobility μ and the electric field E applied to the tube.

7. Deriving the Current in the NMOS

The derivation of the current in the channel of a MOSFET is based on the Gate-channel capacity. This capacitance is the linchpin of the explanation of the saturation or limitation of the current: The electrodes of the "capacitor" are formed by the Gate on the one hand and the electrons in the channel underneath the Gate-Insulator on the other hand. It is important to mention that the electron layer, also called inversion layer, is in contact with the Source and the Drain at the same time (cf. Figure 8)

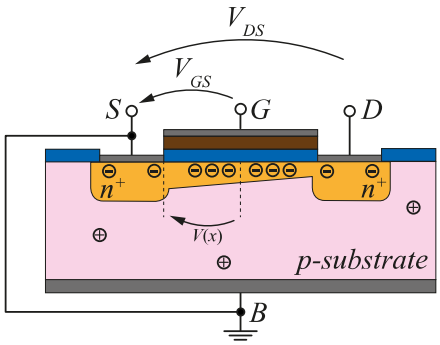


Figure 8: MOSFET with channel (NMOS)

The potential difference within the channel (and therefore the inversion layer) is determined by the potential difference between Source and Drain. The emerging capacity between Gate and channel can be modelled via the equation

$$Q = C \cdot V , \tag{6}$$

The real mathematical description of the charge within the Gate-channel capacity has a slightly shifted outcome as the threshold voltage V_{th} needs to be considered, since significant conductivity of the channel only occurs when this voltage is exceeded.

As previously indicated, the Gate-channel capacitance will be modelled as a plate capacitor with negligible boundary effects. The simple equation for the capacitance design applies. Adapting (6) to the current case gives

$$Q = C_{Ox} \cdot \left(V_{GC} - V_{th}\right) = \varepsilon \frac{A_{SiO_2}}{d_{SiO_2}} \left(V_{GC} - V_{th}\right) , \tag{7}$$

where V_{GC} represents the voltage across the Gate channel capacitance and $\varepsilon = \varepsilon_0 \cdot \varepsilon_{SiO2}$ the permittivity of the Oxide (insulator) layer. The area A describes the area covered by the Gate-insulator and can be approximated as

$$A_{SiO_2} = W \cdot L , \tag{8}$$

using the Gate width W and its length L, where length refers to the distance between the Source and Drain pn-junctions. We assume, as can be seen in Figure 8, that the channel length is equal to the Gate length. In reality, the effective channel length is assumed to be a little less than the full channel length. This is due, for example, to the fact that the ion implantation of the n^+ regions of the Source and Drain contacts cannot be focused precisely, resulting in a slight underdiffusion under the Gate (but still in the substrate), which leads to a reduction of the actual effective channel length. The thickness of the Oxide (distance between the Gate and substrate) is described by the parameter d_{SiO2} which is called TOX in Simulation Programs with Integrated Circuit Emphasis (SPICE).

As the capacitance is nearly constant, the number of charge carriers is enforced by the Gate-Source voltage (equivalent charge principle) and, hence, limited ($Q = C \cdot V$). Thus, it cannot be further increased by charge carriers from the battery.

To determine the current characteristic, the voltage $V(x)$ underneath the Oxide is introduced that varies between $V_S = 0$ und $V_D = V_{DS}$ as the Source potential is set as the reference potential. Therefore, the charge is also a function of the location x. If the voltage V_{GS} is not considered from the point of view of the Gate, but from the point of view of the channel, i. e., the counter-electrode, then due to the principle of equivalent charge, the voltage is reversed. Due to this, the voltage V_{GS} in the substrate points from Source to Drain and $V(x)$ is oppositely directed. In the next step, to determine the channel voltage at a certain location x, all voltages can be superimposed according to the superposition principle. Thus, the Gate channel voltage V_{GC} is composed of the difference between V_{GS} in the substrate and $V(x)$. Considering the infinitesimally small area

$$\mathrm{d}A_{SiO_2} = W \cdot \mathrm{d}x , \tag{9}$$

the charge can be evaluated to

$$\mathrm{d}Q = -\frac{\varepsilon_0 \varepsilon_{SiO_2} W}{d_{SiO_2}} \left[V_{GS} - V_{th} - V(x)\right] \mathrm{d}x , \tag{10}$$

The negative sign is due to electrons being the charge carriers in an n-channel MOSFET. As a preparation for the integration to get the Source-Drain current an essential case study must be executed: linear region and saturation region need to be looked at separately. The essential argumentation for these two regions is that the voltage term

$$V_{GS} - V_{th} - V(x) , \tag{11}$$

may never get negative or change its sign because otherwise the sign of the charge $\mathrm{d}Q$ would be reversed and the electrons being the charge carriers would suddenly become positrons.

It is also possible to look at the energy of a capacitor which cannot become negative, too. The current

$$I_{DS} = I_D = \frac{\mathrm{d}Q}{\mathrm{d}t} , \tag{12}$$

flows against the direction of travel of the electrons. Their speed is described by the drift velocity v_D. Henceforth, it shall be additionally assumed that the electron mobility in the channel is constant and does not depend on the electric field in the x-direction. Mobility reduction due to the vertical electric field will not be

considered here. This assumption works well for long-channel MOSFETs, but accuracy is reduced, especially when considering nanometre MOSFETs, since constant mobility and thus a linear velocity-(electrical)field characteristic is assumed throughout the entire device. This is remedied by the two-region MOSFET model, which will not be discussed here, but can be looked up in [18]. The minus sign is a consequence of the direction of the electric field, which points from Drain (positive potential) to Source (negative potential; grounded), where the electrons flow in the opposite direction. It is additionally noteworthy that the mobility μ_n is not the substrate mobility, but the effective mobility, which is explained in more detail in Appendix C.

$$v_D = -\frac{dx}{dt} = -\mu_n E = -\mu_n \frac{dV(x)}{dx}, \tag{13}$$

Since – as already mentioned in the introduction – this work is restricted to long channel MOSFETs, the longitudinal field along the channel is not sufficient to achieve a charge carrier saturation velocity. Under these circumstances, the velocity is coupled to and limited by the mobility (and thus within this model constant). Otherwise, the short-channel effects explained in more detail in [18] would have to be considered.

Using (12) and differentiating (10) regarding the time the differential equation for the current I_D,

$$I_D = -\frac{\varepsilon \cdot \mu_n W}{d_{SiO_2}}\left[V_{GS} - V_{th} - V(x)\right]\frac{dV(x)}{dx}, \tag{14}$$

can be obtained. Equation (14) can be solved by integrating from 0 to L along the channel and from V_{DS} to 0 because of the voltage directed in reverse to the flow of electrons. Changing the integration variable by multiplying (14) with dx the voltage $V(x)$ can be replaced by V as the integration incorporates solely the voltage difference and no longer the location along the channel.

$$\begin{aligned} \int_0^L I_D dx &= \int_{V_{DS}}^0 \left\{-\frac{\varepsilon \cdot \mu_n W}{d_{SiO_2}}\left[V_{GS} - V_{th} - V\right]\right\} dV \\ I_D \cdot L &= \frac{\varepsilon \cdot \mu_n W}{d_{SiO_2}}\left[\left(V_{GS} - V_{th}\right)V_{DS} - \frac{1}{2}V_{DS}^2\right] \end{aligned} \tag{15}$$

The factor

$$\mu_n \frac{\varepsilon}{d_{SiO_2}} = \mu_n \frac{\varepsilon_0 \varepsilon_{SiO_2}}{d_{SiO_2}}$$

is called K' or KP, whereby the fraction (without the mobility μ_n) is called C'_{ox} (effective oxide capacitance per unit area).

Looking at the Drain current I_D it is clear to see that the characteristic curve has a quadratic behaviour at the beginning (cf. Figure 10). The distribution of the charge carriers in the channel has an approximate triangular shape (at least shown in many textbooks, e. g. [20]), which forms the channel, as the electrons are pulled stronger towards the Drain the higher the Drain-Source voltage V_{DS} is. At the peak point of the characteristic curve, that is when

$$V_{DS} = V_{GS} - V_{th}, \tag{16}$$

the current does not sink but remains nearly constant.

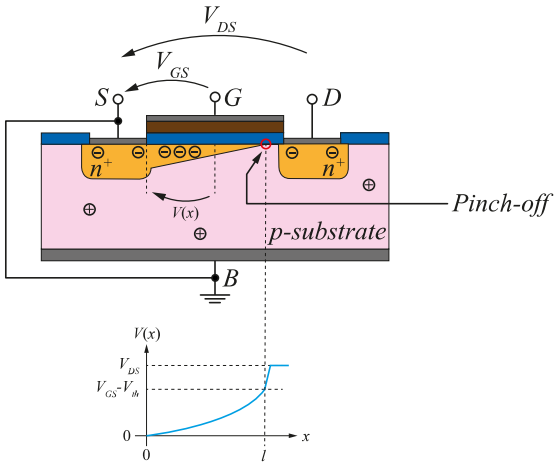


Figure 9: NMOSFET with Pinch-off

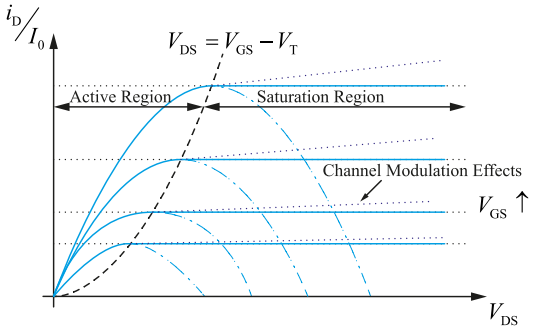


Figure 10: I_D as a Function of V_{GS} and V_{DS}

This peak is the point where the so-called Pinch-off takes place. No more charge carriers can reach for the Drain. If the Drain-Source voltage is further increased the length of the Pinch-off region varies in a way that only a certain amount of charge carriers can reach the Drain. The remaining electrons recombine as minority charge carriers with the holes in the p-substrate. Here one finds the missing arguments in most of the books and lectures: There is no inversion regarding the polarity of the charge carriers nor negative energy can be stored on a capacitor ((10), (11)). The resulting voltage for the current respectively the charge carriers remains constant. This is the explaining argument clarifying the trouble of the students in understanding the saturation region, and why it does not follow the parabola (15) after its peak (cf. Figure 10, blue dashed line).

$$V_{DS} = V_{GS} - V_{th} = V_{DS,Sat} \tag{17}$$

If $V_{DS} \leq V_{GS} - V_{th}$ (15) shows a linear correlation with V_{GS}. This section of the characteristic curve is therefore called linear, ohmic or active region. When $V_{DS} > V_{GS} - V_{th} = V_{DS,Sat}$ the voltage term in (15) needs to be set to $V_{DS} = V_{GS} - V_{th} = V_{DS,Sat}$. The equation for I_D in the saturation region can then be evaluated to

$$I_D = K' \frac{W}{2L}\left(V_{GS} - V_{th}\right)^2, \tag{18}$$

The current I_D stays nearly constant (regarding this model approach) and has a quadratic dependence on the Gate-Source voltage V_{GS}, which can be perfectly modelled by a VCCS. The quadratic dependence on V_{GS} can be seen in Figure 10 as the

distances between each blue curve within the saturation region are not uniform.

Within the device, the Pinch-off effect can be thought of as follows:

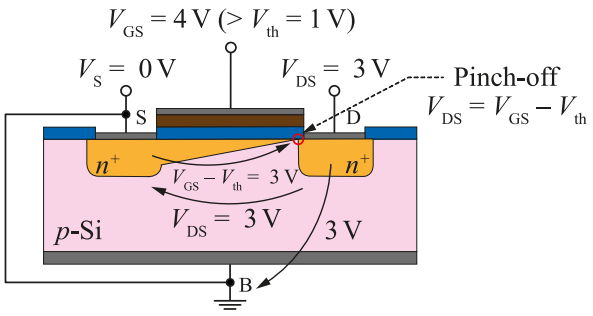


Figure 11: Calculation Example for the Superposition of the Voltage in the Substrate

The Pinch-off effect within the MOSFET device is a result of voltage drops within the substrate. This effect is illustrated by a small calculation example, where the voltage drops superpose within the substrate, thereby affecting the shape of the channel. This calculation example is only intended to illustrate the facts of the voltage drops and is by no means an accurate model. However, as can be seen very well, two voltages overlap within the substrate. Firstly, the one which is impressed by V_{DS} from Drain to Source and secondly, the one which drops due to the principle of equivalent charge between the interface substrate / Oxide and Source. The latter goes from Source to substrate / Oxide interface, since for an NMOS $V_{DS} > 0$ applies and thus the other plate of the plate capacitor (corresponding to the modelling of the MOS structure) must be negatively charged and thus the Source potential V_S = 0 V is larger. Consequently, the two voltages are oriented oppositely. They cancel each other out exactly when $V_{DS} = V_{GS} - V_{th}$ is fulfilled. This point is the so called "Pinch-off point", which occurs when there is a reduction in the relative voltage between the Gate and the substrate, as explained and shown in Figure 11. The point marks the x-coordinate at which the channel is pinched off and the strong inversion changes into a depletion zone. If the Drain voltage V_{DS} is now further increased (above $V_{DS,sat}$) at constant Gate voltage (V_{GS} = const.), the voltage drop from the Drain to the Source in the substrate predominates. As a result, the channel tends to be pinched off even sooner with respect to the x-coordinate. For $V_{DS} > V_{DS,sat}$, the Pinch-off point moves closer to the Source, but the voltage at the Pinch-off point remains the same at $V_{DS,sat}$. So, the saturation voltage is the voltage at which just such a channel and thus the strong inversion still exists in the substrate. This voltage cannot change with a fixed Gate voltage V_{GS}, since the charge carriers in the channel (minority charge carriers with respect to the substrate) are determined by the MOS capacitance C_{ox} – as already explained in the derivation. Thus, the total number of charge carriers (proportional to the overall charge) in the channel cannot change (otherwise the charge neutrality would be violated), whereby also according to $Q = C \cdot V$ with $C = C_{ox}$ = const. and Q = const. the voltage drop, that needs to be compensated for the onset of saturation (channel Pinch-off) by the Drain-Source voltage, cannot change.

In summary, the Pinch-off phenomenon in MOSFETs is influenced by the interplay between the Gate-Source and Drain-Source electric fields. The Gate-Source field creates the channel, while the Drain-Source field causes a voltage gradient along the channel that directly affects the electron concentration in the substrate. When V_{DS} is small and $V_G > V_{th}$ is fulfilled, the Drain-Source electric field is relatively weak, and the channel remains uniformly populated with electrons. As V_{DS} increases, the Drain-Source electric field becomes stronger, leading to a reduction in electron concentration near the Drain region (cf. Figure 12). When the Drain-Source field becomes strong enough to oppose the Gate-Source field at the Drain end, the channel is pinched off, and the transistor enters the saturation region. Beyond the Pinch-off voltage, the channel's resistance increases, and the Drain current I_D remains relatively constant despite further raise of V_{DS}.

It is important to note that as the Drain potential V_D increases, the voltage drop from the Drain to the Bulk also increases due to the internal connection of Source and Bulk. Consequently, the space charge region and the area of influence of the depletion induced by the Drain-pn-junction increase on the Drain side. This enlargement of the space charge zone results in more occupiable states for the same number of minority charge carriers, leading to a decrease in charge carrier density on the Drain side. Thus, the area of strong inversion (channel) decreases. Figure 12 shows the described situation, where the wedge-shaped region marks the area of strong inversion, where depletion comprises the area where more minority charge carriers are present than in the Bulk. Sufficient minority carriers are also present outside the strong inversion, i. e., in the depletion region, where most of the substrate majority carriers are compensated. Therefore, in the saturation region, the Drain current can still reach the Drain-side n^+-island. However, the electrical resistance within the depletion zone is higher than that within the local inversion (conductive channel). Thus, within the saturation, the resistance that the electrons "see" increases as they flow through the substrate from the Source to the Drain region (physical current direction). This shows on the one hand, why the current does not abruptly drop to zero and on the other hand, why the current remains almost constant within the saturation. Due to the constant voltage $V_{D,sat}$, the number of charge carriers arriving at the Pinch-off point remains the same. Consequently, despite the reduction of the effective channel length from L to L', approximately the same current I_D flows. Only if the shortened amount of the effective channel length is a substantial fraction of the channel length, an increase in the Drain current can be observed. As the channel length is reduced, the electric field in the channel increases, which causes the depletion region to expand towards the Source. This expansion reduces the effective length of the channel, and the Pinch-off voltage also reduces. As a result, the Drain current increases as the channel length decreases, assuming the voltage between the Source and Drain remains constant. This effect is considered within the channel length modulation in a more detailed model description, as discussed in the next section.

In summary to understand the difference between Pinch-off voltage and threshold voltage, the Pinch-off voltage is the voltage at which the depletion region around the Drain meets the depletion region around the Source, causing the channel to be "pinched off" and the Drain current to saturate. It is also sometimes referred to as the saturation voltage $V_{DS,sat}$. On the other hand, the threshold voltage is the voltage at which the MOSFET just starts to conduct current, with the channel beginning to form under the Gate. It is the minimum voltage required at the Gate to induce a channel and allow current flow between the Source and Drain. Hence, the threshold voltage V_{th} refers to the Gate-Source-voltage and Pinch-off voltage $V_{DS,sat}$ refers to the Drain-Source-voltage. They can be seen in the equations that describe the device's behaviour, such as the Drain current equation. In the saturation region, the Drain

current is essentially independent of the Drain voltage and is limited by the Pinch-off voltage. As mentioned before, this effect is particularly exploited in integrated semiconductor technology, since the MOSFET resembles a current source within saturation, i. e. almost constant current over a wide voltage range. Both voltages are important parameters in the MOSFET model and are influenced by various, such as the Gate-Oxide thickness and the doping concentration of the semiconductor material.

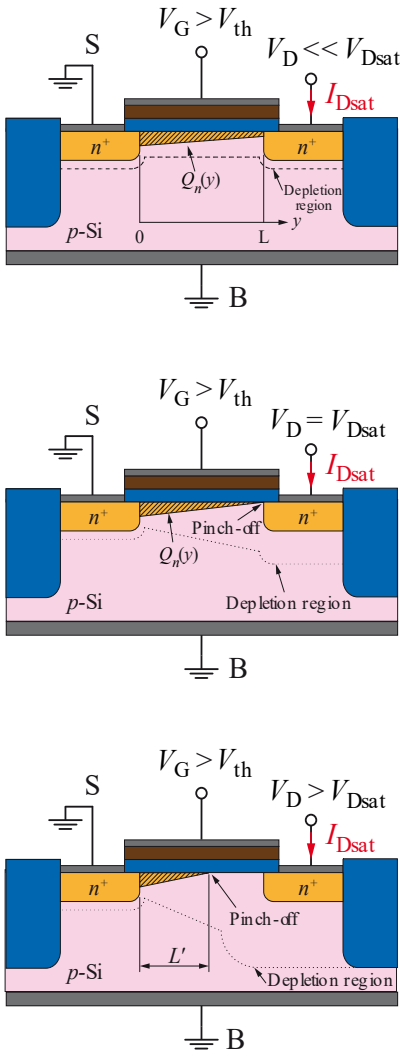


Figure 12: MOSFET operated (a) in the linear Region, (b) at Onset of Saturation, and (c) beyond Saturation (reduced effective Channel Length) with technical Current Direction. (taken from [19])

8. A more accurate Modelling of the Characteristic Curves

The ideas developed beforehand aimed to provide an easy derivation and intuitive understanding of the MOSFET characteristics. Another important parameter that has not been discussed so far is the threshold voltage V_{th}. It describes the Gate-Source voltage from where the inversion layer is built up and a significant current flow starts.

Using the equality (Shichman-Hodges resp. Level 1 model; no consideration of device random noise (thermal and flicker), sub-threshold behaviour, and high frequency effects)

$$V_{th} = V_{th0} + \gamma\left(\sqrt{2\phi_F - V_{BS}} - \sqrt{2\phi_F}\right), \tag{19}$$

With V_{th0} as zero threshold voltage, γ as the body effect parameter, and ϕ_F as the Fermi potential remaining constant, when the semiconductor is in balance. The zero-threshold voltage is the value of the threshold voltage V_{th} when $V_{BS} = 0$, i. e., the substrate has Source potential. The Fermi potential is the energy at which the Fermi-Dirac distribution has the value 1/2. Since this energy corresponds directly to a distribution function, it is (in very simplified terms) a measure of the number of free-moving charge carriers in semiconductor and is strongly influenced by the doping.

Equation (19) reveals an influence of the Bulk-Source voltage V_{BS} on the concrete value of the threshold voltage V_{th}. Consequently, the threshold voltage varies as a function of the Bulk-Source-voltage. This effect is commonly referred to as the "body effect". In the small-signal model (cf. Figure 19, Figure 20, Figure 24), this must be taken into account (Bulk-control). As mentioned before, usually the Bulk-potential is set to a fixed value (e. g. ground) or connected to Source, which reduces or compensates the body effect. It allows further control of the transistor and is especially important in integrated circuit technology. Less commonly, through hole technology (THT) MOSFETs also have the Bulk terminals pulled out of the package as a fourth pin.

Another important effect within the MOSFET and the Shichman-Hodges model is the channel-length modulation (CLM). Increasing the Drain-Source voltage over the value of $V_{DS,sat} = V_{GS} - V_{th}$ causes the Pinch-Off inducing a reduction of the effective channel length and therefore raises the current. Modelling the rising current when the Drain-Source voltage increases can be done with the channel length modulation factor λ. Within an empirical model, λ results from the common intersection point of the voltage axis and the extensions (tangents created in the saturation region) of all branches in the output characteristic field (I_D as a function of V_{GS}).

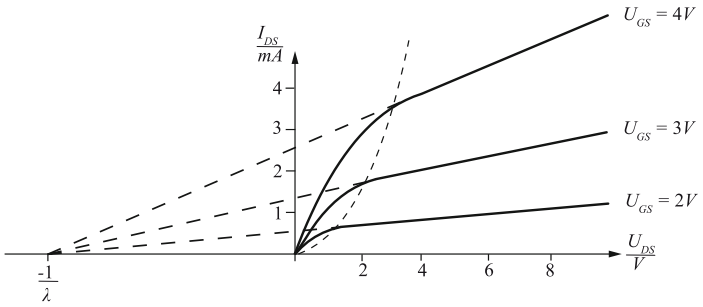


Figure 13: Channel Length Modulation, Asymptotes in Saturation Region meet at the Point of Early Voltage.

Using the intercept theorem, the equation for the characteristic curve of the MOSFET in saturation region can now be written as

$$I_D = \frac{\mu C'_{ox}}{2}\frac{W}{L}\left(V_{GS} - V_{th}\right)^2\left(1 + \lambda \cdot V_{DS}\right), \tag{20}$$

Considering the whole characteristic, the linear region also needs to be modified:

$$I_D = \frac{\mu C'_{ox} W}{L}\left[\left(V_{GS} - V_{th}\right)V_{DS} - \frac{V_{DS}^2}{2}\right]\left(1 + \lambda V_{DS}\right), \tag{21}$$

Typically, λ has a value of 0.05 V^{-1}. λ induces an r_{DS} in the alternating current (AC) model. The effect on the characteristics of the MOSFET resulting from the channel-length modulation can be compared with the Early effect in bipolar transistors. However, their origins are different, which is why they are not identical.

Up to now, only the case where $V_{GS} > V_{th}$ has been considered analytically for NMOSFETs. In this range, the presented model works very well, but it lacks accuracy for Gate-Source voltages smaller than V_{th}. When this condition is met, only a very small Drain current I_D (not zero) flows. This is called the subthreshold regime. For the Drain current in the subthreshold regime of an NMOSFET, a rather long derivation, which will not be explained in detail here but can be found in reference [18], results in the following expression:

$$I_D = \frac{\mu W}{L}\sqrt{\frac{\varepsilon_{Si} e N_A}{4V_T \ln\left(\frac{N_A}{n_i}\right)}}\cdot V_T^{\,2}\cdot e^{\frac{V_{GS}-V_{th}}{m\cdot V_T}}\left(1-e^{-\frac{V_{DS}}{V_T}}\right), \tag{22}$$

with V_T as thermal voltage: $V_T = k \cdot T / e$ and m as body factor:

$$m = 1+\sqrt{\frac{\varepsilon_{Si} e N_A}{4V_T \ln\left(\frac{N_A}{n_i}\right)}}\cdot\frac{1}{C_{ox}'}, \tag{23}$$

Because of the resulting impractical expression, a new widely used parameter has been introduced to characterize subthreshold MOSFET behaviour: sub-threshold slope S. This emerges from (22) assuming $V_{DS} >> V_T$, since under this condition the last exponential term converges to zero. S is then defined as the derivative of the logarithm of I_D with respect to V_{GS}.

$$S = \left(\frac{\mathrm{d}}{\mathrm{d}V_{GS}}\log I_D\right)^{-1}\begin{cases}=\ln 10\cdot m\cdot V_T\\ \approx 2.3\cdot m\cdot V_T\end{cases}, \tag{24}$$

The unit of S is typically given as mV/dec. The subthreshold slope gives an indication of how much the Drain current varies as a function of the Gate-Source voltage. For example, if S is given as 100 mV/dec and the Gate voltage is changed by 100 mV, the subthreshold Drain current will change by a factor of 10, i. e., a decade. The theoretical lower limit of S is 60 mV/dec, since an ideal case of $m = 1$ would then be encountered. In addition, it should be noted in the considerations that the equations only yield ideal values, since, for example, short-channel effects are neglected. These effects will not be discussed further in this paper. However, [18] can be cited as a suitable source for further reading.

What has also not been discussed in the previous MOSFET models are parasitic resistances. So far, it was always assumed that the voltages V_{GS} and V_{DS} applied to the external terminals are equal to those appearing at the intrinsic transistor. However, in reality there are series resistances between the intrinsic transistor and the external terminals, since the Drain current that flows through the channel must also pass through the Source and Drain series resistances R_S and R_D. These are caused, for example, by path resistances of the Drain and Source regions or contacting resistances. This leads to parasitic voltage drops, which is why the voltages across the intrinsic transistor are usually lower than the voltages applied to the terminals. The relationship between the external voltages (subscript: "ext") and the internal voltages (subscript: "int") are given by:

$$V_{DS,ext} = V_{DS,int} + I_D\cdot\left(R_S + R_D\right), \tag{25}$$

$$V_{GS,ext} = V_{GS,int} + I_D\cdot R_S, \tag{26}$$

The resistors have only a negligible effect on the MOSFET current-voltage characterization (I-V-characteristic) below the threshold voltage, i. e., in the subthreshold region. In contrast, in the on-state, i. e., beyond threshold, they lead to reduced Drain currents compared to the ideal case of the intrinsic transistor. This effect of reduced Drain current in the on-state can play a substantial role in nanometre MOSFET. For this reason, one of the design goals is to minimize R_S and R_D. In addition, this reduces the thermal losses of the transistors, which is why this can also be quite relevant for circuit design.

9. Models for Computer-aided and manual Analysis

Analog circuit designers are continuously facing the challenge to squeeze as much performance out of their circuits as the underlying semiconductor technology permits. Pushing a circuit topology to its technological limits requires extensive knowledge of previous designs and deep insight into both the intended functional behaviour of the device under construction and the parasitic effects that degrade its performance. Since conventional numerical circuit simulators cannot provide qualitative insight into the functional dependencies between circuit parameters and behavioural characteristics, it is often necessary to perform a manual analysis of the circuit. At the end of such an analysis the goal is to get an understanding of how the circuit works by interpretation of the mathematical (symbolic) expressions. In this context, the most important application for symbolic (manual) analysis is to gain design knowledge about undesired circuit behaviour observed in numerical simulations, e. g., in the form of resonance effects and instabilities due to parasitic poles or poor power-supply rejection behaviour.

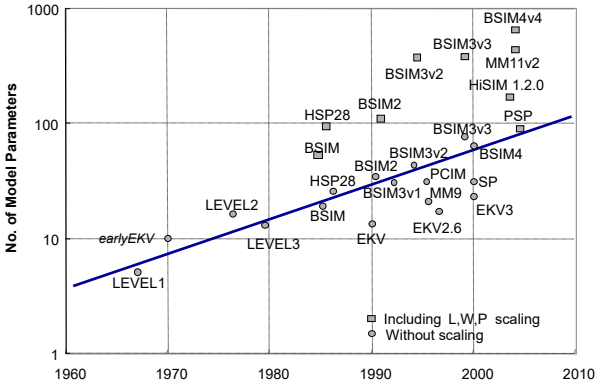


Figure 14: Evolution Model Parameters for MOS Transistors. (taken from [21])

For a fully comprehensive description of the internal operation of a MOSFET to be possible, three-dimensional complete quantum mechanical and atomistic simulations would be required. Since this currently seems unattainable due to a lack of computational resources, various device engineers and researchers have developed simplified abstract MOSFET models with different levels of complexity over the past decades. Because the share of CMOS chips is the largest, a lot of effort has been put into the modelling of MOSFETs, especially pushed by semiconductor manufacturers. Today, models (e. g., BSIM) with several hundred parameters are not uncommon (cf. Figure 14).

On the other hand, designers do not think in the BSIM, especially for the small-signal behaviour as shown in Figure 21. To gain insight into the operation of the circuits, many analogue designers use (15) or (18) for Direct Current (DC), since these equations describe the static behaviour of the MOSFET. Here, the fundamental transistor function of forming and controlling the channel current is modelled. As mentioned above, from a circuit point of view, this model can be represented by a VCCS. However, the dynamic behaviour is neglected and remains unconsidered. This can be remedied by inserting additional internal capacitances. The most important capacitances of a MOSFET and their summary (on the right) are shown in Figure 15.

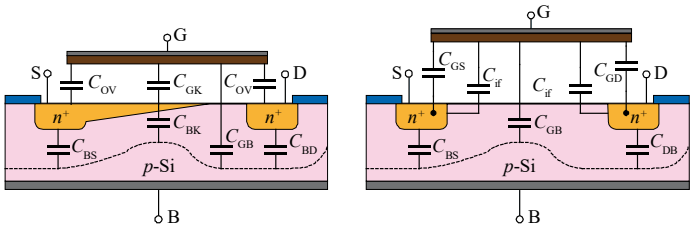


Figure 15: Sectional View of the MOSFET with the most important Capacitances, summarized on the right.

It should be noted that the effect of the individual capacitances is largely dependent on the operating range in which the transistor is located. For example, voltage-dependent capacitances can be defined depending on the Gate-source voltage V_{GS} (cf. Figure 16).

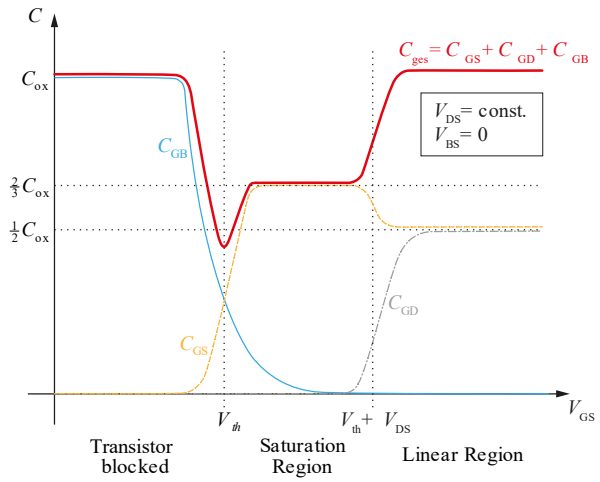


Figure 16: Compilation of the qualitative Capacitance Curves as a Function of the Gate-Source Voltage V_{GS}.

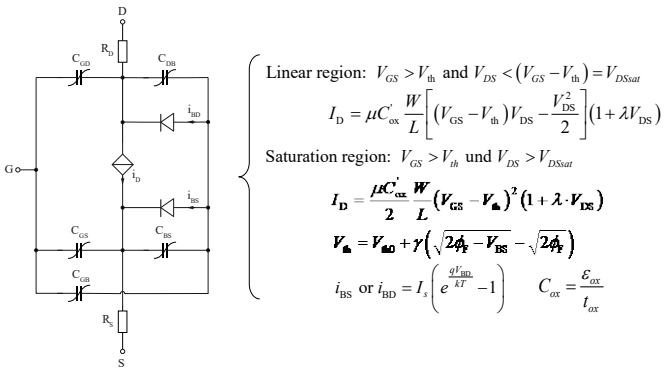


Figure 17: MOSFET Model with the static large-signal Equations.

With additional consideration of the pn-junctions from the Source or Drain to the Bulk, which are abstracted to junction diodes, as well as the terminal resistances, a complete (dynamic) large-signal equivalent circuit can be assembled (cf. Figure 17).

For example, after the operating point has been found using the large-signal equivalent circuit diagram, which roughly corresponds to the SPICE Level 1 model (cf. Figure 18), circuit engineers are then interested in the behaviour of the circuit when excited with signals of small amplitude and power. This is referred to as small-signal behaviour. The corresponding equivalent circuit diagrams can be derived from the large-signal model.

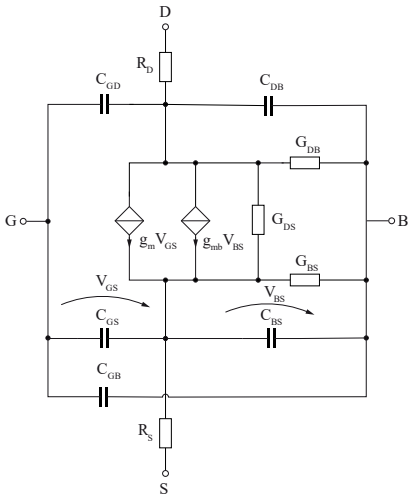


Figure 18: SPICE MOSFET Level 1 small-signal Model.

Using most manual analysis, the goal is to obtain good descriptions of the circuit functional behaviour with as little computational effort as possible. For this reason, the SPICE MOSFET Level 1 model is often already far too complex. A possible simplification for manual analysis is shown in Figure 19. In most cases, however, it will find application without g_{mb}, since Source is mostly connected with Bulk, which makes V_{BS} zero and therefore the corresponding controlled current source can be neglected. The transconductance g_{mb} describes the dependence of the Drain current on the Bulk-Source voltage, i. e., the previously explained body-effect.

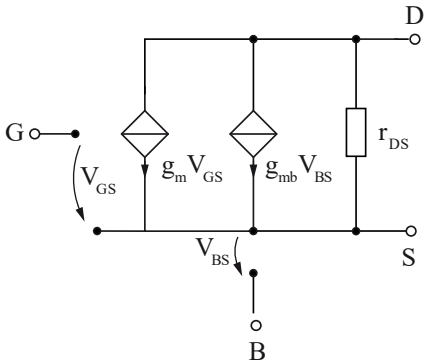


Figure 19: Simple static MOSFET small-signal Model for manual Analysis.

Herby the small-signal quantities are calculated as follows:

Linear / Triode range:

$$\text{NMOS}: V_{GS} > V_{th} \text{ and } |V_{DS}| \le |V_{DS,sat}|$$
$$\text{PMOS}: V_{GS} < V_{th} \text{ and } |V_{DS}| \le |V_{DS,sat}|$$
$$g_m = \frac{\partial I_D}{\partial V_{GS}} = \mu C'_{ox} \frac{W}{L} V_{DS}$$
$$g_{mb} = \frac{\partial I_D}{\partial V_{BS}} = \frac{\mu C'_{ox} \gamma}{2\sqrt{2|\Phi_F| - V_{BS}}} \frac{W}{L} V_{DS}$$
$$r_{DS} = \frac{1}{g_{DS}} = \frac{\partial V_{DS}}{\partial I_D} = \frac{1}{\mu C'_{ox} \frac{W}{L}(V_{GS} - V_{th} - V_{DS})}$$

Saturation range:

$$\text{NMOS}: V_{GS} > V_{th} \text{ and } |V_{DS}| \ge |V_{DS,sat}|$$
$$\text{PMOS}: V_{GS} < V_{th} \text{ and } |V_{DS}| \ge |V_{DS,sat}|$$
$$g_m = \frac{\partial I_D}{\partial V_{GS}} = \mu C'_{ox} \frac{W}{L}(V_{GS} - V_{th}) = \sqrt{2\mu C'_{ox} \frac{W}{L} I_D} = \frac{2I_D}{V_{GS} - V_{th}}$$
$$g_{mb} = \frac{\partial I_D}{\partial V_{BS}} = g_m \frac{\gamma}{2\sqrt{2|\Phi_F| - V_{BS}}} = \eta g_m$$
$$r_{DS} = \frac{1}{g_{DS}} = \frac{\partial V_{DS}}{\partial I_D} = \frac{1}{\frac{1}{2}\mu C'_{ox} \frac{W}{L}(V_{GS} - V_{th})^2 \lambda} \approx \frac{1}{\lambda I_D}$$

It is important to note again that most analogue circuit applications are operated in the saturation region because that is where the transistor acts like an almost ideal controlled current source.

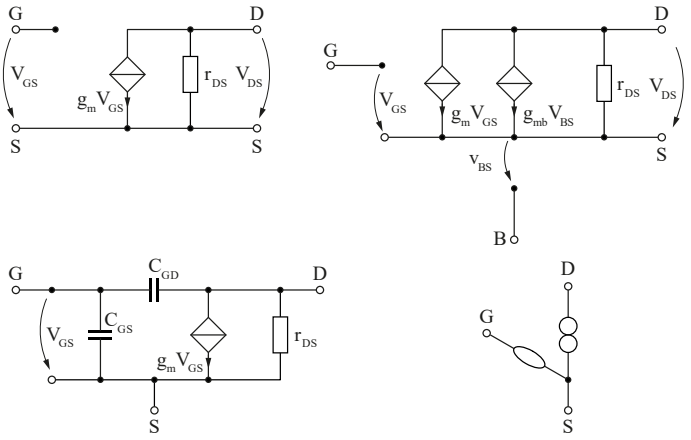


Figure 20: Small-Signal Models for hand calculations.

Other typical small-signal equivalent circuit diagrams may include those shown in Figure 20, including the g_m-only, g_m-r_{DS} and those extended by parasitic capacitances and/or the Bulk-Source induced g_{mb} if necessary and / or needed. Not well known is the Nullor model in the right lower corner, which is often beneficial and very efficient if dynamic analysis is to be performed.

10. Overview of MOS Models

Accurate modelling of MOSFET devices is critical for the design and simulation of electronic circuits, as well as for the optimization of device performance. Over the years, various MOSFET models have been developed to address the evolving requirements of semiconductor technologies. In this overview, we will discuss the development of MOSFET models, including Level 3, BSIM3, BSIM4, BSIM6, Enz-Krummenacher-Vittoz (EKV), and Penn-State Philips (PSP), highlighting their advantages and disadvantages. [22], [23]

MOS Level 1 to Level 3 are the earliest MOSFET models, utilizing simple equations to describe the basic operation of MOSFETs. The advantages of this model include simple calculations and low computational effort. In this article, we utilize the simpler Level 1 to Level 3 MOS models for a specific purpose: to obtain a straightforward understanding of MOSFET functionality and easily interpretable formula expressions for circuits. This is particularly important in the context of education, where students need to grasp fundamental concepts and relationships. Furthermore, circuit designers also rely on simple and easily interpretable relationships for qualitative circuit explanations, making the use of these basic models valuable in both teaching and practical applications. This is especially the case for small signal and frequency behaviour. However, the model suffers from inaccuracies for modern semiconductor devices, e. g., for short channel lengths and modern process technologies. It should be noted that these inaccuracies mainly refer to large signal behaviour and parasitic effects.

The BSIM3v3 model, specifically version 3.3.0, became an industry standard for accurately describing short-channel MOSFETs down to 180 nm. It accounts for various effects, such as Drain Induced Barrier Lowering (DIBL) that reduces threshold voltage with increasing V_{DS}, short-channel and narrow-channel effects impacting threshold voltage variations, mobility reduction due to vertical electric fields, and velocity saturation. Furthermore, the model considers channel length modulation, weak inversion conduction, parasitic resistances in Source and Drain regions, and "hot electron" effects that influence output resistance and threshold voltage over time.

The BSIM4 model significantly enhances the BSIM3v3 model, offering improvements in various aspects, such as better DC modelling accuracy, improved noise modelling crucial for radio frequency (RF) design, an enhanced capacitance-voltage (C-V) model for a wider range of operating conditions, and a new material model accounting for non-SiO_2 insulators, non-poly-Si Gates, and non-Si channels. These advancements make the BSIM4 model more versatile and suitable for modern process technologies and a broader array of applications compared to its predecessor.

BSIM6 is an extension of the BSIM4 MOSFET model and is a charge based symmetric MOSFET model with a charge-based core. BSIM6 has been designed to improve the accuracy of transistor simulations at nanometre scales and to model advanced device structures like FinFETs, nanowire FETs, and double-Gate MOSFETs. Some of the key differences between BSIM4 and BSIM6 include the modelling of carrier-induced voltage effects, addition of symmetry-breaking in mobility models, improved modelling of weak inversion region, and better modelling of back bias dependence. BSIM6 also provides more accurate modelling of sub-threshold slope dependence on Gate length, and improved noise models. Additionally, BSIM6 includes new parameters to

capture short-channel effects and improved models for DIBL and threshold voltage roll-off.

The EKV Model was developed for use in analogue and mixed-signal circuits, particularly in submicron CMOS technology. The model accounts for effects such as CLM, substrate resistance, and subthreshold conduction. The advantages of the EKV model include a good compromise between accuracy and computational effort, simpler mathematical formalism compared to BSIM models, and suitability for low-power applications. The model's disadvantages include not being as detailed as the newer BSIM models, especially for very short channels and cutting-edge process technologies.

The PSP Model is a compact MOSFET model intended for digital, analogue, and RF design, which is jointly developed in the early 2000s by NXP Semiconductors (formerly part of Philips) and Arizona State University (formerly at The Pennsylvania State University). [24] It was designed for improved accuracy, scalability, and predictability for advanced process technologies. The advantages of the PSP model include its physically based nature, good scalability, relatively compact structure, and less complexity than some newer BSIM models. However, the model may not provide the same level of detail as the latest BSIM models.

11. The Crux with the Small-Signal Models

As today's device models are very complex (cf. Figure 14), such as in the case of BSIM3 or higher, their structure is often based entirely on mathematical considerations instead of the underlying geometrical properties of the device. This makes interpretation of the resulting expressions more difficult, as the BSIM AC models involves transcapacitances (cf. Figure 21), i. e., differentiating VCCS instead of capacitances like C_{GS}.

These transcapacitances, which are based on charge derivatives on the various terminal voltages of the transistor may also become negative. This is difficult to be interpreted and, hence, might be confusing especially if stability problems are investigated being caused by parasitic capacitances.

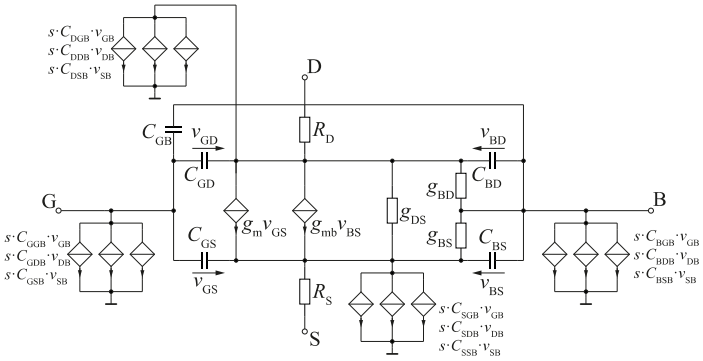


Figure 21: BSIM3 Small-Signal Model with Transcapacitances.

The operational amplifier (artificial term: OpAmp) depicted in Figure 22 is implemented in a 250 nm technology node and BSIM3v3 models were used to analyse its behaviour. The focus of this analysis was to study the sizing of the transistors for obtaining a good operating point and to investigate the AC response based on the small-signal parameters of the circuit.

In Figure 23, the output file provides information about the transcapacitances, which may sometimes exhibit negative values. While in Cadence Spectre, these values are denoted as Cxxxx (e. g., Cbgb), they are represented in Infineon's in-house simulator Titan as DQxDVyy depicted in the simulator output file in Figure 23. This naming convention represents the charge derivative at x with respect to the voltage yy. This naming is, thus, more informative, as it clearly identifies the transcapacitances as charge derivatives, and reduces confusion for those who may not be familiar with the concept of transcapacitances. It is worth mentioning that the BSIM small-signal models used to extract the transcapacitances are – in contrast to SPICE Level 1-3 – not publicly published, and different simulator providers may have different implementations. For example, some providers use branch-based derivatives, while others use node potential-based derivatives. This leads to the possibility of incompatible parameters between different simulators, adding to the difficulty of physically interpreting the impact of the transcapacitances on the frequency response of the circuit.

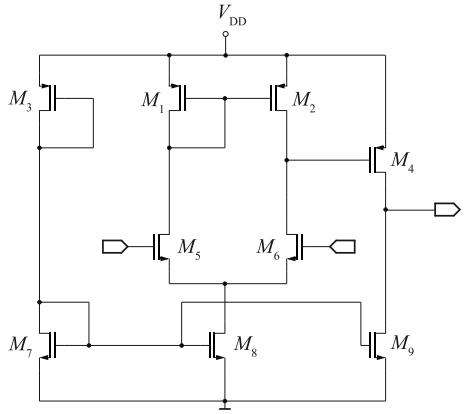


Figure 22: Operational Amplifier in 250nm Technology Node incorporating BSIM3v3 Models.

Despite these challenges, an AC analysis that provides interpretable information about the involved transistor capacitances is important for understanding the performance of OpAmps, especially in advanced technology. This information can help designers understand and adjust the frequency behaviour of their circuits, leading to more optimal designs.

```
**** MOSFETS

M_M1   MODEL: PMOD
      CBD          CBS          CBTOT        CDTOT        CGBOVL       CGD          CGDOVL       CGS          CGSOVL       CGTOT
      8.602E-015   6.737E-015   1.494E-013   2.194E-013   5.911E-026   2.593E-013   4.071E-015   4.322E-015   4.071E-015   3.197E-013
      CSTOT        DQBDVDB      DQBDVGB      DQBDVSB      DQDDVDB      DQDDVGB      DQDDVSB      DQGDVDB      DQGDVGB      DQGDVSB
      1.109E-014   -8.973E-014  -4.421E-014  -1.133E-016  1.382E-013   -1.072E-013  2.869E-016   -2.552E-013  3.115E-013   -2.511E-016
      DQSDVDB      DQSDVGB      DQSDVSB      GDS     *)   GM      *)   GMB     *)   GSTDS        GSTM         GSTMBS       IBD
      2.067E-013   -1.601E-013  7.745E-017   2.640E-008   1.716E-005   4.927E-006   -3.146E-021  -7.001E-021  -1.708E-021  0.000E+000
      IBS          ID           ISTRAT       VBS          VDS          VDSAT        VGS          VTH
      5.475E-015   1.742E-006   -1.601E-022  5.474E-001   5.474E-001   -1.685E-001  0.000E+000   -3.759E-001
...
 M_M5   MODEL: NMOD
      CBD          CBS          CBTOT        CDTOT        CGBOVL       CGD          CGDOVL       CGS          CGSOVL       CGTOT
      6.213E-015   8.105E-015   1.507E-013   1.350E-014   5.927E-026   7.285E-015   7.282E-015   4.382E-013   7.282E-015   5.165E-013
      CSTOT        DQBDVDB      DQBDVGB      DQBDVSB      DQDDVDB      DQDDVGB      DQDDVSB      DQGDVDB      DQGDVGB      DQGDVSB
      3.016E-013   -9.090E-019  -9.025E-014  -4.611E-014  2.707E-018   -1.647E-013  1.908E-013   -2.490E-018  5.020E-013   -4.309E-013
      DQSDVDB      DQSDVGB      DQSDVSB      GDS          GM           GMB          GSTDS        GSTM         GSTMBS       IBD
      6.914E-019   -2.470E-013  2.862E-013   5.099E-008   3.753E-005   5.700E-006   5.470E-011   1.156E-010   1.785E-011   -2.753E-014
      IBS          ID           ISTRAT       VBS          VDS          VDSAT        VGS          VTH
      -1.010E-014  1.742E-006   6.676E-012   -1.010E+000  1.742E+000   7.587E-002   6.399E-001   6.087E-001
...
```

Figure 23: Operating Point and small-signal Parameter Information for the BSIM3v3 MOSFETs (Titan Simulator).

To solve the problem, one can combine the formulas for calculating the SPICE 5-capacitance MOSFET AC equivalent circuit (cf. Figure 24) with the operating points calculated from the BSIM models to obtain reinterpretable capacitances.

However, for a hand analysis, the calculations are too complicated, so the use of a computer program is useful. Thus, such a conversion was implemented for the symbolic analysis tool Analog Insydes [2], which is based on the computer algebra program Mathematica.

Interestingly, it turned out that the SPICE Level 3 equations showed inaccuracies especially in the subthreshold area, so the formulas for the capacitances were revised using the fringing capacitances from Tsividis [16], among others. Fringing capacitances in a MOSFET are parasitic capacitances that occur at the edges of the Gate electrode, where the electric field extends beyond the edge of the physical Gate structure. These capacitances can affect the performance of the MOSFET by increasing the total Gate capacitance, which can impact the speed and power consumption of the device. An important update was the introduction of a partitioning factor (XPART) to distribute C_{ox} between C_{GS} and C_{GD}.

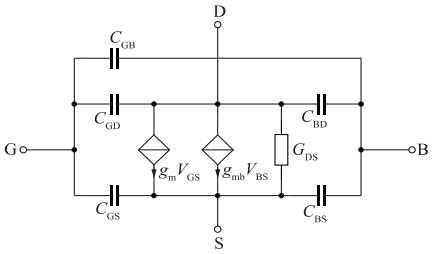


Figure 24: Simplified Level 2 SPICE 5-Capacitance MOSFET AC Model.

To show the steps involved in modifying the SPICE Level 2 5-capacitance MOSFET small-signal model (cf. Figure 24) to match the results of the BSIM small-signal model (cf. Figure 21), an industrial CMOS folded-cascode OpAmp (180 nm technology) is shown in Figure 25, and Figure 26 displays the frequency response of the OpAmp's open-loop differential-mode voltage gain. Here, the red curve shows the original simulation with the BSIM model, while the green curve shows the AC simulation performed with the SPICE Level 2 AC model, where the parameters were determined from the operating point using the full BSIM model. The agreement is very good for low frequencies but shows a significant deviation for higher frequencies.

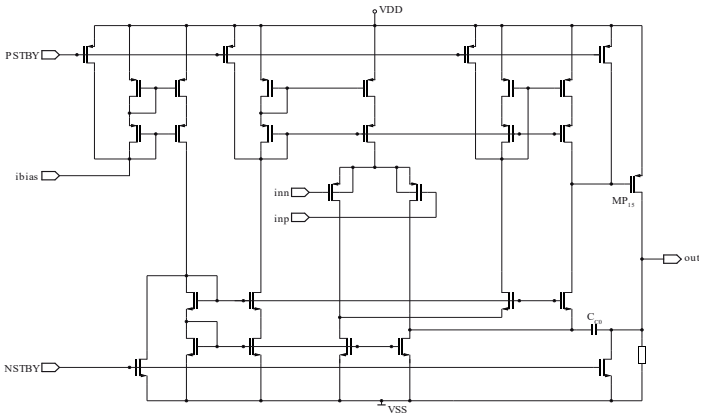


Figure 25: CMOS folded-Cascode Operational Amplifier.

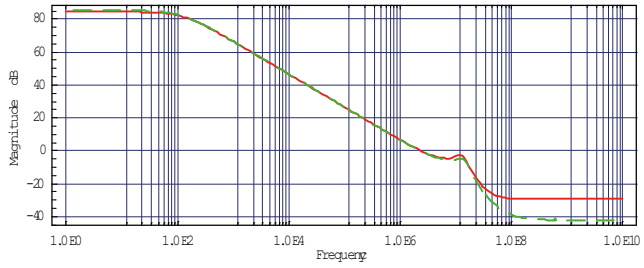


Figure 26: Frequency Response of the OpAmp's open-loop differential-mode Voltage Gain with BSIM (red) and Level 3-AC Model (green).

The deviation between BSIM model AC simulation and SPICE Level 2 AC analysis becomes even more significant for the power-supply feedthrough (PSF) characteristic in Figure 29 of the amplifier shown in Figure 27 (top-level circuit) and Figure 28 (transistor-level circuit).

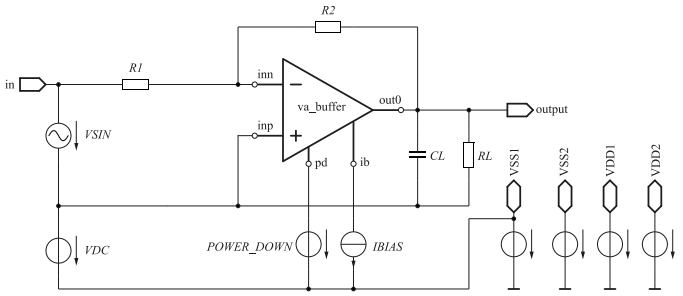


Figure 27: Top-level Circuit Schematic.

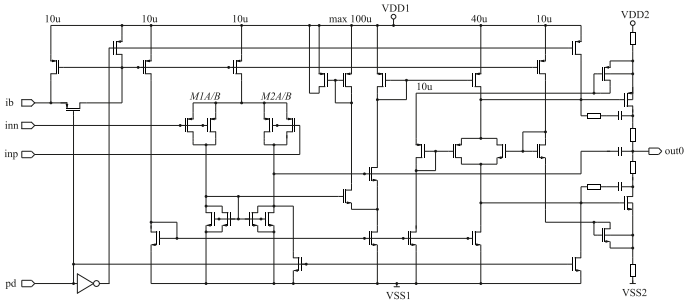


Figure 28: Transistor-level Circuit of the OpAmp with the Power supply rejection ratio (PSRR) Problem.

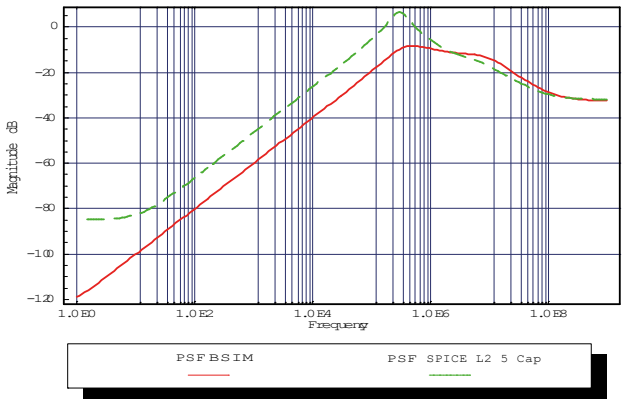


Figure 29: Frequency response of the OpAmp's power-supply feedthrough (PSF) with BSIM (red) and Level 3-AC model (green).

An analysis of the deviations led to a modification of the underlying SPICE level 2 AC model for the intrinsic Gate-Source capacitance C_{GS} and intrinsic Gate-Drain capacitance C_{GD} for the saturated MOSFET by introducing a partitioning factor XPART for C_{ox} between both capacitances.

Another problem is rooted back to the fact that SPICE Level 2 assumes perfectly pinched-off channel and no fringing capacitances. Hence, fringing capacitances C_{if} we added to the model in the subthreshold region (cf. Figure 15). Summing up all modifications the following small-signal equations were

implemented in Analog Insydes allowing symbolic analysis with interpretable results with an acceptable error.

Sub-threshold region: $V_{GS} \le V_{th}$

$$C_{GB} = C_{OX}\frac{0.5\gamma}{\sqrt{0.25\gamma^2 + V_{GB} - V_{GB}}} + CGBOVL$$
$$C_{GS} = C_{if} + CGSOVL;$$
$$C_{GD} = C_{if} + CGDOVL$$
$$C_{if} = W\frac{2}{\pi}\cdot\varepsilon_{GB}\ln\left[1+\frac{XJ}{TOX}\sin\left(\frac{2}{\pi}\frac{\varepsilon_{OX}}{\varepsilon_{Si}}\right)\right]$$

Linear region: $V_{GS} > V_{th} + V_{DS}$

$$C_{GB} = CGBOVL$$
$$C_{GS} = CGSOVL + \frac{2}{3}C_{OX}\left(1-\frac{(V_{GS}-V_{ON}-V_{DS})^2\left(4-\frac{3}{xpart}\right)}{(2(V_{GS}-V_{ON})-V_{DS})^2}\right)xpart$$
$$C_{GD} = CGDOVL + C_{if}\left(1-\frac{2C_{GD}}{C_{OX}}\right)$$
$$+\frac{2}{3}C_{OX}\left(1+\left(1-\frac{(V_{GS}-V_{ON})^2}{(2(V_{GS}-V_{ON})-V_{DS})^2}\right)\left(1+\frac{4}{3}(xpart-1)\right)-xpart\right)$$

Saturation region: $V_{th} < V_{GS}$ and $V_{GS} \le V_{th} + V_{DS}$

$$C_{GB} = CGBOVL$$
$$C_{GS} = CGSOVL + \frac{2}{3}C_{OX}\cdot xpart$$
$$C_{GD} = \frac{2}{3}C_{OX}(1-xpart) + CGDOVL + C_{if}\left(1-\frac{2C_{GD}}{C_{OX}}\right)$$

12. Application Examples using the modified SPICE Level 2 Model to solve industrial Circuit Problems in a CMOS Technology requiring BSIM MOS Models

The root cause of many industrial circuit problems is often traced back to their dynamic behaviour, particularly regarding frequency compensation and stability issues. Hence, symbolic extraction of poles and zeros is a crucial aspect of using symbolic analysis in the design of industrial integrated circuits. As designers wanted to understand the cause of the circuit problems it is essential for them to get interpretable results. Resulting in interpretable formulas, which can then be utilized to identify the appropriate frequency compensations for the circuits being analysed. This is achieved by adjusting certain circuit parameters that cause the poles to move in a manner that guarantees stability and eliminates peaking effects in the frequency response, maximizing bandwidth.

The Bode diagram (cf. Figure 26) of the OpAmp shown in Figure 25 displays a prominent peak at approximately 10 MHz, resulting from a pair of parasitic complex poles located close to the imaginary axis. Note, that in the frequency domain complex pole pairs can cause resonance peaks and phase shifts, which can affect the bandwidth and stability of the system. In the time domain, complex pole pairs can lead to oscillatory and damped responses, which can affect the settling time and overshoot of the system. The objective is to extract a simplified symbolic formula for these poles, to identify the components that significantly contribute to the peak. Therefore, it is necessary to shift the conjugate complex pole pair away from the imaginary axis, i. e., either by decreasing the imaginary part or by increasing the real part. The goal is to push the pole into the region bounded by the 45° axes in quadrants 2 and 3 of the root locus plot, cf. Figure 32 (dashed grey lines). The complexity of this issue is evident from the fact that after expanding the model, the netlist comprises of 321 primitive components, leading to a system of 29×29 modified nodal equations. Instead of the BSIM3 AC model, the modified SPICE Level 2 AC model was utilized.

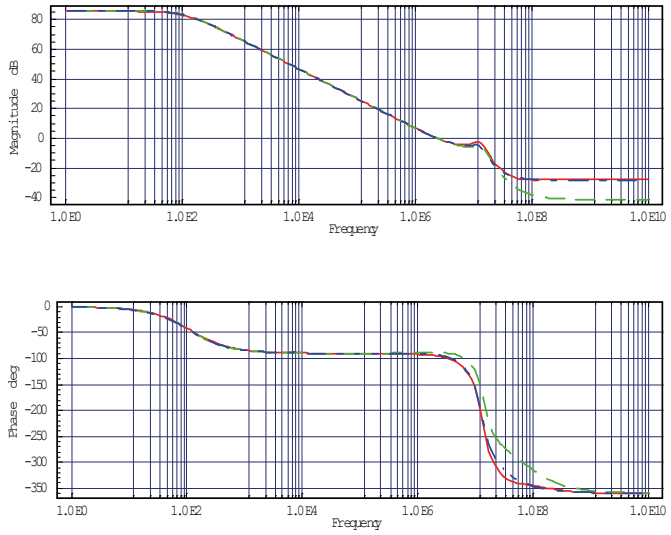


Figure 30: Frequency Response of the OpAmp's open-loop differential-mode Voltage Gain with BSIM (red) and Level 2-AC Model (green) and modified Level 2-AC Model (blue).

The differential-mode voltage transfer function, shown in Figure 30, has 19 poles and 19 zeros. In its fully expanded symbolic form, it would comprise more than a multitude of $5 \cdot 10^{19}$ product terms. Hence, to understand the cause of the peak, the symbolic extraction of the pole pair at $s = (-2.1 \pm 8.3j) \cdot 10^7$ (with j as the imaginary unit) is essential. To symbolically determine the pole pair, a symbolic approximation algorithm [25] was applied, yielding the following formula:

$$-\frac{(C_{C0}+C_L)g_{m\$MN6}}{2C_{C0}C_L} \pm \frac{\sqrt{C_{gs\$MP15}g_{m\$MN6}\left(C_{gs\$MP15}(C_{C0}+C_L)^2 g_{m\$MN6} - 4C_{C0}^2 C_L g_{m\$MP15}\right)}}{2C_{C0}C_L C_{gs\$MP15}}$$

The formula shows that, given a fixed load C_L and operating conditions, an increase in the Gate-Source capacitance of PMOS transistor MP15 $C_{gs\$MP15}$ will result in a reduction of the pole pair's imaginary components. It should be noted that altering the compensation capacitance C_{CO} will not affect the resonance peak, as its contribution is in the same order of magnitude in the numerator and the dominator in the square-root expression that yields the imaginary part of the pole pair.

Hence, reducing the imaginary part can be achieved by adding a shunt capacitor between the Gate and Source terminals of MP15 (cf. Figure 31). Figure 32 presents a root locus plot of the amplifier, calculated from the original (unsimplified) system with BSIM models as is varied from 1 pF to 10 pF. The plot confirms the

validity of the conclusion drawn from the approximated symbolic pole expression with physical interpretable small-signal capacitance.

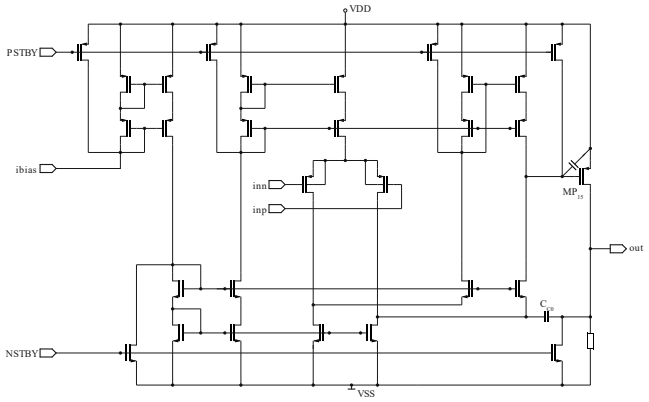


Figure 31: CMOS folded-Cascode OpAmp with Compensation Capacitor derived from symbolic Analysis.

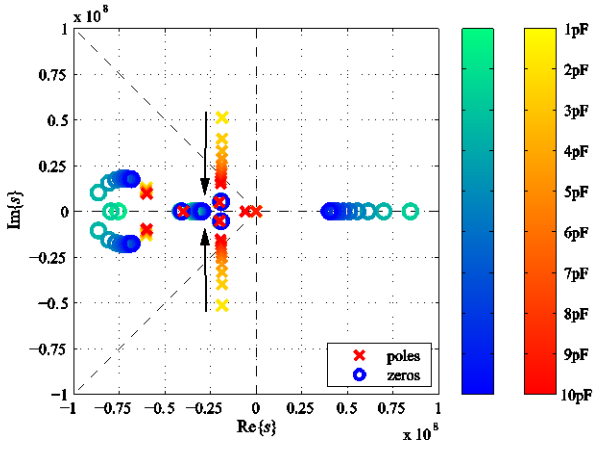


Figure 32: Root locus Analysis of the Voltage Transfer Function as $C_{GS,MP15}$ is swept from 1 pF to 10 pF.

The second example for the application of Analog Insydes [2] in industrial circuit design using the modified SPICE Level 2 model is the CMOS OpAmp, shown in Figure 25. The circuit operates as an inverting amplifier (in Figure 27).

The AC PSF characteristic of the circuit, depicted in Figure 29, demonstrates an unexpectedly large PSF value of over at 1 kHz frequency, which was predicted to be better than based on rough manual calculations by the designer. PSF occurs when unwanted signals from the power supply source leak into the output of an electronic circuit, hence it is the transfer function from the voltage supply to the output. Power supply rejection ratio (PSRR) is a measure of how well a circuit rejects ripple coming from the input power supply at various frequencies. It is defined as the ratio of the change in supply voltage to the equivalent (differential) output voltage it produces. To identify the reasons for the discrepancy and enhance power-supply rejection, Analog Insydes was again utilized for a symbolic analysis of the circuit behaviour. Solving the unapproximated equations symbolically for the transfer function from V_{DD} to the output node would result in a complex expression with over 10^{27} terms. Hence, Analog Insydes' symbolic approximation algorithms are utilized to estimate the PSF behaviour at low frequencies. With an error limit of 10 % at $f = 1$ kHz, we quickly obtained the following straightforward description of the amplifier's PSF characteristic, using only a few seconds of central processing unit (CPU) time.

$$V_{\text{outbyVDD}} = -2C_{\text{gb1}}R_2 \cdot s \cdot V_{\text{DD}}, \tag{27}$$

Additionally, this formula of the PSF can avoid misinterpretation of the real signal flow. The signal path for the PSF is not through the amplifier, but instead, it goes in reverse over the outer feedback resistance R_2 to the Bulk of the input pair. The designer neglected these Bulk capacitances in his manual calculations, as he believed them to be insignificant. That also explains that optimizing the OpAmp does not solve the PSF or PSRR problem, and oversimplifying the problem beforehand prevented the true cause of the unwanted circuit behaviour from being discovered. Once the true reason due to (27) was found, introducing a separated well for the input pair, which reduces the influence of the Gate-Bulk capacitance, solved the issue so that the PSF could be reduced. More details on the methodology especially on the generation of approximated symbolic formulas can be found in [25] and [26].

13. Conclusion

The study of MOSFETs is an important part of electrical engineering education as they are the most used components in integrated circuits. Therefore, a systematic and logical explanation of the behaviour of MOSFETs is necessary to facilitate comprehension among students in the context of circuit design. Despite the availability of substantial resources on MOSFETs, students continue to face difficulties in comprehending the concepts of charge carrier limitation and saturation. To address this challenge, a systematic and logical derivation of the Level 1 behaviour of MOSFETs is presented, incorporating simple equations and clear illustrations, to clarify the questions commonly raised by students without adding complexity through the introduction of additional semiconductor effects. Additionally, the use of modern standard MOSFET models, such as BSIM, may also pose difficulties in interpretation. A special problem is rooted back to the fact that – in contrast to the well-known AC models of SPICE Level 1 to 3 – BSIM and PSP small signal models are not published and that they include transcapacitances due to their charge-based modelling instead of physical capacitances. The acquisition of knowledge and insight regarding the behaviour of circuits is imperative for both designing circuits and resolving issues such as instability, ringing and other related phenomena. Symbolic analysis is a useful tool to facilitate this understanding, however, it is important to ensure that the formulas are comprehensible. The utilization of BSIM model and its parameters in conjunction with a modified SPICE Level 3 AC model can help in bridging the divide between technical formulas and their practical interpretation and applications. This methodology can enhance the understanding of MOSFET behaviour and its influence within the circuit design tasks for both students and experienced designers alike.

14. Acknowledgement

We would like to thank Dominik Krauße and Wladek Grabinski for providing valuable material for this contribution as well as Marius Steindel for his support in creating the illustrations and diagrams that helped to clarify and enhance our explanations of MOSFET behaviour.

15. Appendix A

Definitions of the threshold voltage V_{th} [18]:

Classical V_{th}

Conventionally, the threshold voltage is defined as the Gate voltage that must be applied to the MOS structure to reach a carrier concentration near the substrate-insulator (Si / SiO_2) interface that is equal to the majority carrier concentration in the substrate (i. e., far from the interface). In the case of an NMOS with applied threshold voltage and a p-doped substrate, this means that the electron concentration near the interface would be equal to the hole concentration in the Bulk. This condition marks the onset of the strong inversion. In particular, the band bending in the energy band model reaches a value twice as large as the Bulk potential at the location of the interface.

However, this definition is problematic due to the experimental measurement uncertainties of carrier concentration and band bending at the interface.

Extrapolated V_{th}

As can be seen in the transfer characteristic of a MOSFET, i. e., Drain current I_D plotted as a function of Gate voltage V_{GS}, the characteristic behaves almost linearly over a wide Gate voltage range in the on-state. Extrapolating the linear part of the curve to a current $I_D = 0$, the threshold voltage V_{th} can be obtained. Because the slope is not constant everywhere, consequently the threshold voltage also depends on V_{GS}. The point at which the extrapolation was applied decides the intersection location with the x-axis. To avoid this ambiguity, the point at which the slope is maximum is often selected for the linear extrapolation.

Constant current

A more pragmatic approach to defining the threshold voltage V_{th} would be to have a certain predefined current flow through the transistor at the Gate voltage $V_{GS} = V_{th}$. During this, two variants can be formulated:

The first formulation is independent of the Gate length:

$$I_D(V_{GS} = V_{th}, V_{DS}) = C_1 \cdot W , \tag{28}$$

where C_1 is the specified current threshold per unit Gate length (values around 10^{-7} A / µm) and W is the Gate width.

A second variant additionally considers the Gate length L and is given by:

$$I_D(V_{GS} = V_{th}, V_{DS}) = C_2 \cdot \frac{W}{L} , \tag{29}$$

with C_2 as a constant current specification (values around 10^{-7} A).

Constant sheet concentration

Due to the proportional dependence of the Drain current in a MOSFET on the inversion layer sheet concentration, the constant C_1 from (28) can be converted to such, as an inversion layer sheet concentration threshold.

16. Appendix B

The relation between energy E and potential φ is given by:

$$\varphi = -\frac{E}{e} , \tag{30}$$

where e is the elementary charge. Thus, the potentials corresponding to the intrinsic Fermi level E_i and the Fermi level E_F can be defined as the electrostatic potential φ_i and the Fermi potential φ_F, respectively.

$$\varphi_i = -\frac{E_i}{e} \text{ and } \varphi_F = -\frac{E_F}{e} , \tag{31}$$

17. Appendix C

The charge carrier mobility within the Si MOSFET channel can be significantly smaller than the Bulk mobility. The transport properties of the carriers are strongly influenced by the surface, in particular scattering mechanisms play a role. These can be, for example, surface acoustic phonon scattering or surface roughness scattering. Furthermore, the mobility shows a dependence on the electric field perpendicular to the surface. This can be attributed to inversion, since at higher Gate-Source voltages the electric field perpendicular to the surface becomes larger and more electrons are drawn below the surface. Under these circumstances, the electrons interact more and more with each other, which only further enhances the scattering effects at higher Gate-Source voltages. Due to this, an effective mobility is introduced which takes exactly these effects into account.

18. References

[1] C. Gatermann, R. Sommer, "Teaching the MOSFET: A Circuit Designer's View," in Proc. 18th International Conference on Synthesis, Modelling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2022, doi:10.1109/SMACD55068.2022.9816264.

[2] Fraunhofer ITWM, Analog Insydes, Accessed: Feb. 07, 2022, [online] available: https://www.itwm.fraunhofer.de/en/departments/sys/products-and-services/analog-insydes.html.

[3] J. E. Lilienfeld, "Method and Apparatus for Controlling Electric Currents," U.S. Patent 1 745 175, filed 1926, granted 1930.

[4] J. E. Lilienfeld, "Amplifier for Electric Currents," U.S. Patent 1 877 140, filed 1928, granted 1932.

[5] J. E. Lilienfeld, "Device for Controlling Electric Currents," U.S. Patent 1 900 018, filed 1928, granted 1933.

[6] O. Heil, "Improvements in or Relating to Electrical Amplifiers and other Control Arrangements and Devices," British Patent 439 457, filed and granted 1935.

[7] W. Shockley, G. L. Pearson, "Modulation of Conductance of Thin Films of Semiconductors by Surface Charges," Physical Review, **74**, 232, 1948, doi: 10.1103/PhysRev.74.232.

[8] J. R. Ligenza, W.G. Spitzer, "The Mechanisms for Silicon Oxidation in Steam and Oxygen," Journal of Physics and Chemistry of Solids, **14**, 131-136, 1960, doi:10.1016/0022-3697(60)90219-5.

[9] M. M. Atalla, "Semiconductor Devices Having Dielectric Coatings," U.S. Patent 3 206 670, filed 1960, granted 1965.

[10] D. Khang, M. M. Antalla, "Silicon-Silicon Dioxide Field Induced Surface Devices," in 1960 IRE-AIEE Solid-State Device Research Conference, 1960.

[11] D. Khang, "A Historical Perspective on the Development of MOS Transistors and Related Devices," IEEE Transactions Electron Devices, **23**(7), 655-657, 1976, doi:10.1109/T-ED.1976.18468.

[12] C. T. Sah, "Evolution of the MOS Transistor – From Conception to VLSI", in 1988 Proceedings of the IEEE, 1280-1326, 1988, doi: 10.1109/5.16328.

[13] D. Kahng, Applied Solid State Science Supplement 2: Silicon Integrated Circuits Part A, Academic Press, 1981.

[14] Y. Taur, T. H. Ning, Fundamentals of Modern VLSI Devices, Cambridge University Press, 2009.

[15] R. M. Warner, B. L. Grung, MOSFET Theory and Design, Oxford University Press, 1999.
[16] Y. Tsividis, Operation and Modelling of the MOS Transistor, Oxford University Press, 2012.
[17] G. N. Lewis, "The Atom and the Molecule," Journal of the American Chemical Society, **38**(4), 762-785, 1916, doi:10.1021/ja02261a002.
[18] F. Schwierz, H. Wong, J. J. Liou, Nanometer CMOS, Pan Standford Publishing, 2010.
[19] S. M. Sze, Kwok K. Ng, Physics of Semiconductor Devices, John Wiley & Sons, 2006.
[20] L. Stiny, Grundwissen Elektrotechnik und Elektronik, Springer Vieweg Wiesbaden, 2018.
[21] W. Grabinski, "EKV v2.6 Parameter Extraction Tutorial," in 2001 ICCAP Users' Web Conference, 2001.
[22] P. Antognetti, G. Massobrio, Semiconductor Device Modeling with Spice, McGraw-Hill, 1993.
[23] S. K. Saha, Compact Models for Integrated Circuit Design, CRC Press, 2016.
[24] X. Li, W. Wu, G. Gildenblat, G.D.J. Smit, A.J. Scholten, D.B.M. Klaassen, R. van Langevelde, NXP Semiconductors PSP 102.3, 2008, Technical Note: NXP-R-TN-2008/00162.
[25] R. Sommer, D. Krauße, E. Schäfer, E. Hennig, "Application of Symbolic Circuit Analysis for Failure Detection and Optimization of Industrial Integrated Circuits," in Design of Analog Circuits through Symbolic Analysis, pp. 445-477, Bentham Science Publishers, 2012, doi:10.2174/978160805095611201010445.
[26] E Hennig, R Sommer, "A reliable iterative error tracking method for approximate symbolic pole/zero analysis", in 2001 European Conference on Circuit Theory and Design (ECCTD'01), 2001, lib.tkk.fi.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 2, 44-63 (2023)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Hybrid Intrusion Detection Using the AEN Graph Model

Paulo Gustavo Quinan[*,1], Issa Traoré[1], Isaac Woungang[2], Ujwal Reddy Gondhi[1], Chenyang Nie[1]

[1] *University of Victoria, Department of Electrical and Computer Engineering, Victoria, B.C., Canada*

[2] *Ryerson University, Department of Computer Science, Toronto, ON, Canada*

ARTICLE INFO

Article history:
Received: 28 November, 2022
Accepted: 21 February, 2023
Online: 11 March, 2023

Keywords:
Attack fingerprint
Anomaly detection
Intrusion detection system
Subgraph matching
Unsupervised machine learning
Graph database

ABSTRACT

The Activity and Event Network (AEN) is a new dynamic knowledge graph that models different network entities and the relationships between them. The graph is generated by processing various network security logs, such as network packets, system logs, and intrusion detection alerts, which allows the graph to capture security-relevant activity and events in the network. In this paper, we show how the AEN graph model can be used for threat identification by introducing an unsupervised ensemble detection mechanism composed of two detection schemes, one signature-based and one anomaly-based. The signature-based scheme employs an isomorphic subgraph matching algorithm to search for generic attack patterns, called attack fingerprints, in the AEN graph. As a proof of concept, we describe fingerprints for three main attack categories: scanning, denial of service, and password guessing. The anomaly-based scheme, in turn, works by extracting statistical features from the graph upon which anomaly scores, based on the bits of meta-rarity metric first proposed by Ferragut et al., are calculated. In total, 15 features are proposed. The performance of the proposed model was assessed using two intrusion detection datasets yielding very encouraging results.

1 Introduction

The Activity and Event Network (AEN) is a new graph that models a computer network by capturing various network security events that occur in the network perimeter. The AEN has the purpose of providing a base for the detection of both novel and known attack patterns, including long-term and stealth attack methods, which have been on the rise but have proven difficult to detect.

This paper is an extension of work originally presented in the 3rd Workshop on Secure IoT, Edge and Cloud systems (SIoTEC) of the 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid 2022) [1]. In the present paper, we present an unsupervised ensemble intrusion detection mechanism composed of two detection schemes, one signature-based and one anomaly-based, with the goal of leveraging the strengths of both types of detection methods and mitigating their weaknesses.

Signature-based detection, also known as rule-based detection, works by searching data for specific characteristics of previously seen attacks. This makes it good at detecting known attack patterns, but at the same time renders it ineffective when confronted with new and unseen attacks. In contrast, anomaly detection methods rely on the assumption that events deviating from normal usage patterns or behaviours are potentially malicious. This method has the potential to detect novel attack patterns but may generate a large number of false positives due to the fact that atypical events are not necessarily malicious [2, 3].

To validate the scheme, we provide a collection of attack fingerprints covering a small subset of known scanning, denial of service (DoS) and password guessing attacks.

The fingerprints are described using Property Graph Query Language (PGQL) because it provides a standardized language for describing graphical patterns, which we believe makes comprehension easier than describing the fingerprints algorithmically. We also provide a subgraph matching algorithm specifically for finding subgraphs that are isomorphic to the fingerprints.

The anomaly-based scheme, in turn, involves calculating anomaly scores based on the bits of meta-rarity metric introduced by [4] for a set of 15 statistical features and underlying distributions extracted from the AEN graph.

To evaluate the proposed intrusion detection mechanism, we conducted experiments with two datasets: the Information Security and Object Technology (ISOT) Cloud Intrusion Detection (ISOT-CID)

[*]Corresponding Author: Paulo Gustavo Quinan, quinan@uvic.ca

Phase 1 dataset [5] and the 2017 Canadian Institute for Cybersecurity (CIC) Intrusion Detection Evaluation Dataset (CIC-IDS2017) [6]. First, each of the two schemes were separately evaluated, and then an ensemble classification was created that fuses the two results. The obtained results were promising for both the individual detection schemes and for the combined method.

The remainder of this paper is structured as follows. In section 2, we review the literature on graph-based intrusion detection, anomaly detection and subgraph matching. In section 3 we give a brief overview of the AEN graph. In section 4, we present the fingerprint model and explain how the fingerprints are described and searched for in the graph. In section 5, we provide detail about the anomaly detection model, including a description of the anomaly score calculation and the proposed features model. In section 6, we present the experimental evaluation of the proposed scheme and discuss the obtained performance results. Finally, in section 7, we make concluding remarks.

2 Literature Review

2.1 Graph-based Intrusion Detection

Many different graphical models have been proposed for intrusion detection or forensic analysis. One traditional focus is on non-probabilistic models, such as attack graphs. These include state attack graphs [7]–[9], logical attack graphs [10] and multiple prerequisite graphs [11], each of which aims to either elucidate different aspects of the system or network's security issues or, at a minimum, fix the limitations of the previous models. In general, there are many open challenges when working with attack graphs [12].

Current approaches are plagued by the exponential growth of the graph according to its vulnerabilities and network size, making generation intractable, even for a few dozen hosts. They are limited in scope, and while they provide static information about the attack paths and the probability of a vulnerability exploitation, they do not provide any information about other effective parameters, such as current intrusion alerts, active responses or network dependencies. Furthermore, current attack graphs do not capture the dynamic and evolving nature of the long-term threat landscape. In practice, each change to the network requires a complete recreation of the graph and a restart of the analysis, which means they can only be applied for offline detection.

Another important area of research is in probability or belief-based models used for signature-based intrusion detection such as those employing Bayesian networks (BNs) [13]–[15] and Markov random fields (MRFs) [16, 17]. These models provide good results for the modelled attacks but have some important limitations. The first stems from the fact that the entity being modelled is not the network itself. Instead, the graph is constructed based on predefined features, which limits the extensibility of the models. This is because each new attack type requires the definition of more predefined features that must be incorporated into the graph as new elements. Moreover, these methods require a training phase used to define the graph's probabilities, making them fully supervised methods. Finally, like attack graphs, they need to be reconstructed whenever the graph changes, which can be time consuming, and in practice make them unable to perform online detection.

Moving to anomaly detection applied for intrusion detection, numerous models have been proposed that have used a diverse range of techniques, such as decision trees [18] and neural networks [19]–[21]. These models obtain good performance but suffer from the previously mentioned issues, including necessitating a training phase, requiring multiple rounds of training in some cases, and not supporting online detection. Moreover, despite being anomaly-based, some models have a limited capability to identify novel attacks due to their structure based on predefined features.

In our work we overcame these problems by modelling the network itself. This allows for greater extensibility in describing new attacks because, rather than attack features being predefined, they may be extracted from the graph. Moreover, the AEN graph is fully dynamic and in constant change. Each new subgraph matching operation can be performed against the graph online without the need to recreate it after every change. Finally, the proposed schemes are all unsupervised, which eliminates the need for a training phase.

2.2 Isomorphic Subgraph Matching

Isomorphic subgraph matching is used to search graphs for subgraphs that match a particular pattern. It has been employed extensively in diverse areas, including computer vision, biology, electronics and social networks. However, to the best of our knowledge, our work is the first to employ isomorphic subgraph matching for signature-based intrusion detection.

The general form of this problem is known to be NP-complete [22]; however, its complexity has been proven to be polynomial for specific types of graphs, such as planar graphs [23].

Different algorithms have been proposed for this problem. For example, Ullmann's algorithm [24] uses a depth-first search algorithm to enumerate all mappings of the pattern. Over the years, many improvements have been proposed for that algorithm, such as the VF2 algorithm [25], in attempts to more effectively prune search paths. More modern algorithms, such as the Turboiso [26] and the DAF [27], employ pre-built auxiliary indexes to accelerate searches and facilitate search-space pruning. These algorithms can perform several orders of magnitude faster than index-less algorithms like Ullmann's but require more memory to store the index and also some pre-processing time to build the indexes.

In our work, we leveraged these ideas to design a custom-made matching algorithm specifically to match fingerprints, given the specific characteristics of the AEN and the proposed fingerprints.

3 AEN Graph Overview

The AEN graph was designed to model the variety, complexity and dynamicity of network activity, along with the uncertainty of its data, something that is intrinsic to the collection process, through a time-varying uncertain multigraph. The graph is composed of different types of nodes and edges, with the nodes describing different types of network elements, such as hosts, domains and accounts, and the edges describing their relationships, such as sessions (sets of traffic between two hosts of the same protocol, ports, etc.) and authentication attempts (an account trying to authentication on a host). Furthermore, each element type has different sets of proper-

ties, including domain name, account identifier, session protocol and start and stop time. To build such a graph, data from heterogeneous sources (*e.g.* network traffic, flow data) and system and application logs (*e.g.* syslog, auditd) are used.

The graph is built online, with elements added or modified as soon as they are observed in the data and old elements removed once they are considered stale. Consequently, the graph serves as a stateful model of the network, and as such, can be used as a basis for many different types of analyses and inferences. Interested readers are referred to [28] for more details on the AEN graph model's elements and construction.

4 Attack Fingerprints in the AEN Graph Model

4.1 AEN Fingerprints Framework

4.1.1 Attack Fingerprint Visualization

As a visual example of how attack fingerprints can be mined from the AEN graph, Figure 1 shows how certain network activity can create evident patterns in the graph. Specifically, the figure shows a visualization of a subset of a graph generated from an example dataset containing a distributed password guessing attack. The hosts are represented by blue nodes at the center, accounts that were used in authentication attempts are represented by orange nodes and the attempts themselves (edges) are either blue when successful or orange when unsuccessful.

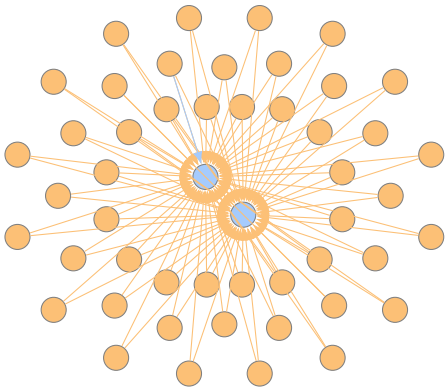


Figure 1: Visualization of an AEN graph containing a password guessing attack.

The figure shows a "cloud" of failed authentication attempts against the two central hosts using the same set of accounts. Furthermore, the majority of the edges are orange, indicating failed attempts, but there is a single blue (successful) edge. This pattern makes evident a successful combined credential stuffing and spraying attack where, after several failed attempts, one login was successful.

Likewise, other types of attacks insert their own distinct attack patterns in the graph. It follows that those patterns serve as fingerprints of these attacks and can, therefore, be mined using subgraph isomorphism matching algorithms to identify instances of an attack.

The statefulness of the AEN plays an important role here because it permits the formation of long term patterns. That is in contrast with traditional intrusion detection systems (IDSs) which can only identify short-term patterns. In the given example, the attack could have been carried out over several weeks, which would have created a challenge for traditional detection mechanisms. In contrast, because the AEN maintains the relationships over a longer term, those patterns can emerge and be identified.

4.1.2 Problem Definition

Given a graph $G = (N_G, E_G)$ where N_G is the set of nodes and $E_G : N_G \times N_G$ is the set of edges, the graph $F = (N_F, E_F)$ is isomorphic to a subgraph $G\prime$ of G if all nodes and edges of F can be mapped to nodes and edges of G. More formally, $F \cong G\prime \sqsubseteq G$ if there is a bijective function $f : N_F \mapsto N_G$ such that $\forall u \in N_F,\ f(u) \in N_G$ and $\forall (u_i, u_j) \in E_F,\ (f(u_i), f(u_j)) \in E_G$.

The definition above can easily be extended to apply to more complex graphs that contain labels and properties, such as the AEN, by applying f to those labels and properties as well.

Finally, the problem of matching a fingerprint is defined as the following: Given an AEN graph G and a fingerprint F, find all distinct subgraphs of G that are isomorphic to F.

4.1.3 Describing Attack Fingerprints

In this study, the attack fingerprints are described using the PGQL query language [29] because it provides a standardized language for describing queries, or patterns, that we wish to search for in a property graph. We believe this makes comprehension easier than when the fingerprints are described algorithmically. This is because PGQL's syntax follows SQL where possible, except that instead of querying tables, it aims to find matches in the nodes and edges of a graph. Doing so requires specific symbols and constructs for that purpose, but is still easily understandable by those who already know SQL.

Other graph query languages, such as Cypher [30], are also descriptive for this purpose but are less like SQL and have a distinct set of supported features. Still, in most cases, PGQL queries can easily be adapted to other graph query languages.

A simple example of PGQL query is as follows:

Algorithm 1: PGQL query example

```
SELECT s, d
  MATCH (s:HOST)-[e:SESSION]->(d:HOST)
  WHERE e.duration > 30
```

The `SELECT` clause specifies what values are to be returned, while the `MATCH` clause specifies the pattern to match. The parentheses are used to describe nodes, while the square brackets describe edges, with the arrow specifying the direction, if any. Inside the brackets, the colon separates the variable name to the left and the optional label, or type, to the right. The above example matches any pattern in the graph that involves two nodes, `s` and `d`, of type `HOST` connected by a directed edge, `e`, of type `SESSION`, from `s` to `d`, whose duration is greater than 30, and then returns the two nodes for each match.

In general, PGQL allows for a rich description of graph patterns; however, it has limitations which make it impossible to fully express certain attack patterns and, in particular, the information we wish to

retrieve from it. For our specific use case, PGQL has the following key limitations:

1. Subquery is not supported in the FROM clause.

2. There is limited array aggregation support: In some cases, it is desirable to group matches by destination (the victim) and get an array of sources. However, the current PGQL specification supports only array aggregation of primitive types in paths (using the `ARRAY_AGG` function). Therefore, only properties like IDs can be aggregated in this way. Note that there are some cases where only the `LISTAGG` function is supported. In those cases, we use the `ARRAY_AGG` function to substitute for `LISTAGG` as if the former had similar support as the latter for simpler pattern description.

To overcome these limitations, the fingerprints contain a post-processing phase during which the query results are processed to reach a final match result.

4.1.4 Attack Fingerprint Matching

Finding matches to fingerprints in the AEN graph requires the application of an isomorphic matching algorithm. How this is accomplished depends on the graph engine used to store the AEN graph.

In engines that natively support PGQL, such as PGX and Oracle's RDBMS with the OPG extension, the fingerprints can be used directly to query the database, with only the post-processing phase requiring further implementation. In this case, the matching algorithm is implemented by the graph engine itself.

Similarly, in engines that support other graph query languages, such as neo4j, the fingerprints need first to be converted to the supported query language, but after that, only the post-processing phase requires implementation.

In contrast, when using any graph engine that does not support a graph query language, the whole fingerprint matching algorithm must be implemented. There are several general isomorphic matching algorithms, including Ullmann's algorithm [24] and its derivations (*e.g.* VF2 [25]), the Turboiso [26] algorithm and the DAF [27] algorithm. However, the specific characteristics of the AEN graph and the proposed fingerprints means that a simpler searching algorithm can be employed.

Specifically, the small diameter of the fingerprints means that recursion or any type of partial matching is unnecessary, while the types and properties of the nodes and edges allow for large swathes of the search space to be quickly pruned. In practice, the custom graph engine implemented for the AEN speeds up searching at the cost of extra memory by maintaining separate sets of nodes per type, as well as separate sets of edges per type and per source and destination pair. This can be considered analogous to the indexes employed by the general algorithms mentioned previously, such as the Turboiso and the DAF.

Consequently, finding pairs of nodes per type and edges between them is a constant-time operation. Conversely, iterating the set of edges or groups is done linearly because no index per property is maintained. However, this operation can be trivially parallelized. Aggregating values, such as summing up properties of elements in groups, must also be done linearly.

A generic fingerprint matching algorithm equivalent is presented in Algorithm 2. The algorithm starts by pairing nodes of the desired types (note that each pair is directed). Then, for each pair, it finds all edges between the source and the destination nodes. For each of those edges, the `matches` function is used to test whether the edge matches all of the `WHERE` clauses of the fingerprint and then accumulates the matching edges.

Afterwards, edges are grouped into sets according to the `GROUP BY` expression specified in the fingerprint. Subsequently, each set (group) is tested for matches to all of the `HAVING` clauses of the fingerprint. If true, the results are extracted from the elements in the set based on the `SELECT` expression. These results map to what is returned by the PGQL queries.

Algorithm 2: Generic fingerprint matching algorithm

```
pairs ← pairNodes(nodes, srcType, destType)

matched ← {}

foreach pair in pairs do
  s, d ← pair
  edges ← getEdges(s, d, edgeType)

  foreach e in edges do
    if matches(e, whereClauses) then
      matched ← matched ∪ {e}

groups ← group(matched, groupingExpr)

preResults ← {}

foreach g in groups do
  if matches(g, havingClauses) then
    gr ← extractPreResult(g, selectExpr)
    preResults ← preResults ∪ {gr}

return preResults
```

As mentioned previously, many fingerprints also require a post-processing phase, for which Algorithm 3 is the generic algorithm used. It returns a set of matches of the fingerprint as described in the respective section of each fingerprint.

Algorithm 3: Generic fingerprint post-processing phase algorithm

```
postProcGroups ← group(preResults,
    postProcGroupingExpr)

results ← {}

foreach ppg in postProcGroups do
  if matches(ppg, postProcClauses) then
    r ← extractFinalResult(ppg,
        postProcSelectExpr)
    results ← results ∪ {r}

return results
```

Finally, as explained in the following sections, some fingerprints

employ the sliding window algorithm to define time windows. In those cases, Algorithms 2 and 3 are applied for each time window separately, although it would be straightforward to apply the post-processing phase to the combined results of all time windows and group them accordingly. Because the consecutive time windows share some elements, it is possible that the same results will appear on different windows. Therefore, an optional final step when the sliding window is used is the deduplication of results in different time windows, which can be applied to the results returned by Algorithm 3.

Also, to speed up searching in those cases, the sets of edges between nodes are sorted by the desired sliding property so that the start and end indexes of each window can be quickly identified through the application of a binary search. Moreover, the algorithm maintains a cursor to the initial position of the previous window so that older elements do not need to be searched.

4.2 Scanning Attacks

Scanning attacks consist of probing machines for openings that can be further explored for vulnerabilities and then exploited. They are part of the initial information gathering phase of an attack.

These attacks can target a variety of protocols and applications but are most commonly employed for scanning TCP and UDP ports [31]. They can be deployed by a single source or be distributed among several attackers. In addition, there are many different techniques used for scanning, with each focusing on different layers and using different methods to avoid detection [32, 31].

Scanning attacks can be classified based on several different properties. With regard to their footprints, they can be classified into three major types [32]:

- Vertical scan, which scans multiple ports on a single host
- Horizontal scan, which scans the same port across multiple hosts
- Block scan, which is a combination of both vertical and horizontal scans, whereby multiple ports are scanned across multiple target hosts

With regard to their timing, they can be classified as a slow scan or a fast scan, with the latter being easier to spot than the former, given its speed and short duration [33].

In the following, we propose a fingerprint for single-source fast vertical scans. Fingerprints for other types of scans can be derived by slightly modifying the fingerprint parameters as demonstrated in subsection 4.3 for the different DoS attacks.

As already mentioned, vertical scans target a specific host by sweeping across the port space, looking for open ports and running services. Unique characteristics can be summarized as follows [31]:

- The packets are sent from one source host to one destination host.
- The packets have several different destination ports.
- The amount of data/bytes exchanged is never large. For TCP scans, for instance, connections are almost never even established.
- The time frame of each single session is very short.

Taking into consideration these characteristics, we can define a typical attack as one with a short duration and a small amount of data exchanged, particularly from the victim. Otherwise the attack would be too heavy and easier to spot, but with a large number of ports involved. This definition can be described by the following fingerprint:

Algorithm 4: Fingerprint for scanning attack

```
SELECT s, d
  MATCH (s:HOST)-[e:SESSION]->(d:HOST)
  WHERE e.destSize < sizeThr
    AND e.duration < durThr
  GROUP BY s,d
  HAVING count(DISTINCT e.destPort) > portThr
```

where:

- `destSize` corresponds to the cumulative size of the packets of the session sent by the destination of the session, which in this case is the target host;
- `sizeThr` defines a threshold for a maximum expected `destSize`;
- `duration` corresponds to the total time duration of a session;
- `durThr` defines a threshold for a maximum expected `duration`;
- `destPort` corresponds to the target ports; and
- `portThr` defines a threshold for the minimum number of distinct destination ports.

When applied, the fingerprint returns all pairs of source and destination hosts where the sources and destinations correspond to the attackers and the victims, respectively, according to the aforementioned characteristics.

4.3 Denial of Service

DoS is a family of attacks that aim to disrupt the service of a target server or network resource and make it completely or partially unavailable to users. They are broadly divided into two categories [34, 35]:

- Volumetric attacks, where the target is inundated with huge amounts of traffic that overwhelm its capabilities. These include most flood attacks and amplification attacks.
- Semantic attacks, also known as resource depletion attacks, where weaknesses in applications or protocols are exploited in order to render a resource inoperable without requiring the same large volume of traffic as pure volumetric attacks. These include attacks like TCP SYN flood and slow-rate attacks like Slowloris [36].

Based on the source of attack, DoS attacks can be single-source or distributed, in which case they are commonly referred to as distributed DoS (DDoS). In this section, we use DoS to refer to both types.

Another common characteristic of many DoS attacks is that the source IP address can be spoofed in order to hide the true source of the attack and to deflect replies away from the attacker. This introduces asymmetry into the traffic load between the attacker and the victim [37, 38]. In other words, the IP addresses identified by the fingerprints as sources of attacks might be spoofed IP addresses in many cases.

In this section, we focus on selected flood attacks covering both categories of DoS attacks under different layers of the OSI model, specifically layers 3 (network layer), 4 (transport layer) and 7 (application layer).

The fingerprints follow a basic pattern of counting the number of matching sessions of a specific attack type within a short time frame. For this reason, we employed a sliding window mechanism with large overlaps between each window and applied the fingerprints separately for each window. Sliding windows were used instead of simply slicing the timeline so that any short duration attack that would otherwise be divided between two windows could be fully inside at least one window. This had no effect on long duration attacks, as they would fully cover at least one window regardless. In the fingerprints, the start and end times of a time window are represented by `twStart` and `twEnd` parameters, respectively.

4.3.1 *ICMP ping flood*

ICMP ping flood is an attack where a high volume of ICMP echo/ping requests are sent to a target IP address in the expectation of flooding the victim with more traffic than it is capable of handling [38].

Based on this, we identified the primary typical characteristics of an ICMP ping flood as the following:

- The attacker host sends a large number of ping requests (i.e. ICMP packets) to the target host.
- The packets correspond to echo requests and replies and thus are small.
- The time frame for any single session is very short.

These characteristics can be expressed by the following fingerprint:

Algorithm 5: Fingerprint for ICMP Flood DoS attack

```
SELECT s, d, count(e)
  MATCH (s:HOST)-[e:SESSION]->(d:HOST)
  WHERE e.protocol = 'icmp'
    AND e.destSize < sizeThr
    AND e.startTime > TIMESTAMP 'twStart'
    AND e.stopTime < TIMESTAMP 'twEnd'
  GROUP BY s,d
  HAVING count(e) > sessionThr
```

where `sessionThr` defines a threshold for the minimum number of distinct sessions to trigger the fingerprint.

The query returns the number of sessions between each pair of hosts matching the defined conditions, where the source is the attacker, or the spoofed host, and the destination is the victim. These results might be enough for single-source attacks, but to obtain a final result for distributed attacks, they need to be further processed.

This post-processing step is completed by aggregating the results for each destination host in each time window and applying a further threshold, `cntThr`, on the aggregated count (sum) of matching session per destination. The final result is then a set of attack instances, each one containing the victim host, the cumulative sum of matching sessions and a set of attacker hosts.

Since each time window is considered separately, longer attack instances can end up being reported repeatedly in multiple adjacent windows. To improve on that, the results can be deduplicated by aggregating the results of a same target that fall in contiguous windows.

4.3.2 *IP Fragmentation Attack*

IP packet fragmentation is a normal event whereby packets larger than the maximum transmission unit (MTU) of the route (normally 1500 bytes) are fragmented into smaller packets that are reassembled by the receiver. A problem arises when systems have trouble reassembling the packets or will expend too many resources doing so. Attackers take advantage of the situation by crafting special fragmented packets that are impossible to reassemble, causing targets to either crash due to related bugs or to expend more and more resources trying to handle the reassembly of these degenerate packets [35, 39].

Different protocols can be used for fragmented attacks, including UDP, ICMP and TCP. Moreover, fragmented packets can be used to deceive IDSs by crafting fragmented packets that are rejected by the IDS but not by the end system, or vice versa, such that the extra or missing packets prevent the IDS from identifying an attack it otherwise would [39].

From that, we identified the general characteristics of an IP fragmentation attack as follows:

- A medium to high absolute number of fragmented packets can be observed.
- The ratio of fragmented packets to all packets is high.
- The time frame of a single session is very short.

To be able to capture the ratio of fragmented packets, we introduced two properties to the session edge: one that tracks the number of packets, `pktCnt`, comprising the session and another that tracks the number of fragmented packets among those, `fragPktCnt`.

The fingerprint can be expressed as the following query:

Algorithm 6: Fingerprint for IP Fragmentation attack

```
SELECT s, d, count(e), sum(e.fragPktCnt)
  MATCH (s:HOST)-[e:SESSION]->(d:HOST)
  WHERE e.fragPktCnt / e.pktCnt > fragRatioThr
    AND e.startTime > TIMESTAMP 'twStart'
    AND e.stopTime < TIMESTAMP 'twEnd'
  GROUP BY s, d
  HAVING count(e) > sessionThr
```

where `fragRatioThr` defines a threshold for the minimum ratio of fragmented packets to all packets of each session that is considered to be matching the fingerprint.

As before, this query returns the number of sessions matching the defined conditions between each pair of hosts, where the source is the attacker and the destination is the victim. In addition, it also returns the sum of the fragmented packet counts from all grouped sessions.

A post-processing phase is included where the results are aggregated by destination host in each time window so that distributed attacks can be identified. A further threshold, `fragPckCntThr`, was applied to the aggregated sum of fragmented packet counts to guarantee that normal absolute amounts of fragmented packets exchanged between hosts are filtered out.

The final result is then a set of attack instances, each containing the victim host, the cumulative sum of matching sessions and fragmented packet counts, and a set of attacker hosts.

Finally, a deduplication step can also be executed to combine instances from adjacent time windows.

4.3.3 TCP SYN Flood

TCP SYN flood attacks exploit the three-way TCP handshake process by sending a large volume SYN requests to a target host without ever completing the handshake process with the expected ACK requests. This causes the target server to hold multiple partially initiated connections, eventually filling its connection buffer and thus preventing subsequent real connections from being established. In some cases, this will result in crashes due to unhandled resource starvation [37].

Therefore, for this attack, we needed to keep track of the state of the TCP connection. After the initial SYN packet is sent and a session is created, we defined four possible states for the connection, with the first three mirroring the TCP states related to connection establishment [40], only with a slight change of semantics because the client and server states are combined:

- `SYN_SENT`: The initial session state when the session is created from a SYN packet sent by the source host. This means the SYN packet was sent and the source host is now waiting for the SYN-ACK packet.
- `SYN_RECEIVED`: With the session at the `SYN_SENT` state, the destination host has sent the SYN-ACK packet meaning it received the original SYN packet and is now waiting for the ACK packet that will conclude the handshake.
- `ESTABLISHED`: The source host has sent the ACK packet while the session was at the `SYN_RECEIVED` state, which concluded the three-way handshake, establishing the connection. Once established, only FIN and RST packets can change the state of the session.
- `OTHER`: A catch-all state that indicates any other scenario, such as when the first packet of the session is not a SYN packet.

Moreover, two other related properties, `synFlagCount` and `ackFlagCount`, were added to the session edge to track the number of packets added to the session that had the SYN flag and the ACK flag set, respectively.

A limitation of this technique is that it requires the packet information in the graph to be correct, which is not guaranteed in all cases. Examples include cases where the network data injected into the model is not complete, whether due to sampling or an unexpected data loss, and also cases where the system has just gone live and thus only started receiving the network data after the connections were established. Another is the case where the system is fed with NetFlow data instead of raw network data and the NetFlow application did not properly track the TCP state or the number of packets containing each flag in any given flow. In these cases, the model is not able to properly track the correct state of the connections. This makes fingerprints that rely on that information ineffective in identifying attacks.

To mitigate those issues, some correction heuristics were employed to change a session's attributes, such as the TCP state, in cases where inconsistencies between the data and the attributes are encountered. An example is when it is observed that large amounts of data are being exchanged between two hosts on a TCP session, indicating a fully established connection, but the state of the connection indicates otherwise.

In short, the characteristics of a TCP SYN flood attack can be summarized as follows:

- The attacker keeps sending SYN packets to the victim and never replies to the SYN-ACK packet, resulting in a large number of sessions in the `SYN_RECEIVED` state.
- The time frame of any single session is very short.

The above pattern can be expressed in the following query:

Algorithm 7: Fingerprint for TCP SYN flood attack

```
SELECT s, d, count(e)
  MATCH (s:HOST)-[e:SESSION]->(d:HOST)
  WHERE e.tcpState = SYN\_RECEIVED
    AND e.synFlagCount / e.ackFlagCount >
        synAckRatio
    AND e.startTime > TIMESTAMP 'twStart'
    AND e.stopTime < TIMESTAMP 'twEnd'
  GROUP BY s, d
  HAVING count(e) > sessionThr
```

where:

- `destSize` corresponds to the cumulative size of the packets in the session sent by the destination of the session, in this case, the target host.
- `sizeThr` defines a threshold for the maximum expected `destSize`.
- `duration` corresponds to the total length of a session.
- `durThr` defines a threshold for the maximum expected `duration`.
- `destPort` corresponds to the target ports.
- `portThr` defines a threshold for the minimum number of distinct destination ports.

This query mostly follows the same pattern as the preceding ones, with the distinction that it has a condition to only select a session if its TCP state is `SYN_RECEIVED`.

The post-processing phase follows the same pattern as the preceding ones as well, so distributed attacks can be identified and the `cntThr` applied.

4.3.4 Other TCP "Out-of-State" Flood Attacks

Aside from the aforementioned SYN flood attack, there are many other less common TCP-based layer 4 flood attacks variants that exploit illegal or unexpected combinations of TCP packet flags sent without first establishing a TCP connection (thus the "out-of-state" term) with the objective of causing a DoS [41]. The lack of a prior connection causes some systems to return RST packets, which can exacerbate bandwidth consumption problems related to the attack. Finally, bugs stemming from unexpected conditions can also cause issues. Examples of flag combinations used in these attacks include:

- ACK-PSH
- PSH-RST-SYN-FIN
- ACK-RST
- URG-ACK-PSH-FIN
- URG

Because these attacks involve out-of-state packets that form the initial packets of the sessions, it is possible to refine the session's TCP state property to track these cases. Specifically, a new property called `tcpFirstPktFlags` was added to the session edge to track the flags of the first packet of the session if it is a TCP packet and the TCP state is set to `OTHER`.

With that, it is possible to define a generic query following the same pattern as the SYN flood attack query, but parameterized by the first packet flags corresponding to the sought after attacks:

Algorithm 8: Fingerprint for TCP "Out-of-State" flood attacks

```
SELECT s, d, count(e)
  MATCH (s:HOST)-[e:SESSION]->(d:HOST)
  WHERE e.protocol = 'TCP'
    AND e.tcpState = OTHER
    AND e.tcpFirstPktFlags = attackFlags
    AND e.startTime > TIMESTAMP 'twStart'
    AND e.stopTime < TIMESTAMP 'twEnd'
  GROUP BY s, d
  HAVING count(e) > sessionThr
```

The query requires the same post-processing as the SYN flood attack. It is also possible to modify the fingerprint so that it matches any of the possible out-of-state attack flag combinations instead of matching each one individually by allowing `tcpFirstPktFlags` to be equal to any of the known invalid flag combinations.

4.3.5 UDP Flood

UDP flood attacks are flood attack aimed at UDP datagrams. It is considered a volumetric attack because it does not exploit any specific characteristic of the UDP protocol. Instead, it works by sending a large volumes of UDP packets to random or fixed ports on a target host, depleting its available bandwidth, which makes it unreachable by other clients. The attack can also consume a lot of the target's processing power as it tries to determine how to handle the UDP packets [42].

In summary, the key characteristics of a UDP flood attack are as follows:

- The attacker sends UDP packets to the victim at a high rate of frequency.
- The amount of data exchanged per session is relatively fixed and mostly the same.
- The time frame for any single session is very short.

These characteristics can be expressed by the following fingerprint:

Algorithm 9: Fingerprint for UDP flood attack

```
SELECT s, d, count(e)
  MATCH (s:HOST)-[e:SESSION]->(d:HOST)
  WHERE e.protocol = 'UDP'
    AND e.destSize < sizeThr
    AND e.startTime > TIMESTAMP 'twStart'
    AND e.stopTime < TIMESTAMP 'twEnd'
  GROUP BY s,d
  HAVING count(e) > sessionThr
```

Once again, this query mostly follows the same pattern as the preceding ones, with the distinction being the condition to only select UDP sessions.

The post-processing phase follows the same pattern as the preceding ones as well, so distributed attacks can be identified and the `cntThr` applied.

4.3.6 HTTP Flood

HTTP flood is a layer 7 DoS attack in which a target server is saturated with a high volume of HTTP requests. This can slow the server as it tries to handle the high volume and eventually makes the servers unable to handle legitimate traffic [43].

Because a TCP connection must be established for these attacks to be performed, the spoofing of IP addresses is not possible [35], which makes the identification of source IP addresses more reliable.

An HTTP flood attack can use different types of requests and methods (*e.g.* GET, POST), with the most damaging ones being the heaviest requests for a server to handle, such as those involving heavy processing of input or pushing large amounts of data into a database [43, 35]. Consequently, less bandwidth is required to bring down a web server using an HTTP flood attack than is required for another type of DoS attack.

Several different techniques are employed in HTTP flood attacks. Some send a large number of requests, while others send fewer, but very large or very focused, requests. In either case, the attack involves sending large amounts of IP packets to the target. Therefore, the key characteristics of an HTTP flood attack as follows:

- The attacker sends HTTP packets to the victim at a high rate of frequency.
- The amount of data exchanged per session is high.

Naturally, the fingerprint for these attacks needs to be able to identify HTTP sessions. For this reason, a `service` property was added to the session edge so that sessions can be marked as HTTP-related. Note that this property can be used for other reasons as well, such as identifying SSH or FTP sessions.

The challenge in this case is populating the field, given that HTTP is a layer 7 protocol and can, in many cases, be encrypted. We employed two techniques for this purpose. The first was performing deep packet inspection (DPI) to search for identifiers of HTTP messages, such as the version, in packet content. Once a session is identified as being "HTTP-related", it is marked as such and no DPI is required thereafter. A limitation of this technique is that it requires clear text traffic, which in most scenarios today would require the AEN to be deployed after a TLS termination proxy. DPI is also computationally expensive.

For those reasons, a second technique was employed using a service registry comprised of IP addresses and ports of services of interest, such as web services and SSH. This allows for a quick discovery of services but also for some false positives if invalid packets are sent to those servers, such as when non-HTTP packets are sent to an HTTP service.

With the capacity to identify HTTP sessions, a query can be defined as follows:

Algorithm 10: Fingerprint for HTTP flood attack

```
SELECT s, d, count(e), sum(e.pktCount)
  MATCH (s:HOST)-[e:SESSION]->(d:HOST)
  WHERE e.protocol = 'TCP'
    AND e.service = 'HTTP'
    AND e.srcSize < sizeThr
    AND e.startTime > TIMESTAMP 'twStart'
    AND e.startTime < TIMESTAMP 'twEnd'
  GROUP BY s, d
  HAVING sum(e.pktCount) > pktCntThr
```

This query is somewhat distinct from the preceding ones because it is based on the number of packets exchanged in a given time window rather than the number of sessions and also because it does not consider short-term sessions. Both differences are a consequence of the fact that HTTP flood attacks require fully established connections. The seemingly redundant clause to select only TCP sessions when there is already a clause to select only HTTP sessions is included to filter out part of the invalid packets, such as UDP packets sent to that specific service, in case the service registry was used.

Finally, the post-processing phase follows the same pattern as the preceding queries as well, with a further aggregation per destination host so that distributed attacks can be identified and the `cntThr` applied.

4.4 Password Guessing

Password guessing is when the attacker tries to gain access to a system by persistently attempting to guess user passwords [44, 45]. The passwords attempted are normally derived from either leaked password associated with a particular user or dictionaries of common passwords, in which case the attack is also known as a brute-force attack.

There are a few different types of password guessing attacks, but they all share the main characteristic of generating a high volume of failed login attempts, which are normally logged by the applications into which the authentications are attempted. For this reason, the AEN ingests application and system logs, like those from SSH, to extract authentication information and insert that it into the graph through nodes of type `ACCOUNT` and edges of type `AUTH_ATT` ("authentication attempt") that link an account with the target host of the authentication. To track whether the authentication attempt was successful, the edge has a Boolean property called `succ`.

Another important characteristic of password guessing attacks is that they do not necessarily happen in a short time frame. Sometimes the whole process can last for days, or even longer. That means there is no need to consider the time frame of the attempts. Incidentally, that means the AEN must keep authentication-related elements for longer than it would for many other types of elements.

In this study, we investigated three types of password guessing attacks:

- Basic: One account on one host is targeted with a brute-force attack.
- Spraying: Multiple accounts on one host are each attacked one or a few times.
- Stuffing: The same account is targeted on multiple hosts one or a few times per host.

4.4.1 *Basic password guessing*

A basic password guessing attack is one where a single account on a single host is targeted with a brute-force attack. It is the most common type of password guessing attack [44, 45]. Because it only focuses on one-to-one relations, the graph patterns should be `(:ACCOUNT)-[:AUTH_ATT]->(:HOST)`. For the sake of brevity the whole fingerprint will not be described because it is a generalization of the *spraying password guessing* fingerprint that follows.

4.4.2 *Spraying password guessing*

In a spraying password guessing attack, instead of multiple passwords being tried with a single account, the attacker tries to breach multiple accounts with a single password or a few passwords [46]. In this manner, the attacker can circumvent the most common authentication protection measures, such as account lockouts.

These attacks can be performed either from a single source, in which case tracking attempts per IP address might be a useful detection method, or from distributed sources, which makes detection harder. This is one of the strengths of the proposed fingerprint, as it focuses exclusively on authentication attempts per victim host, which makes it, by design, effective regardless of the number of source hosts involved.

In summary, spraying password guessing attacks have the following characteristics:

- One host is targeted with a high number of authentication attempts.
- The attempts are spread over several accounts such that no account has more than a few attempts.

- One or more hosts can participate in the attack.
- The time frame of an attack can be very long.

Based on the above characteristics, the query is defined as follows:

Algorithm 11: Fingerprint for spraying password guessing attack

```
SELECT h, count(DISTINCT a), ARRAY_AGG(e.id) as
    attempts
  MATCH (a:ACCOUNT)-[e:AUTH_ATT]->(h:HOST)
  WHERE count(!e.succ) > attemptThr
  GROUP BY h
  HAVING count(!e.succ) / count(e) >
      authFailRatioThr
    AND count(DISTINCT a) > accountThr
```

where:

- `attemptThr` defines a threshold for the minimum number of failed attempts per account to exclude regular login attempt failures from real users.
- `authFailRatioThr` defines a threshold for the minimum ratio of failed attempts in relation to the total number of authentication attempts.
- `accountThr` defines a threshold for the minimum number of distinct accounts that will trigger the fingerprint.

Each match returned by the query contains the victim host, the associated number of distinct target accounts on the host and, as a special output, the array of identifiers of the matching edges, called `attempts`, which is used in the post-processing step to retrieve the list of targeted accounts by fetching from the graph the source nodes of the edges in the array.

Furthermore, if retrieving the hosts responsible for the attack is desired, another query can be performed to find the source hosts of the authentication attempts in the `attempts` array based on its source properties.

4.4.3 *Credential Stuffing*

The credential stuffing attack, also known as a targeted password guessing attack, consists of trying the same credentials (*e.g.* user name and password combination) on multiple hosts [45]. The credentials used in these attack are traditionally obtained from leaks of previous attacks or are default passwords in systems that have them. The latter type is particularly common in IoT devices [47]. More advanced attacks use slight variations of the passwords for cases where the user has the same base password with small modifications per host. Ultimately, this type of attack targets password reuse by the same user on different websites and hosts or poorly designed systems. This consequently means the attacker will not try more than a few different combinations per account but will try the same combinations against multiple targets.

As with any password guessing attack, credential stuffing can be performed from either a single source or distributed sources. However, because this attack targets multiple hosts that normally do not coordinate their detection and prevention efforts, the attacker is more likely to be able to carry out the attack using a single source than with other types of password guessing attacks.

In practice, a single attack campaign can perform credential stuffing on several accounts at once. This type of attack would be detected by the fingerprint as multiple credential stuffing attacks happening together, or possibly as multiple spraying attacks, depending on the number of accounts targeted. Also note that the AEN does not have access to the password, so it cannot determine if the same passwords are being used in multiple hosts, only that multiple failed authentication attempts are being performed for a given account on multiple hosts.

In summary, credential stuffing attacks have the following characteristics:

- Multiple hosts are targeted with a high number of authentication attempts across them.
- Only one account is targeted.
- Only a few attempts are made per host.
- One or more hosts can participate in the attack.
- The time frame of an attack can be very long.

Based on the above characteristics, the query is defined as follows:

Algorithm 12: Fingerprint for credential stuffing attack

```
SELECT a, count(DISTINCT h), ARRAY_AGG(e.id) as
    attempts
  MATCH (a:ACCOUNT)-[e:AUTH_ATT]->(h:HOST)
  WHERE count(!e.succ) > attemptThr
  GROUP BY a
  HAVING count(!e.succ) / count(e) >
      authFailRatioThr
    AND count(DISTINCT h) > hostThr
```

where `hostThr` defines a threshold for the minimum number of distinct hosts that will trigger the fingerprint.

In contrast to the spraying query, this query groups by account rather than host and counts the number of matching hosts instead of the number of matching accounts. As a result, each match returned by the query contains the targeted account, the associated number of distinct hosts where authentications were attempted and the `attempts` array, which in this case is used to retrieve the list of victim hosts by fetching from the graph the destination nodes of the edges in the array in the post-processing phase.

As with the previous query, retrieving the hosts responsible for the attack is possible by performing another query to find the source hosts of the authentication attempts in the `attempts` array, based on its source properties.

5 Anomaly Detection based on the AEN Graph Model

5.1 *Measure of Anomalousness*

Anomaly detection approaches use statistical methods to help identify outliers or rare events, which are flagged as anomalous.

Anomaly is defined as something that deviates from what is standard, normal or expected [3]. When applied to intrusion detection, this involves assuming that anomalous events are more likely to be malicious.

In [48], the authors defined the anomaly score of an event as the negative log likelihood of that event, which was later adapted by Ferragut et al. [4] as the bits of rarity metric. Formally, given a random variable X with probability density or mass function f, the rarity of an event x is defined as:

$$R(x) = -\log_2 P_f(x) \tag{1}$$

The negative means that rarer events have a higher rarity value. Moreover, using the log helps with numerical stability, while the base 2 causes the rarity of the event to be measured in bits. Finally, note that the negative log of zero is defined by convention to be positive infinity.

In [4], the authors also demonstrated that the bits of rarity metric has some important limitations when used for anomaly detection because it is not *regulatable* or *comparable* between two different types of data. As a supporting example, the authors defined two uniform discrete distributions, one with 100 values and one with 2000 values. If a threshold of 10 is chosen to define what is anomalous or not, then no event of the first distribution will be considered anomalous while all events of the second distribution will be even though they are all equally likely.

Note how, in this example, a predefined threshold cannot be used to regulate the number of anomalous events identified in a sample of any distribution. Furthermore, note how it is not possible to compare the rarity of the events of two distinct distributions because the rarity metric of an event is an absolute value that does not describe the rarity of that event relative to its distribution.

For these reasons, the authors proposed a regulatable and comparable anomalousness metric called bits of meta-rarity based on the "probability of the probability" of the event rather than just the probability of an event. More formally, given a random variable X with probability density or mass function f defined on the domain $\mathcal{D}$, the bits of meta-rarity anomaly score $A : \mathcal{D} \to \mathbb{R}_{\geq 0}$ of en event x is defined as:

$$A(x) = -\log_2 P_f(f(X) \leq f(x)) \tag{2}$$

Going back to the previous example, note how for any value x of either distribution, $P_f(f(X) \leq f(x)) = 1$ and consequently $A(x) = -\log_2 1 = 0$. This implies that neither distribution has any anomalous event, regardless of the threshold used (as long as it is greater than 0), and also that the anomaly scores of the different distributions can be compared.

Moreover, for a given threshold θ, the probability that $A(x)$ exceeds that value is bounded by $2^{-\theta}$ such that the ratio of events flagged as anomalous in a sample is never more than $2^{-\theta}$ as long as f fits the sample well. This condition applies for any f which makes the anomaly regulatable through θ.

Note here the importance of a high goodness of fit of f, without that the above condition will not hold.

For a continuous variable, $P_f(f(X) \leq f(x))$ is defined as the area under f restricted to those t such that $f(t) \leq f(x)$, that is

$$P_f(f(X) \leq f(x)) = \int_{\{t|f(t)\leq f(x)\}} f(t)\mathrm{d}t \tag{3}$$

For a discrete variable, $P_f(f(X) \leq f(x))$ is defined as the sum of all probabilities less than or equal to $P_f(x)$, that is

$$P_f(f(X) \leq f(x)) = \sum_{\{t|f(t)\leq f(x)\}} f(t) \tag{4}$$

As a further example, consider the discrete variable $X = \{x_1,\ x_2,\ x_3\}$, such that $f(x_1) < f(x_2) < f(x_3)$. Thus, the anomaly scores of these events are given by

$$\begin{aligned} A(x_1) &= -\log_2(f(x_1)) \\ A(x_2) &= -\log_2(f(x_1) + f(x_2)) \\ A(x_3) &= -\log_2(f(x_1) + f(x_2) + f(x_3)) \end{aligned}$$

In this case, it is clear that $A(x_1) > A(x_2) > A(x_3)$; in other words, as the events become more common, they become less anomalous.

Bringing this to the scope of our work, the variables are the features we extract from the graph as described in subsection 5.2. Each feature is a multinomial variable with k categories, each defined by an n-tuple. For instance, the categories of feature `totalsessions` are defined by the 2-tuple ($SourceHost$, $DestinationHost$), whose values are defined by the total number of sessions between those hosts.

Now recall the importance of a suitable distribution for each variable. This is normally obtained through a training phase. However, because our model is fully unsupervised, it does not contain a training phase. Instead, we estimate the probability of each observed value online based on the frequency of that observation in the sample extracted from the graph, which collectively describe the probability mass function of the feature.

This implies that, although the values of each category of a variable might be continuous in theory, they are discrete in practice, which may cause some issues. Consider, for example, the following sample of a feature: $\{x_1 = 22,\ x_2 = 11,\ x_3 = 22,\ x_4 = 555,\ x_5 = 10,\ x_6 = 9\}$.

Intuitively, x_4 should have the higher anomaly score as its value is farther from the values of the others, but that is not the case. Instead, when considering the frequency of each value, x_1 and x_3 have the same values and thus the same higher probability (i.e. 2/6) and the same anomaly score. The other values, including the value for x_4, each appear only once and thus result in the same lower probability (i.e. 1/6) and the same anomaly score.

To overcome this issue, we discretize the values into bins such that neighbouring values will be mapped to the same bins and thus have a higher probability. In practice, given a bin width h, the binned value of x is defined as

$$b(x) = h\left\lfloor \frac{x}{h} \right\rfloor \tag{5}$$

Note that multiplying by h is only done to keep the binned values near to the original values but is not required in practice.

To help with understanding, Table 1 shows the binned values of a sample using different bin widths. The values with lower probabilities, meaning the most anomalous, for each bin width are bolded.

Table 1: Value binning examples

Category	Bin Width		
	1	5	25
x_1	22	20	0
x_2	**11**	10	0
x_3	22	20	0
x_4	**555**	**555**	**550**
x_5	**10**	10	0
x_6	**9**	**5**	0

It can be seen that as the bin width increases, more similar values are mapped into the same bins, which increases their probabilities. With the bin width of 25, all values except that of x_4 are mapped to the same bin, resulting in a probability of $5/6$, while x_4 continues to have a probability of $1/6$, making it the most anomalous event in the sample for this bin width.

To compute the features, we first split the time range of the data into time windows, with the windows treated independently from each other. For that, we employed the previously discussed sliding window mechanism. Then, for each time window, we extract the features from the AEN graph.

Afterwards, for each feature $X = \{x_1, \ldots, x_k\}$ where k is the size or number of categories of X, its values are binned according to (5). After this process, there will be n bins, with each bin representing a value range. The probability of each bin, $p(b_j)$, where $j = \{1, \ldots, n\}$, is defined as the ratio of the number of elements in the bin over the total number of categories of the feature:

$$p(b_j) = \frac{|b_j|}{k} \tag{6}$$

Clearly, the distribution of p approximates the distribution of f such that $P_f(f(X) \leq f(x_i))$ can be approximated through $p(b(x_i))$. Therefore, it is useful to define the anomaly score of a bin b_j following (2) and (4):

$$A(b_j) = -\log_2 \sum_{\{b_m | p(b_m) \leq p(b_j)\}} p(b_m) \tag{7}$$

From that, the anomaly score of x_i is defined as equal to the anomaly score of its binned value:

$$\begin{aligned} A(x_i) &= A(b(x_i)) \\ &= -\log_2 \sum_{\{b_m | p(b_m) \leq p(b(x_i))\}} p(b_m) \end{aligned} \tag{8}$$

The last step of the anomaly detection is to compare the anomaly scores with a predefined threshold such that if the anomaly score of an element is greater than the threshold, that element is considered anomalous. Specifically, the source element of the anomalous feature tuples are the ones actually considered anomalous. For instance, for feature `totalsessions`, it is the source host, not the destination host, that is reported as anomalous.

5.2 Feature Model

In this subsection, we describe the proposed feature model, which contains a wide range of features extracted from the AEN graph. The features are categorized into session features, which are extracted from session data, and authentication features, which are extracted from authentication data.

Note that all features are contained within a time window, meaning that each operation described below is performed only on the edges that were created in that time window. Also note that the features are directed, so any feature extracted for a pair of hosts h_1 and h_2 is different from that same feature between h_2 and h_1.

5.2.1 Session Features

The main type of edge in the AEN graph model is the session edge, which represents a communication session between two hosts. Our detection model leverages session edges to extract useful features that can support threat identification.

There are a total of nine session features:

- Total sessions: The total number of sessions between a pair of hosts.
- Unique destination ports: The number of unique ports of a destination host for which there are sessions from a source host.
- Unique destination hosts with same destination port: The number of unique destination hosts to which a source host connected with the same destination port.
- Unique destination ports for a source host: The number of unique destination ports for which a host has sessions.
- Mean time between sessions: The mean time between the start of the subsequent sessions between a pair of hosts.
- Mean session duration: The mean duration of the sessions between a pair of hosts.
- Mean session size ratio: The mean ratio of the destination size (bytes sent from the destination host of the session) over the source size (bytes sent from the source host of the session) for a pair of hosts.
- Mean session velocity: The mean velocity of the sessions between a pair of hosts. The session velocity is defined as the ratio of the total number of packets of a session to the total duration of the session, which is expressed in packets/sec.
- Mean session source size: The mean source size of the sessions sent from a host.

5.2.2 Authentication Features

The authentication data contained in the AEN graph are potential sources of useful information for the detection of anomalous authentication behaviour. There are a total of six different authentication features:

- Total authentication failures: The total number of failed authentication attempts between a pair of hosts.
- Total authentication failures per account per host: The total number of failed authentication attempts by a host using a specific account to all other hosts.

- Total authentication failures per account: The total number of failed authentication attempts between a pair of hosts using a specific account.
- Unique accounts: The number of unique accounts that were used in failed authentication attempts between a pair of hosts.
- Unique accounts per host: The number of unique accounts that were used in failed authentication attempts by a host to all other hosts.
- Unique target hosts per host per account: The total number of hosts that a host attempted, but failed, to authenticated using a specific account.

6 Experimental Evaluation

6.1 Setup and Procedures

The proposed ensemble intrusion detection mechanism was evaluated in two separate sets of experiments, one using the ISOT-CID Phase 1 dataset and one using the CIC-IDS2017 dataset. These datasets were selected because they contain benign and malicious network traffic data that can be used to build an AEN graph. Additionally, both datasets include examples of the attacks for which fingerprints have been developed, allowing their performance to be assessed.

The goal of the experiments was to evaluate the model's performance in correctly classifying hosts as malicious or benign. Specifically, each of the two detection schemes were assessed individually, and afterwards, a final ensemble classification was performed.

Separate experiments were performed for each day of each of the datasets according to the following procedure:

- An AEN graph was generated based on the available data of the specific day.
- Each of the two schemes were executed against the generated graph, and the results were collected.
- Each *host* node was given three classifications, one for each of the two detection schemes and one for the combined classification. The rules employed for each classifier are described later.
- The classification performance of each of the two individual schemes and that of the ensemble classifier were calculated based on the actual and predicted classifications.

The details for each of the two schemes are described in the following sections.

6.1.1 Fingerprint Matching

The fingerprint matching scheme classifies a host as malicious if the host is found to be part of an attack by at least one fingerprint.

Table 2 shows the parameters adopted for the experiment. Parameters with the same values used by multiple fingerprints are marked as "Multiple/Default", while parameters that are unique to a single fingerprint or values that are distinct from the default are marked according to the fingerprint.

Table 2: Attack fingerprint experiment parameters

Fingerprint	Parameter	Value
Multiple/Default	`attemptThr`	50
	`authFailRatioThr`	0.8
	`cntThr`	700
	`sessionThr`	100
	`sizeThr`	600 bytes
	`twSize`	20 seconds
	`twStep`	10 seconds
Basic Pwd Guessing	`attemptThr`	4
Credential Stuffing	`hostThr`	4
HTTP Flood	`pktCntThr`	15000
	`sizeThr`	1200 bytes
	`twSize`	2 minutes
	`twStep`	1 minute
IP Fragmentation Attack	`fragPckCntThr`	600
	`fragRatioThr`	0.8
	`sessionThr`	20
Spraying Pwd Guessing	`accountThr`	4
TCP SYN Flood	`synAckRatio`	100
UDP Flood	`sessionThr`	300
Vertical Port Scanning	`durThr`	1 second
	`portThr`	50

Finally, to measure the performance of the fingerprints, we used precision (positive predicted value – PPV) and sensitivity (true positive rate – TPR) because they describe the detection performance of the scheme without taking into consideration the true negatives, which is desirable for evaluating the performance of a signature-based intrusion detection scheme. Specifically, precision is better suited for the task because it describes the ratio of true positives among all predicted positive elements, which is expected to be high for a signature-based intrusion detection scheme. In contrast, sensitivity, indicates the ratio of true positives among all actual positive elements. This is not necessarily expected to be high given that the provided fingerprints only cover a few specific types of attacks and not all attack types that exist in the dataset.

6.1.2 Anomaly Detection

The anomaly detection scheme classifies a host as malicious if it is found to be the source of any anomalous behaviour, that is, if any of the features reports a score for the host above the experiment's threshold.

The algorithm has four parameters, bin width, time window size, time window step and threshold. To assess the performance of the scheme under different combinations of parameters, and identify the optimal threshold value, we defined the following set of values:

- Bin width: 1, 2, 4, 12 and 64.
- Time window size: 30 minutes, 1 hour, 4 hours and 12 hours.
- Time window step: Half the time window size.

- Threshold: From 0 to 20 bits at 0.5 intervals.

As a consequence, for each day experiment, the anomaly detection was executed 20 times, once for each parameter combination (excluding the threshold). Afterwards, the scores were evaluated against the 40 thresholds, resulting in a total of 800 sets of results.

Finally, to measure the performance of the anomaly detection scheme, three metrics were chosen: F1 score, bookmaker informedness (BM), and the Matthews correlation coefficient (MCC). As discussed later, both datasets are unbalanced; therefore, traditional metrics like the accuracy, precision and recall were not suitable, and as such, they were not used for the evaluation. Conversely, the three chosen metrics, particularly the latter two, were chosen because they are generally considered to be better suited for this scenario [49, 50].

Nonetheless, we provide the resulting receiver operating characteristic (ROC) curve for each of the parameter combinations, representing the classifications of the separate days combined together, to show the general performance behaviour of the scheme under different parameters and different thresholds. This serves as a basis for selecting the best parameter combinations and, accordingly, discussing the performance of the algorithm. We also present the sensitivity and the false positive rate (1 − specificity) when discussing the performance of the best parameter combination so it can be correlated to the ROC curves.

6.1.3 *Ensemble Classification*

The ensemble classification was performed by fusing the classification of the individual schemes (classifiers) in two ways: One with an *and* rule, meaning a host is only considered malicious if both classifiers agree that it is malicious, and the other with an *or* rule, meaning that a host is considered malicious if either of the classifiers consider it malicious.

Naturally, the *and* rule is expected to generate few false positives and more false negatives. In contrast, the *or* rule is expected to generate more false positives and few false negatives. In practice, the classifier with fewer positive predictions sets an upper limit on those numbers when using the *and* rule and a lower limit when the *or* rule is applied.

6.2 *ISOT-CID Phase 1*

The ISOT-CID Phase 1 dataset [5] contains systems calls, system and event logs, memory dumps and network traffic (TCPdump) data extracted from Windows and Linux virtual machines (VMs) and OpenStack Hypervisors collected from a production cloud computing environment, more specifically, Compute Canada's WestGrid. It includes both benign and malicious traces of several human-generated attacks and of unsolicited Internet traffic.

The dataset includes the time stamps and IP addresses related to each attack, as well as the IP addresses that generated benign traffic. There is also a label file that labels each packet in the dataset's network traffic data as benign or malicious, and the malicious packets are labelled by the type of attack. Unsolicited traffic is labelled as malicious but does not have an attack type label.

6.2.1 *Graph Generation*

In this study, we used only the network traffic data from which we extracted communication patterns between hosts, and system logs from which we extracted authentication information.

The graph elements are labelled based on the dataset labels, with *host* nodes labelled as malicious if they are the source of at least one packet labelled as malicious. The labels of other elements are derived from the host labels such that elements related to the host inherit its labels. For instance, a *session* edge is labelled as malicious if its source host is labelled as malicious. The malicious session edges are also labelled with the attack type when available. Note that labels are independent for each day of the data, meaning that they are not maintained from one day to the next.

The details of the generated AEN graphs for each day are shown in Table 3, and the sessions' attack type labels are shown in Table 4. As can be seen in both tables, there is a high prevalence of malicious hosts and sessions; however, most malicious sessions are not labelled with an attack type. Moreover, each day has at least a few samples of different known types of attacks, but there were no samples of any DoS attacks.

Table 3: Graph details for ISOT-CID Phase 1 dataset

Day	Nodes	Hosts (malicious)	Edges	Sessions (malicious)
Day 1	376	78 (60 – 77%)	12432	8313 (7279 – 88%)
Day 2	635	134 (116 – 87%)	45334	17544 (14276 – 81%)
Day 3	653	86 (70 – 81%)	31405	9355 (8741 – 93%)
Day 4	491	94 (78 – 83%)	8258	4637 (3882 – 84%)
Combined	2155	392 (324 – 83%)	97429	39849 (34178 – 86%)

Table 4: Malicious session attack type labels for ISOT-CID Phase 1 dataset. PG stands for password guessing, PS for post scanning and UL for unauthorized login.

Day	PG	Ping	PS	UL	Unknown
Day 1	21	1	3	13	7241
Day 2	38	–	11	4	14223
Day 3	20	–	12	2	8707
Day 4	–	–	–	2	3880
Combined	79	1	26	21	34178

6.2.2 *Fingerprint Matching Results*

The results obtained after running the fingerprints on the generated graph for each of the four days of the dataset are shown in Tables 5 and 6, with the former showing the classification performance of the proposed fingerprints combined for each day and the latter showing the individual performance of each fingerprint. Fingerprints for which no matches were found are omitted.

Table 5: Classification performance of the proposed fingerprint matching scheme for ISOT-CID Phase 1 dataset

Day	TP	TN	FP	FN	PPV	TPR
Day 1	28	17	1	32	0.97	0.47
Day 2	24	18	0	92	1.00	0.21
Day 3	38	16	0	32	1.00	0.54
Day 4	23	16	0	55	1.00	0.29
Combined	113	67	1	211	0.99	0.35

As shown in Table 5, the fingerprints had very high precision for all days of the dataset with only a single false positive match resulting in a combined precision of over 0.99, which, as discussed previously, was expected given that the scheme is signature-based and given the high prevalence of malicious hosts in the dataset. The sensitivity was medium to low, depending on the day, which was once again expected, given the small number of attack types covered by the fingerprints.

Table 6: Individual fingerprint performance for ISOT-CID Phase 1 dataset. PG stands for password guessing. Fingerprints for which no matches were found are omitted.

Fingerprint	Day 1		Day 2		Day 3		Day 4	
	TP	FP	TP	FP	TP	FP	TP	FP
Basic PG	24	0	19	0	35	0	22	0
Spraying PG	28	1	24	0	38	0	23	0

Looking at the performance of the individual fingerprints in Table 6, we can see that there were only two fingerprints for which matches were found, namely the basic password guessing fingerprint and the spraying password guessing fingerprint. To understand the reason for that, we need to refer back to Table 4, where two notable pieces of information are shown. The first is that there are no known samples of DoS attacks in the dataset, which means that none of those fingerprints were expected to be matched. The second is that, while there were a few port scanning attacks, they involved a very small number of sessions (the maximum being 26 on day 4), which maps to a small number of ports scanned in total since each session only has one destination port. Moreover, as the dataset documentation states, these attacks were horizontal scans targeting only a few ports across several hosts in the network, while the available port scanning fingerprint is designed for vertical scans. Therefore, it was expected to find no matches for that fingerprint.

Note that it would be possible for samples of those attacks to be unknowingly present in the dataset from the collected unsolicited traffic. However, no instances of those attacks were observed, except for some instances of password guessing attacks.

Also of note is the fact that the network data in the dataset were sampled and thus contain gaps that can skew some of the graph elements. This can also explain some false negatives and even cause false positives.

6.2.3 *Anomaly Detection Results*

After running the experiments as previously described, the results from different days were combined, and an ROC curve was plotted for each of the parameter combinations. The curves are shown in Figure 2. Marked in each plot is the point where the threshold is equal to 0.5.

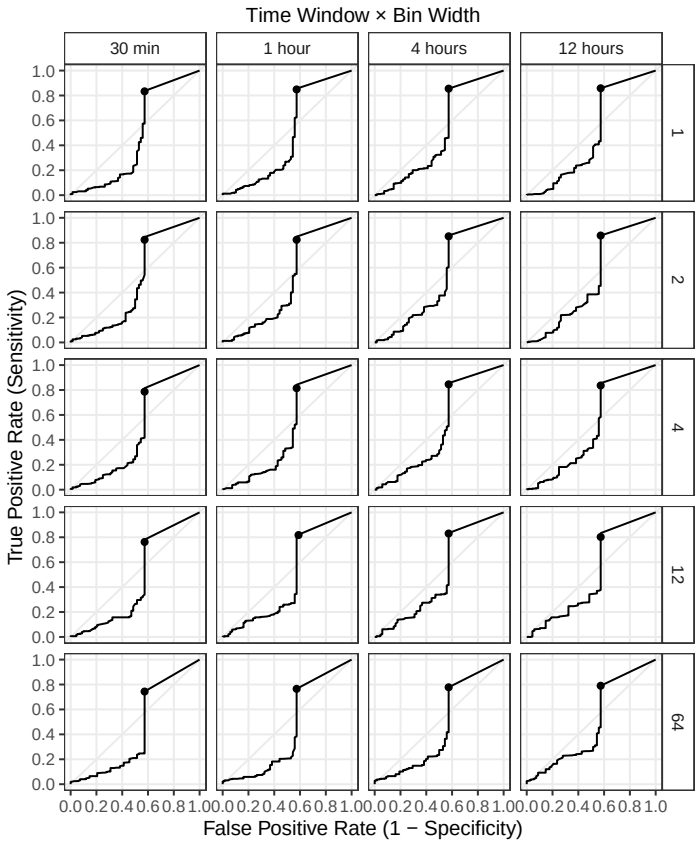


Figure 2: ROC curves of the anomaly detection scheme for the ISOT-CID Phase 1 dataset under different parameter combinations. The point in each plot marks where the threshold is equal to 0.5.

All curves show a similar pattern in which the sensitivity is poor while the threshold is high, until a point where it sharply rises until reaching its peak performance close to where the threshold is equal to 0.5. After that, it just goes straight to the top-right endpoint. Both of these characteristics can be explained by the exponential nature of the score, which means that a linear increase in the threshold will cause an exponential decrease in the true positives identified by the model. For this reason, it is common for the maximum score reported by any feature to be between 0 and 0.5, but the probability is exponentially smaller for higher scores.

When comparing the different ROC curves with regard to the other two parameters, a slightly better performance can be observed with longer time windows of 4 and 12 hours and with smaller bin widths of 1 and 2. That demonstrates that the extra information available with longer time windows allows the model to better distinguish anomalous behaviour. However, there is a limit to this, considering that too-long time windows could possibly be hiding shorter duration attacks. Also, the smaller bin widths allow for a greater variability of behaviours to be modelled. In practice, bin widths that are too large result in low resolutions of the distributions of variables that is caused by very diverse values being binned together.

These findings become even more clear when analyzing the

other performance metrics, which are more suitable for the unbalanced nature of this dataset. Therefore, we selected a threshold of 0.5, bin width of 2 and time window size of 12 hours to discuss the findings further. Table 7 shows the daily and combined results of the model under that specific parameter combination.

Table 7: Performance of the proposed anomaly detection model for the ISOT-CID Phase 1 dataset

Day	TP	TN	FP	FN	TPR	FPR	F1	BM	MCC
Day 1	52	8	10	8	0.87	0.56	0.85	0.31	0.32
Day 2	94	7	11	22	0.81	0.61	0.85	0.20	0.16
Day 3	61	7	9	9	0.87	0.56	0.87	0.31	0.31
Day 4	71	7	9	7	0.91	0.56	0.90	0.35	0.37
Comb.	278	29	39	46	0.86	0.57	0.87	0.28	0.27

As can be seen, under the selected parameters, the model was able to detect the majority of the malicious hosts but also generated a relatively high number of false positives. That behaviour can be observed in the other metrics as well, with a combined F1-score of 0.87, a combined BM of 0.28 and a combined MCC of 0.27. Another notable aspect is the mostly consistent performance observed for each individual day, with only the results for day 2 having a greater deviation from the average.

In general, the observed behaviour was expected, given that the model is anomaly-based and thus prone to generating false alarms. Moreover, the high prevalence of hosts in the dataset means that the malicious behaviour is in fact not anomalous in the dataset. On the contrary, most of the traffic and hosts are labelled as malicious, which explains why the scheme generated a high number of false positives and points to a general limitation of anomaly-based detection, which can produce degraded results when the malicious behaviour is not uncommon.

6.2.4 *Ensemble Classification Results*

The ensemble classification using the *and* rule resulted in the same predictions as the fingerprint matching already shown in Table 5. This outcome means that all hosts classified as malicious by fingerprint matching were also classified as malicious with the anomaly detection. In contrast, but for the same reason, the ensemble classification using the *or* rule resulted in the same predictions as the as anomaly detection, which are summarized in Table 7.

In practice, in a real environment, where the anomaly detection threshold is not optimally chosen, the divergence between the classifiers will be greater, thus causing the performance of the ensemble classification to be distinct. Other ensemble classification rules, such as soft voting could also be used as a middle ground between the two rules evaluated in this paper.

6.3 *CIC-IDS2017*

The CIC-IDS2017 dataset [6] contains network traffic data (both pcap and NetFlow) of benign traffic, as well as several samples of attack scenarios including SSH and FTP password guessing, DoS, web attacks and instances of host infiltration. The dataset is labelled on the flow level, with each flow being labelled as either benign or with the attack performed.

6.3.1 *Graph Generation*

To build the graph, the packet data (pcap files) were used to extract the communication patterns between hosts since packets are more finely detailed than flow data. Because no system or application logs were available, we could not extract authentication information from the data. As a consequence, no authentication-related elements (*account* nodes and *authentication attempt* edges) are present in the graphs of this dataset which in turn means that no password guessing instances can be found with the fingerprints as designed.

The graph labelling followed the same rules as the previous dataset, with *host* nodes labelled as malicious if they were the source of at least one flow labelled as malicious and with the labels of other elements being derived from the host labels. One distinction was how to define the attack type labels. For that, we first combined the dataset labels into generic attack type labels. For instance, the "DoS slowloris" and "DoS GoldenEye" labels were combined into the "Denial of Service" label. Then, each flow was matched with its respective *session* edge that was labelled with the generic attack type.

In a few cases, sessions had more than one attack type. This was an artifact of how the sessions are created from packets such that one session can map to more than one flow. Moreover, some malicious sessions have no attack type labels in cases where none of their mapped flows were labelled as malicious, even though their source hosts were labelled as such.

The details of the generated AEN graphs for each day are shown in Table 8, and the sessions' attack type labels are shown in Table 9. As shown by the tables, there was a very small prevalence of malicious hosts, while the prevalence of malicious sessions varied from low (approximately 3% on day 1) to high (approximately 87% on day 4). As for the attack type labels, there was a high number of attack samples from all days. However, the types of attacks present for each day varied with most types of attacks only present for one day. Moreover, the number of sessions with unknown attack was high on days 3 and 4, which stems from the fact that those days had combined multiple attack sessions.

Table 8: Graph details for the CIC-IDS2017 dataset

Day	Nodes	Hosts (malicious)	Edges	Sessions (malicious)
Day 1	27653	8498 (1 – 0.01%)	279271	243264 (6954 – 3%)
Day 2	29216	9017 (1 – 0.01%)	298033	259938 (16571 – 6%)
Day 3	27828	8545 (2 – 0.02%)	323502	287272 (89523 – 31%)
Day 4	27035	8331 (10 – 0.12%)	460893	425649 (370297 – 87%)
Comb.	111732	34391 (14 – 0.04%)	1361699	1216123 (483345 – 40%)

Table 9: Malicious session attack type labels for CIC-IDS2017 dataset. BN stands for botnet, Inf for infiltration, PG for password guessing, PS for port scanning, WA for web attack and Unk for unknown.

Day	BN	DoS	Inf	PG	PS	WA	Unk
Day 1	–	–	–	6953	–	–	1
Day 2	–	16537	1	–	–	–	33
Day 3	–	–	6	1363	–	643	87511
Day 4	1228	45392	–	–	158678	–	165026
Comb.	1228	61929	7	8316	158678	643	252571

6.3.2 *Fingerprint Matching Results*

The results obtained after running the fingerprints on the generated graph for each of the four days of the dataset are shown in Tables 10 and 11, with the former table showing the classification performance of the proposed fingerprints combined for each day and the latter showing the individual performance of each fingerprint. Fingerprints for which no matches were found are omitted.

Table 10: Classification performance of the proposed fingerprints for the CIC-IDS2017 dataset

Day	TP	TN	FP	FN	PPV	TPR
Day 1	0	8494	3	1	0.00	0.00
Day 2	0	9011	5	1	0.00	0.00
Day 3	2	8542	1	0	0.67	1.00
Day 4	6	8320	1	4	0.86	0.60
Combined	8	34367	10	6	0.44	0.57

Table 11: Individual fingerprint performance for the CIC-IDS2017 dataset. PS stands for port scanning. Fingerprints for which no matches were found are omitted.

Fingerprint	Day 1		Day 2		Day 3		Day 4	
	TP	FP	TP	FP	TP	FP	TP	FP
HTTP Flood	0	0	0	0	0	0	1	0
TCP SYN Flood	0	0	0	0	2	0	1	0
UDP Flood	0	1	0	2	0	0	1	0
Vertical PS	0	2	0	4	2	1	5	1

As shown in Table 10, the general performance was not as high as with the previous dataset. There were no true positive matches on days 1 and 2, resulting is precision and sensitivity of 0 for those days. In contrast, days 3 and 4 had better performance, particularly day 4, with a precision of 0.86.

Having no true positives was expected for day 1 because this day only had password guessing attacks but graph had no authentication elements, which are part of the password guessing fingerprints. This was not the case for day 2, which had DoS attacks that were HTTP-based, but no matches were found for the HTTP flood fingerprint. Still, this can be explained by the types of attacks performed, such as Heartbleed and Slowloris, which are not flood attacks, making the HTTP flood fingerprint unsuitable for this case. Tuning the fingerprint parameters might allow for these attacks to be found but might also result in some false positives. Moreover, the very low prevalence of malicious hosts, with only a single one for both days 1 and 2, means that not finding that host will result in a precision and sensitivity of 0 as observed.

Continuing onto day 3, there were matches for the two malicious hosts for both the TCP SYN flood fingerprint and the vertical port scanning fingerprint. Note that according to the dataset labels, as shown in Table 9, neither type of attack was expected to be present. However, the dataset documentation states that both port scans and nmap scans were performed on that day, although not labelled, which explains the positive matches.

As for day 4, there were matches for four different fingerprints, including three DoS fingerprints and the port scanning fingerprint. The day's data are labelled as having both of those types of attacks, as well as a botnet attack. Sessions with all three labels were matched. Note that the dataset's documentation is not clear on exactly which attacks were executed as part of the botnet attack, but in any case, the attack's data were the source of some of the matches, too.

Finally, all days had false positives, which would not be expected from a signature-based scheme; however, the absolute number of false positives was low compared to the total number of hosts in the dataset. Moreover, as shown in Table 11, this was mostly from the vertical port scanning fingerprint. When analyzing the benign sessions that were matched by that fingerprint, almost 60% had port UDP/137, which is used by the NetBIOS name service. However, in these cases, it was not the destination port that was fixed at 137, but instead, the source port was 137, while the destination port varied. This behaviour is an artifact of how name queries are broadcast in NetBIOS, but the replies are directed to the host that made the query on what was originally the source port of the broadcast query. The rest of the false positives do not have such a clear explanation. Tuning the fingerprint parameters could reduce them but would also reduce the detection rate.

6.3.3 *Anomaly Detection Results*

After running the experiments as previously described, the results from different days were combined, and an ROC curve was plotted for each of the parameter combinations. The curves are shown in Figure 3. Marked in each plot is the point where the threshold is equal to 0.5.

All parameter combinations showed high levels of performance, with an area under the ROC curve (AUC) of over 0.99. However, that metric by itself is deceiving, given the highly unbalanced nature of the dataset. In reality, the model was able to detect all 14 malicious hosts with thresholds between 0.5 to 5.5 under most parameter configurations, but it reported a decreasing number of false positives as the threshold increased. Nonetheless, the relative variation was small in terms of the total number of benign hosts in the dataset.

Although not distinguishable from the ROC curves, the behaviour observed for the previous dataset with regards to the bin width can also be observed for this dataset when analyzing the other metrics, with values of 1 and 2 resulting in a better performance on average than the others. However, the benefit of a longer time window parameter is not as clear for this dataset, which shows a more mixed performance with different values for this parameter.

Based on that, we selected a threshold of 5, bin width of 1 and time window size of 12 hours for further discussion. Table 12 shows the daily and combined results of the model under that specific parameter combination.

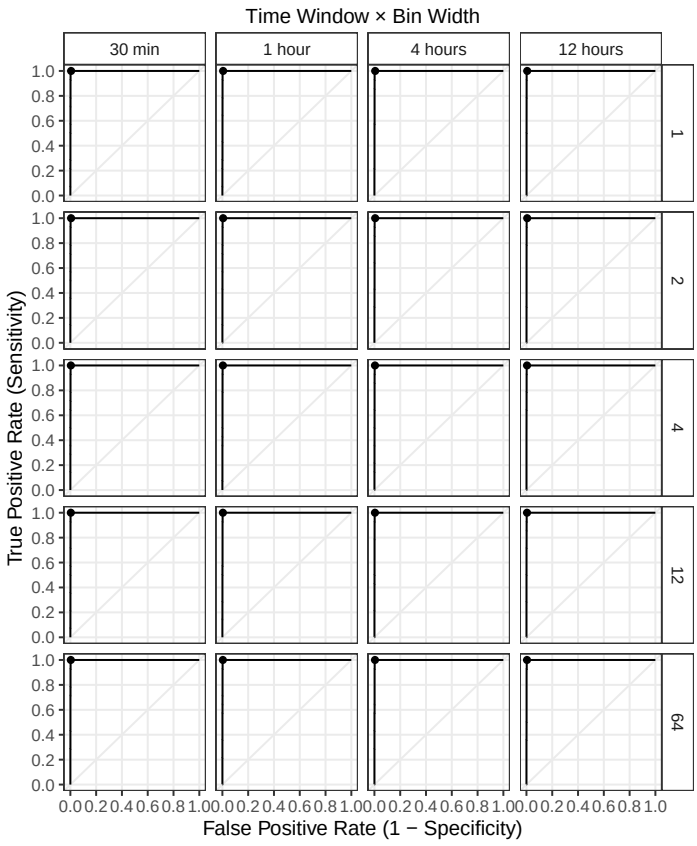


Figure 3: ROC curves of proposed scheme under different parameter combinations for the CIC-IDS2017 dataset. The point in each plot marks where the threshold is equal to 0.5.

Table 12: Performance of the proposed anomaly detection model for the CIC-IDS2017 dataset

Day	TP	TN	FP	FN	TPR	FPR	F1	BM	MCC
Day 1	1	8481	16	0	1.00	0.01	0.11	0.99	0.24
Day 2	1	9001	15	0	1.00	0.01	0.12	0.99	0.25
Day 3	2	8529	14	0	1.00	0.01	0.22	0.99	0.35
Day 4	10	8313	8	0	1.00	0.01	0.71	0.99	0.75
Comb.	14	34324	53	0	1.00	0.01	0.34	0.99	0.46

As can be seen, under the selected parameters, the model was able to detect all malicious hosts in all days with a relatively low number of false positives. As a point of comparison, using the same parameters as the previous dataset (threshold of 0.5 and bin width of 2) would not affect the number of true positives but would result in an extra 145 false positives and only slightly worse performance.

Because of the very small prevalence of malicious hosts in the dataset, the observed performance resulted in a low F1-score and a medium MCC for all days except day 4, which had a higher number of malicious hosts that can compensate for the false positives. Conversely, the BM was above 0.99 for all four days.

In summary, the scheme obtained very good results for this dataset with a varied number of parameter configurations. This shows that the scheme is suitable for detecting malicious behaviour when the prevalence of malicious hosts is low and, therefore anomalous.

6.3.4 *Ensemble Classification Results*

As with the previous dataset, the ensemble classification using the *and* rule resulted in the same predictions as the fingerprint matching shown in Table 10, while the the ensemble classification using the *or* rule resulted in the same predictions as the anomaly detection. These are summarized in Table 12.

This again shows that hosts that were classified as malicious by fingerprint matching were also classified as malicious with the anomaly detection.

7 Conclusion

This paper presented an unsupervised ensemble intrusion detection mechanism composed of two detection schemes, one signature-based that employs isomorphic subgraph matching of graphical patterns of known attacks, called attack fingerprints, and one anomaly-based, which consists of an anomaly score developed based on the work of Ferragut et al. [4].

To validate the proposed scheme we presented a collection of attack fingerprints for the AEN graph model, which were expressed using PGQL and covered common attacks, such as port scanning, DoS and password guessing, along with a subgraph matching algorithm specific for finding subgraphs isomorphic to the fingerprints. Furthermore, a total of 15 anomaly features, including nine extracted from session data and six extracted from authentication data.

The proposed schemes were evaluated individually and as an ensemble in the capacity for identifying malicious hosts using two datasets: The ISOT-CID Phase 1 dataset and the CIC-IDS2017 dataset.

The evaluation of the fingerprint matching scheme showed a combined precision of 0.99 and a combined sensitivity of 0.35 for the former dataset, while the latter dataset resulted in a combined precision of 0.44 and a combined sensitivity of 0.57. The observed results are promising, particularly considering the limited number of fingerprints available and the specific types of errors encountered. Ultimately, they demonstrate that the method is capable of identifying known attacks and is particularly suited to identifying stealth attacks, which is a weakness of traditional signature-based intrusion detection systems.

The evaluation of the anomaly detection was particularly encouraging for the CIC-IDS2017 dataset, with a BM of over 0.99 and an MCC of 0.46. This shows that the scheme has a high capacity for detecting anomalous behaviour when there was a low prevalence of malicious elements in the network.

The evaluation of the ensemble classification showed that one classifier could end up dominating the ensemble classification. This underscores the need for future exploration of possible benefits from using ensemble classification rules, such as soft voting.

As limitations, we identified the amount of effort needed to create the fingerprints, the binning of the values anomaly features that can result in suboptimal distributions and the high computational power required to build and maintain the graph. Another limitation is the high computation cost involved in searching for isomorphic subgraphs given the high complexity of the subgraph matching algorithms, although the use of indexes helps alleviate that issue with the cost of extra memory requirements.

In our future work, we aim to improve and extend the proposed fingerprint database to add other types of attacks, particularly those that are traditionally harder to detect, such as HTTP request smuggling, and to increase the feature space of the anomaly detection model by introducing more features. We believe this can help improve detection accuracy, particularly in environments that have a high prevalence of malicious hosts.

We also plan to implement more advanced classification rules as part of the anomaly detection scheme and also in the ensemble classification of the two detection schemes. In addition, we plan to employ adaptive bin width values according to the value range of each given variable to improve the fitness of the bin distributions.

Finally, we aim to conduct further evaluations using other datasets, such as the ISOT-CID Phase 2 dataset and the 2018 CIC Intrusion Detection Evaluation Dataset (CSE-CIC-IDS2018) [51].

References

[1] C. Nie, P. G. Quinan, I. Traoré, I. Woungang, "Intrusion Detection using a Graphical Fingerprint Model," in 2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid), 806–813, 2022, doi:10.1109/CCGrid54584.2022.00095.

[2] R. Sommer, V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in Proceedings of the 2010 IEEE Symposium on Security and Privacy, SP '10, 305–316, IEEE Computer Society, Washington, DC, USA, 2010, doi:10.1109/SP.2010.25.

[3] A. Aldribi, I. Traoré, B. Moa, O. Nwamuo, "Hypervisor-based cloud intrusion detection through online multivariate statistical change tracking," Computers & Security, **88**, 2020, doi:10.1016/j.cose.2019.101646.

[4] E. M. Ferragut, J. A. Laska, R. A. Bridges, "A New, Principled Approach to Anomaly Detection," 2012 11th International Conference on Machine Learning and Applications, **2**, 210–215, 2012, doi:10.1109/ICMLA.2012.151.

[5] A. Aldribi, I. Traore, B. Moa, Data Sources and Datasets for Cloud Intrusion Detection Modeling and Evaluation, 333–366, Springer International Publishing, Cham, 2018, doi:10.1007/978-3-319-73676-1_13.

[6] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization." in ICISSP, 108–116, 2018, doi:10.5220/0006639801080116.

[7] C. Phillips, L. P. Swiler, "A Graph-based System for Network-vulnerability Analysis," in Proceedings of the 1998 Workshop on New Security Paradigms, NSPW '98, 71–79, ACM, New York, NY, USA, 1998, doi:10.1145/310889.310919.

[8] O. Sheyner, S. Haines, Jand Jha, R. Lippmann, J. M. Wing, "Automated generation and analysis of attack graphs," in Proceedings of the Symposium on Security and Privacy, IEEE, 2002, doi:10.1109/SECPRI.2002.1004377.

[9] S. Jha, O. Sheyner, J. Wing, "Two formal analyses of attack graphs," in Proceedings 15th IEEE Computer Security Foundations Workshop. CSFW-15, 49–63, 2002, doi:10.1109/CSFW.2002.1021806.

[10] X. Ou, G. Sudhakar, A. A. W., "MulVAL: A Logic-based Network Security Analyzer," in Proceedings of USENIX Security Symposium, volume 8, 2005, doi:10.5555/1251398.1251406.

[11] K. Ingols, R. Lippmann, K. Piwowarski, "Practical Attack Graph Generation for Network Defense," in 2006 22nd Annual Computer Security Applications Conference (ACSAC'06), 121–130, 2006, doi:10.1109/ACSAC.2006.39.

[12] L. Akoglu, H. Tong, K. D., "Graph based Anomaly Detection and Description: A Survey," Journal Data Mining and Knowledge Discovery, **29**(3), 626–688, 2015, doi:10.1007/s10618-014-0365-y.

[13] F. Jemili, M. Zaghdoud, M. B. Ahmed, "Intrusion detection based on "Hybrid" propagation in Bayesian Networks," 2009 IEEE International Conference on Intelligence and Security Informatics, 137–142, 2009, doi:10.1109/ISI.2009.5137285.

[14] P. Xie, J. H. Li, X. Ou, P. Liu, R. Levy, "Using Bayesian networks for cyber security analysis," 2010 IEEE/IFIP International Conference on Dependable Systems & Networks (DSN), 211–220, 2010, doi:10.1109/DSN.2010.5544924.

[15] L. Xiao, Y. Chen, C. K. Chang, "Bayesian Model Averaging of Bayesian Network Classifiers for Intrusion Detection," 2014 IEEE 38th International Computer Software and Applications Conference Workshops, 128–133, 2014, doi:10.1109/COMPSACW.2014.25.

[16] K. K. Gupta, B. Nath, K. Ramamohanarao, "Conditional Random Fields for Intrusion Detection," in 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07), volume 1, 203–208, IEEE, 2007, doi:10.1109/AINAW.2007.126.

[17] H. Ma, Y. Xie, S. Tang, J. Hu, X. Liu, "Threat-Event Detection for Distributed Networks Based on Spatiotemporal Markov Random Field," IEEE Transactions on Dependable and Secure Computing, **19**(3), 1735–1752, 2022, doi:10.1109/TDSC.2020.3036664.

[18] K. Peng, V. C. M. Leung, L. Zheng, S. Wang, C. Huang, T. Lin, "Intrusion Detection System Based on Decision Tree over Big Data in Fog Environment," Wireless Communication and Mobile Computing, **2018**, 2018, doi:10.1155/2018/4680867.

[19] C. Yin, Y. Zhu, J. long Fei, X.-Z. He, "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks," IEEE Access, **5**, 21954–21961, 2017, doi:10.1109/ACCESS.2017.2762418.

[20] Y. Zhang, P. Li, X. Wang, "Intrusion Detection for IoT Based on Improved Genetic Algorithm and Deep Belief Network," IEEE Access, **7**, 31711–31722, 2019, doi:10.1109/ACCESS.2019.2903723.

[21] Z. Wang, Y. Zeng, Y. Liu, D. Li, "Deep Belief Network Integrating Improved Kernel-Based Extreme Learning Machine for Network Intrusion Detection," IEEE Access, **9**, 16062–16091, 2021, doi:10.1109/ACCESS.2021.3051074.

[22] S. A. Cook, "The complexity of theorem-proving procedures," in Proceedings of the third annual ACM symposium on Theory of computing, 151–158, 1971, doi:10.1145/800157.805047.

[23] J. E. Hopcroft, J.-K. Wong, "Linear time algorithm for isomorphism of planar graphs (Preliminary Report)," in Proceedings of the sixth annual ACM symposium on Theory of computing, 172–184, 1974, doi:10.1145/800119.803896.

[24] J. R. Ullmann, "An algorithm for subgraph isomorphism," Journal of the ACM (JACM), **23**(1), 31–42, 1976, doi:10.1145/321921.321925.

[25] L. P. Cordella, P. Foggia, C. Sansone, M. Vento, "A (sub)graph isomorphism algorithm for matching large graphs," IEEE Transactions on Pattern Analysis and Machine Intelligence, **26**, 1367–1372, 2004, doi:10.1109/TPAMI.2004.75.

[26] W.-S. Han, J. Lee, J.-H. Lee, "Turboiso: towards ultrafast and robust subgraph isomorphism search in large graph databases," in SIGMOD '13, 2013, doi:10.1145/2463676.2465300.

[27] M. Han, H. Kim, G. Gu, K. Park, W.-S. Han, "Efficient Subgraph Matching: Harmonizing Dynamic Programming, Adaptive Matching Order, and Failing Set Together," Proceedings of the 2019 International Conference on Management of Data, 2019, doi:10.1145/3299869.3319880.

[28] P. G. Quinan, I. Traoré, I. Woungang, "Activity and Event Network Graph and Application to Cyberphysical Security," in I. Traoré, I. Woungang, S. Saad, editors, Artificial Intelligence for Cyber-Physical Systems Hardening, chapter 10, 217–233, Springer, 2022, doi:10.1007/978-3-031-16237-4_10.

[29] O. van Rest, S. Hong, J. Kim, X. Meng, H. Chafi, "PGQL: a property graph query language," in GRADES '16, 2016, doi:10.1145/2960414.2960421.

[30] N. Francis, A. Green, P. Guagliardo, L. Libkin, T. Lindaaker, V. Marsault, S. Plantikow, M. Rydberg, P. Selmer, A. Taylor, "Cypher: An Evolving Query Language for Property Graphs," Proceedings of the 2018 International Conference on Management of Data, 2018, doi:10.1145/3183713.3190657.

[31] M. H. Bhuyan, D. K. Bhattacharyya, J. K. Kalita, "Surveying Port Scans and Their Detection Methodologies," The Computer Journal, **54**, 1565–1581, 2011, doi:10.1093/comjnl/bxr035.

[32] S. Staniford, J. A. Hoagland, J. M. McAlerney, "Practical Automated Detection of Stealthy Portscans," Journal of Computer Security, **10**, 105–136, 2002, doi:10.3233/JCS-2002-101-205.

[33] M. De Vivo, E. Carrasco, G. Isern, G. O. de Vivo, "A review of port scanning techniques," ACM SIGCOMM Computer Communication Review, **29**(2), 41–48, 1999, doi:10.1145/505733.505737.

[34] J. Mirkovic, P. L. Reiher, "A taxonomy of DDoS attack and DDoS defense mechanisms," Comput. Commun. Rev., **34**, 39–53, 2004, doi:10.1145/997150.997156.

[35] R. Tandon, "A Survey of Distributed Denial of Service Attacks and Defenses," ArXiv, **abs/2008.01345**, 2020, doi:10.48550/arXiv.2008.01345.

[36] E. Cambiaso, G. Papaleo, G. Chiola, M. Aiello, "Slow DoS attacks: definition and categorisation," International Journal Trust Management in Computing and Communications, **1**, 300–319, 2013, doi:10.1504/IJTMCC.2013.056440.

[37] M. Bogdanoski, T. Suminoski, A. Risteski, "Analysis of the SYN flood DoS attack," International Journal of Computer Network and Information Security (IJCNIS), **5**(8), 1–11, 2013, doi:10.5815/IJCNIS.2013.08.01.

[38] V. K. Yadav, M. C. Trivedi, B. Mehtre, "DDA: an approach to handle DDoS (Ping flood) attack," in Proceedings of International Conference on ICT for Sustainable Development, 11–23, Springer, 2016, doi:10.1007/978-981-10-0129-1_2.

[39] T. H. Ptacek, T. N. Newsham, "Insertion, Evasion, and Denial of Service: Eluding Network Intrusion Detection," Technical report, Secure Networks inc Calgary Alberta, 1998.

[40] "Transmission Control Protocol," RFC 793, 1981, doi:10.17487/RFC0793.

[41] MazeBolt, "Layer 4 — MazeBolt Knowledge Base," .

[42] A. Bijalwan, M. Wazid, E. S. Pilli, R. C. Joshi, "Forensics of random-UDP flooding attacks," Journal of Networks, **10**(5), 287, 2015, doi:10.4304/jnw.10.5.287-293.

[43] I. Sreeram, V. P. K. Vuppala, "HTTP flood attack detection in application layer using machine learning metrics and bio inspired bat algorithm," Applied Computing and Informatics, 2019, doi:10.1016/j.aci.2017.10.003.

[44] C. Paar, J. Pelzl, B. Preneel, "Understanding Cryptography: A Textbook for Students and Practitioners," 2010, doi:10.1007/978-3-642-04101-3.

[45] D. Wang, Z. Zhang, P. Wang, J. Yan, X. Huang, "Targeted online password guessing: An underestimated threat," in Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, 1242–1254, 2016, doi:10.1145/2976749.2978339.

[46] Mitre, "Brute Force: Password Spraying," .

[47] M. Patton, E. Gross, R. Chinn, S. Forbis, L. Walker, H. Chen, "Uninvited connections: a study of vulnerable devices on the internet of things (IoT)," in 2014 IEEE joint intelligence and security informatics conference, 232–235, IEEE, 2014, doi:10.1109/JISIC.2014.43.

[48] G. Tandon, P. K. Chan, "Tracking user mobility to detect suspicious behavior," in Proceedings of the 2009 SIAM International Conference on Data Mining, 871–882, SIAM, 2009, doi:10.1137/1.9781611972795.75.

[49] A. Luque, A. Carrasco, A. Martín, A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognit., **91**, 216–231, 2019, doi:10.1016/j.patcog.2019.02.023.

[50] D. Chicco, G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, **21**, 2020, doi:10.1186/s12864-019-6413-7.

[51] Canadian Institute for Cybersecurity, "CSE-CIC-IDS2018 on AWS: A collaborative project between the Communications Security Establishment (CSE) & the Canadian Institute for Cybersecurity (CIC)," .

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 1, 44-56 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Nonlinear Model Predictive Control of Rover Robotics System

Serdar Kalaycioglu*, Anton de Ruiter

Department of Aerospace Engineering, Toronto Metropolitan University, Toronto, M5B 2K3, Canada

ARTICLE INFO

Article history:
Received: 30 September, 2022
Accepted: 23 December, 2022
Online: 24 January, 2023

Keywords:
NMPC
Optimal Control
Multi Rover Control

ABSTRACT

The paper presents two robust and efficient control algorithms based on (i) Optimal Control Allocation (OCA) and (ii) Nonlinear Model Predictive Control (NMPC). The robotics system consists of two rovers with mecanum wheels and mounted two 7-DOF arms carrying a common load. The overall system is an underdetermined one with non-holonomic constraints. The developed control algorithms focus on providing an optimal solution to the wheel and joint torque saturation problem, which is typically encountered while manipulating a large and heavy payload. The first control algorithm based on OCA minimizes a quadratic cost function consisting of robot joint and rover wheel torques, contact forces, and moments using only the current state values and the system dynamics. It is computationally very efficient. The NMPC algorithm minimizes a quadratic cost function which not only includes the current states but also the future state estimates, and the control inputs over a specified prediction horizon. The system consisting of multi-rover with a dual arm is highly non-linear. The linear MPC technique on which most of the previous studies relied is not adequate. On the other hand, the computational difficulties of a generic NMPC algorithm is remarkably high. In this paper, an elegant, discretized technique with exact realization is implemented to take into account the full non-linear model and yet provide a simple real-time solution satisfying a minimum performance index subject to constraints. The results show that the developed control algorithms OCA and NMPC work efficiently, and the minimum the contact moments and forces, and the joint torques are realized while two arms carry a common load and successfully track a reference end-effector trajectory. The results also indicate that although NMPC algorithm is computationally more involved, it provides superior results in reducing joint and wheel torques as well as contact moments and forces.

1. Introduction

This paper is an extension of the work originally presented in IEMTRONICS [1]. The Optimal Control Allocation algorithm (OCA) presented in the original work is further extended to accommodate a Nonlinear Model Predictive Control technique to increase performance of the approach.

There has been a significant interest in exploring complex environments using mobile rovers. Such rovers are commonly used in space exploration, construction, mining, and military.

Especially, there has been a considerable amount of interest in Space Robotics Exploration missions in the last two decades. Similar to on-orbit robotics missions (e.g., servicing, assembly, and manufacturing), the future planetary exploration missions will also include tasks such as assembly of large space structures using multiple coordinating rovers and the rover-mounted robotics manipulators. Recently the Moon and Mars rover missions are the main target of various space agencies including NASA, Canadian Space Agency, ESA, JAXA, etc. Most of these space agencies in collaboration with space industries and research centers are heavily focusing on innovative rover technologies and designs. Autonomous rover motion control capability has been identified as one of the critical and enabling technology requirement for such systems. Although, there is a significant amount of research studies in the fields of control of single rover trajectory and force control of fixed-based arms, there are still major research challenges in the areas of load sharing multi- rovers and arms, particularly, real-time force and motion control when they are carrying a common load.

* Corresponding Author: Serdar Kalaycioglu, TMU, Department of Aerospace Engineering, Toronto, Canada, email: skalay@torontomu.ca

The initial technological challenges that involved designing a mobile rover were related to its mechanics. These included the development of dynamic control systems and collision free trajectories.

In order to develop effective control systems for mobile rovers, a team led by Necsulescu [2,3] studied the free and contact motion of the vehicles. They also developed methods to generate collision free trajectories and perform force control.

Motion control of rovers with nonholonomic constraints were studied using differential wheeled rovers in [4,5] These constraints exist if the constraints cannot be expressed in the form of time derivatives of a function consists of the generalized coordinates.

There have been extensive studies in control of systems with non-holonomic constraints. However, most of the cases, kinematic control is typically achieved by ignoring the dynamics when dealing with systems with non-holonomic constraints [6]. However, it has been shown that a mechanical system with these constraints were controlled in spite of its structure [7]. In addition, it has been shown that a non-holonomic system cannot be brought to a single equilibrium with a smooth time-invariant feedback [8].

In a study conducted in Kalaycioglu [9], a control technique with optimal force distribution for multiple robotic manipulators was demonstrated. However, it only involved two cooperating arms and did not include rovers.

The use of a Model Predictive Control (MPC) framework facilitates the optimization of a given performance index. It also allows for the analysis of the system constraints and dynamics [10–15]. One of the most challenging aspects of implementing a robust model of (MPC) is dealing with the various uncertainties that can impact its performance [16]. Due to the characteristics of the model's receding horizon, standard MPC can provide an adequate level of robustness [17].

Unfortunately, the literature has shown that standard MPC cannot provide an adequate performance in complex robotics systems [18]. To address this issue, various research studies have been conducted to develop novel MPC methods that can provide a robust and stable performance [19–23].

The scope and capabilities of Non-linear Model Predictive Control (NMPC) have significantly improved over the past few years. Due to the increasing number of tools that can be used to implement this type of model, the performance of this algorithm has been greatly improved. Some of these include the ability to perform fast gradient use and input parameterisation [24–27]. The application of NMPS for free-floating space manipulator are provided in [28–31].

The mechanics of wheeled locomotion have also attracted a lot of attention [32–37][. A number of studies have been conducted on the dynamics and kinematics of the mecanum wheel (a subcategory of omnidirectional wheel) [38–43].

There has been a significant amount of research on the various aspects of wheeled locomotion, but it is still not yet feasible to fully understand the mechanisms involved in the movement control of multiple rovers and mounted arms. For instance, the development of systems with multi- rovers with dual manipulators that can perform real-time trajectories while manipulating a common load is still in its early stages.

This paper presents two robust and efficient control algorithms based on (i) Optimal Control Allocation (OCA) and (ii) Nonlinear Model Predictive Control (NMPC) for a rover robotics system with mecanum wheels when the two 7-DOF arms operating a common load. The system is an underdetermined one subject to non-holonomic constraints. The control algorithms focus on providing an optimal solution to the wheel and joint torque saturation problem, which is typically encountered while manipulating a large and heavy payload.

The first control algorithm based on OCA minimizes a quadratic cost function (a performance index) consisting of robot joint and rover wheel torques, contact forces, and moments using only the current state values and the system dynamics. It is computationally very efficient. The NMPC algorithm minimizes a quadratic cost function which not only includes the current states but also the future state estimates, and the control inputs over a specified prediction horizon.

The literature on the application of MPC for robotics is mainly focused on linear models. However, the multi-rover dual arm coordinating system is highly non-linear and MPC lacks robust applications in this area. In this paper, we present a novel NMPC discretized technique that incorporates the full non-linear characteristics of the multi-rover dual arm system.

This paper consists of four sections. The first section provides the mathematical formulations such as the kinematics and dynamics models of the total system including two *n*-degree redundant manipulators, two rovers and a common load. The second section presents two novel control algorithms based on optimal control allocation (OCA) and non-linear model predictive control (NMPC) which are formulated to minimize the wheel moments, the joint torques, and contact moments/forces. The third section provides the simulation results and discussion, and the fourth section provides some concluding remarks and recommendations for future work.

2. Theoretical Formulations

2.1. The Rover Robotics System

The system includes two mobile rovers with four mecanum wheels and two *n*-DOF redundant arms attached on the two rovers carrying a common load. Figure 1 shows an example of such a system with two rovers and two *n*-degree arms.

Table 1 contains the rover and robotics parameters utilized in the computer simulations.. The rotation angle ψ_i and the position vector $\widetilde{\boldsymbol{R}}_{\boldsymbol{ci}}$, provide the pose of the center of mass C_i of the ith rover-in the inertial coordinate system, X, Y, Z. The coordinate axes x_{ci} , y_{ci} , z_{ci} attached to point *Ci* are obtained via a rotation around Z-axis with an angle of ψ_i.

The masses associated with the rovers and the wheels are given as m_{ci} and m_{wij}, respectively for the ith rover and the jth wheel, , j=1...4 and i=1,2 for each rover. The distances between the wheel centers along the y_{ci} and x_{ci}-axes are denoted by *2a* and *2b*, respectively.

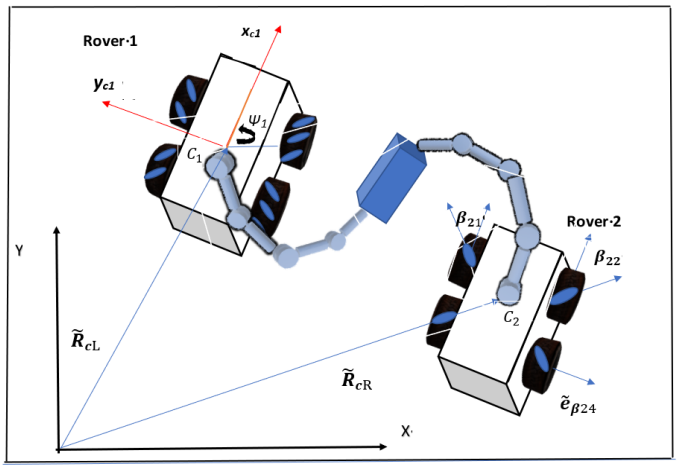


Figure 1: Description of the rover robotics system

The wheels have a radius of s and the angle of rotation, and the angular rate are denoted as $\phi_{ij}$and, ω_{ij} , respectively. The rollers are attached to the outer rims of the mecanum wheels as illustrated in Figure 1. The angle βij is defined as the angle between the axis of rotation of the roller and the x_{ci} of the j[th] wheel of the i[th] rover.

2.2. Model of Kinematics

$\tilde{\boldsymbol{V}}_{\boldsymbol{mij}}$, the velocity vector of the center of the j[th] wheel of the i[th] rover can be determined by the following relationship:

$$\tilde{\boldsymbol{V}}_{\boldsymbol{mij}} = \tilde{\boldsymbol{V}}_{\boldsymbol{ci}} + \tilde{\boldsymbol{\Omega}}_{\boldsymbol{ci}} \times \tilde{\boldsymbol{r}}_{\boldsymbol{wij}} \tag{1}$$

$$\tilde{\boldsymbol{\Omega}}_{\boldsymbol{ci}} = \dot{\psi}_i \, \tilde{\boldsymbol{e}}_{\boldsymbol{z}} \tag{2}$$

where $\tilde{\boldsymbol{V}}_{\boldsymbol{ci}}$ is the velocity of the mass center of the rover, $\tilde{\boldsymbol{\Omega}}_{\boldsymbol{ci}}$ is the angular velocity vector of the rover and $\tilde{\boldsymbol{e}}_{\boldsymbol{z}}$ is a unit vector both along the $z_{ci,}$ - axis while $\tilde{\boldsymbol{r}}_{\boldsymbol{wij}}$ is the position vector from the rover's mass center to the wheel center.

The velocity vector $\tilde{\boldsymbol{V}}_{\boldsymbol{pij}}$, representing the velocity of a point P located at the roller center can be expressed as

$$\tilde{\boldsymbol{V}}_{\boldsymbol{pij}} = \tilde{\boldsymbol{V}}_{mij} + \tilde{\boldsymbol{\omega}}_{ij} \times \tilde{\boldsymbol{\rho}}_{\boldsymbol{ij}} \tag{3}$$

where $\tilde{\boldsymbol{\rho}}_{\boldsymbol{ij}}$ is the position vector from the wheel's center to the point P, the roller center.

If the rollers do not slip, $\tilde{\boldsymbol{V}}_{\boldsymbol{pij}}$ does not have a component in the direction of the axis of roller rotation $\tilde{\boldsymbol{e}}_{\boldsymbol{\beta j}}$, and can be expressed as

$$\tilde{\boldsymbol{V}}_{\boldsymbol{pij}} \cdot \tilde{\boldsymbol{e}}_{\boldsymbol{\beta ij}} = 0 \tag{4}$$

where $\tilde{\boldsymbol{e}}_{\boldsymbol{\beta ij}}$ is a unit vector along the roller's axis of rotation. After carrying out some algebraic manipulations using (3) and (4), one can write the following expressions:

.

$$\tilde{\boldsymbol{V}}_{mij} \cdot \tilde{\boldsymbol{e}}_{\boldsymbol{\beta ij}} + (\tilde{\boldsymbol{\omega}}_{ij} \times \tilde{\boldsymbol{\rho}}_{\boldsymbol{ij}}) \cdot \tilde{\boldsymbol{e}}_{\boldsymbol{\beta ij}} = 0$$

$$(\tilde{\boldsymbol{\omega}}_{ij} \times \tilde{\boldsymbol{\rho}}_{\boldsymbol{ij}}) = -\tilde{\boldsymbol{\omega}}_{ij} s \tilde{\boldsymbol{e}}_{\boldsymbol{xi}}$$

$$\tilde{\boldsymbol{V}}_{mij} \cdot \tilde{\boldsymbol{e}}_{\boldsymbol{\beta ij}} = \tilde{\boldsymbol{\omega}}_{ij} s \, (\tilde{\boldsymbol{e}}_{\boldsymbol{xi}} \cdot \tilde{\boldsymbol{e}}_{\boldsymbol{\beta ij}}) \tag{5}$$

where s is the radius of the wheel and $\tilde{\boldsymbol{e}}_{\boldsymbol{xi}}$ is a unit vector in the direction of the x_{ci} - axis.

Furthermore, rewriting the equations of constraints by utilizing (1) and (5), one can obtain the following relationships:

$$\tilde{\boldsymbol{V}}_{\boldsymbol{ci}} \cdot \tilde{\boldsymbol{e}}_{\boldsymbol{\beta ij}} + (\tilde{\boldsymbol{\Omega}}_{\boldsymbol{ci}} \times \tilde{\boldsymbol{r}}_{\boldsymbol{wij}}) \cdot \tilde{\boldsymbol{e}}_{\boldsymbol{\beta ij}} = \tilde{\boldsymbol{\omega}}_{ij} \, s \, (\tilde{\boldsymbol{e}}_{\boldsymbol{xi}} \cdot \tilde{\boldsymbol{e}}_{\boldsymbol{\beta ij}})$$

$$\tilde{\boldsymbol{V}}_{\boldsymbol{ci}} \cdot \tilde{\boldsymbol{e}}_{\boldsymbol{\beta ij}} + (\tilde{\boldsymbol{r}}_{\boldsymbol{wij}} \times \tilde{\boldsymbol{e}}_{\boldsymbol{\beta ij}}) \cdot \tilde{\boldsymbol{\Omega}}_{\boldsymbol{ci}} = \tilde{\boldsymbol{\omega}}_{ij} \, s \cos(\beta_{ij}) \tag{6}$$

where β_{ij} is defined as the angle between the two unit vectors $\tilde{\boldsymbol{e}}_{\boldsymbol{xi}}$ and $\tilde{\boldsymbol{e}}_{\boldsymbol{\beta ij}}$

$$e_{\beta i1}^T = [cos(\beta_{i1}), -sin(\beta_{i1}), 0]$$
$$e_{\beta i2}^T = [cos(\beta_{i2}), sin(\beta_{i2}), 0]$$
$$e_{\beta i3}^T = [cos(\beta_{i3}), sin(\beta_{i3}), 0]$$
$$e_{\beta i4}^T = [cos(\beta_{i4}), -sin(\beta_{i4}), 0]$$
$$\tilde{r}_{wi1}^T = [a, b, 0]$$
$$\tilde{r}_{wi2}^T = [a, -b, 0]$$
$$\tilde{r}_{wi3}^T = [-a, b, 0]$$
$$\tilde{r}_{wi4}^T = [-a, -b, 0] \tag{7}$$

One can obtain the following expressions by plugging (7) into (6) and substituting 45° for β_{ij}:

$$\tilde{\boldsymbol{V}}_{\boldsymbol{ci}} = \begin{bmatrix} V_{cix} \\ V_{ciy} \\ V_{ciz} \end{bmatrix} = \begin{bmatrix} s(\omega_{i1} + \omega_{i2})/2 \\ s(\omega_{i3} - \omega_{i1})/2 \\ 0 \end{bmatrix}$$

$$\tilde{\boldsymbol{\Omega}}_{\boldsymbol{ci}} = \begin{bmatrix} \Omega_{\mathrm{cix}} \\ \Omega_{\mathrm{ciy}} \\ \Omega_{\mathrm{ciz}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ s(\omega_{i3} - \omega_{i1})/(2(a+b)) \end{bmatrix}$$

$$\omega_{i4} = \omega_{i1} + \omega_{i2} - \omega_{i3} \tag{8}$$

The following rotational matrix represents the rotation between the inertial and the rover body axes:

$$\underline{\Psi}_{zi} = \begin{bmatrix} \cos\psi_i & -\sin\psi_i & 0 \\ \sin\psi_i & \cos\psi_i & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{9}$$

Homogeneous transformation matrix $\underline{\boldsymbol{T}}_{\boldsymbol{f}}^{\boldsymbol{g}}$ which transforms the coordinates between frame-g and frame-f on the robot arm can be obtained by Denavit-Hartenberg (D-H) convention as follows.

$$\underline{\boldsymbol{T}}_{\boldsymbol{f}}^{\boldsymbol{g}} = \underline{\boldsymbol{A}}_{\boldsymbol{f+1}} \, \underline{\boldsymbol{A}}_{\boldsymbol{f+2}} \, \ldots\ldots\ldots \, \underline{\boldsymbol{A}}_{\boldsymbol{g-1}} \, \underline{\boldsymbol{A}}_{\boldsymbol{g}} \qquad \mathbf{f<g}$$

$$\underline{\boldsymbol{A_f}} = \begin{bmatrix} \cos\theta_{fi} & -sin\theta_{fi}cos\alpha_{fi} & sin\theta_{fi}sin\alpha_{fi} & a_{fi}\cos\theta_{fi} \\ sin\theta_{fi} & \cos\theta_{fi}\,cos\alpha_{fi} & -\cos\theta_{fi}\,sin\alpha_{fi} & a_{fi}sin\theta_{fi} \\ 0 & sin\alpha_{fi} & cos\alpha_{fi} & d_{fi} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{10}$$

where $\theta_{fi}, \alpha_{fi}, d_{fi}, a_{fi}$ are the parameters related to joint-f and link-f on the i[th] arm, namely d_{fi} is the offset, a_{fi}, is the f[th] link-length , θ_{fi} is the joint angle and α_{fi} is the twist as defined in DH convention.

The following expressions can be used to obtain the Jacobian matrices and the first-time derivatives of these matrices associated with the rover's center and any arbitrary point-*k* on the arm:

$$\left(\underline{J}_c^k\right)_i = \begin{bmatrix} \widetilde{\boldsymbol{e_z}} & \widetilde{\boldsymbol{e_1}} & \cdots & \widetilde{\boldsymbol{e_7}} \\ \widetilde{\boldsymbol{e_z}} \times \tilde{\boldsymbol{r}}_{\boldsymbol{ck}} & \widetilde{\boldsymbol{e_1}} \times \tilde{\boldsymbol{r}}_{\boldsymbol{1k}} & \cdots & \widetilde{\boldsymbol{e_7}} \times \tilde{\boldsymbol{r}}_{\boldsymbol{7k}} \end{bmatrix} \tag{11}$$

$$\left(\dot{\underline{J}}_c^k\right)_i = \begin{bmatrix} \widetilde{\boldsymbol{e_z}} & \widetilde{\boldsymbol{e_1}} & \cdots & \widetilde{\boldsymbol{e_7}} \\ \widetilde{\boldsymbol{e_z}} \times (\tilde{\Omega}_{ci} \times \tilde{\boldsymbol{r}}_{\boldsymbol{ck}}) & \widetilde{\boldsymbol{e_1}} \times (\dot{\theta}_1 \widetilde{\boldsymbol{e_1}} \times \tilde{\boldsymbol{r}}_{\boldsymbol{1k}}) & \cdots & \widetilde{\boldsymbol{e_7}} \times (\dot{\theta}_7 \widetilde{\boldsymbol{e_7}} \times \tilde{\boldsymbol{r}}_{\boldsymbol{7k}}) \end{bmatrix} \tag{12}$$

Where $\tilde{\boldsymbol{e}}_i$ is the unit vector along the i[th] joint rotation axis, $\widetilde{\boldsymbol{e_z}}$ is the unit vector along $\widetilde{\boldsymbol{\Omega}}_{\boldsymbol{ci}}$, and $\tilde{\boldsymbol{r}}_{\boldsymbol{ik}}$, $\tilde{\boldsymbol{r}}_{\boldsymbol{ck}}$ are the position vectors from i[th] joint and the rover's center to the point k, respectively.

The linear and angular velocities and accelerations of point *k* on the i[th] rover arm can be calculated as follows:

$$\begin{bmatrix} \tilde{\Omega}_k \\ \tilde{V}_k \end{bmatrix}^i = \left(\underline{J}_c^k\right)_i \begin{bmatrix} \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{ci}} \\ \dot{\tilde{\theta}}_i \end{bmatrix} + \begin{bmatrix} 0 \\ \widetilde{\boldsymbol{V}}_{\boldsymbol{ci}} \end{bmatrix} \tag{13}$$

$$\begin{bmatrix} \dot{\tilde{\Omega}}_k \\ \dot{\tilde{V}}_k \end{bmatrix}^i = \left(\underline{J}_c^k\right)_i \begin{bmatrix} \dot{\tilde{\Omega}}_{ci} \\ \ddot{\tilde{\theta}}_i \end{bmatrix} + \left(\dot{\underline{J}}_c^k\right)_i \begin{bmatrix} \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{ci}} \\ \dot{\tilde{\theta}}_i \end{bmatrix} + \begin{bmatrix} 0 \\ \dot{\tilde{V}}_{ci} \end{bmatrix} \tag{14}$$

where $\begin{bmatrix} \tilde{\Omega}_k \\ \tilde{V}_k \end{bmatrix}^i$ is a vector consisting of the angular and linear velocity vectors of the point *k* on the i[th] arm, respectively while $\dot{\tilde{\theta}}_i^T = [\,\dot{\theta}_{i1}, \dot{\theta}_{i2}, \dot{\theta}_{i3,\ldots.} \dot{\theta}_{in}]$ is a vector consists of the i[th] rover-arm joint angular rates.

2.3. Model of Dynamics

The dynamics equations of motions of the system is derived using the Lagrangian formulation. The total kinetic energy T_t consists of two parts, the rotational and translational kinetic energies of the robotics arms and the rovers.

$$T_t = T_{tr} + T_{rt} \tag{15}$$

The angular and translational velocities of the rovers as well as that of the robot links' center of mass can be calculated using (14) and (8). Then, the total kinetic energy of the system can be obtained using (15).

The dynamics equations of motion can be obtained using the following Lagrangian formulation:

$$\frac{d}{dt}\left(\frac{\partial T_t}{\partial \dot{q}_h}\right) - \frac{\partial T_t}{\partial q_h} = Q_h, \qquad h = 1, \ldots 2m \tag{16}$$

where q_h and Q_h are the generalized coordinates and forces, respectively and

$$q^T = [\phi_{11}, \phi_{12}, \phi_{13}, \phi_{21}, \phi_{22}, \phi_{23}, \theta_{11} \ldots \theta_{1n}, \theta_{21} \ldots \theta_{2n}]$$

and *m=(n+3)*, *n* represents the total number of degrees of freedom of the robotics arms.

Applying (16), the dynamics equations of motions for both rovers and the arms can be written in the following form:

$$\begin{bmatrix} \boldsymbol{G_{W_L}} & \boldsymbol{G_{W_{LR}}} & \boldsymbol{G_{W\theta_L}} & \boldsymbol{G_{W\theta_R}} \\ \boldsymbol{G^T_{W_{LR}}} & \boldsymbol{G_{W_R}} & \boldsymbol{G_{W\theta_L}} & \boldsymbol{G_{W\theta_R}} \\ \boldsymbol{G^T_{W\theta_L}} & \boldsymbol{G^T_{W\theta_L}} & \boldsymbol{G_{\theta_L}} & \boldsymbol{G_{\theta_{LR}}} \\ \boldsymbol{G^T_{W\theta_R}} & \boldsymbol{G^T_{W\theta_R}} & \boldsymbol{G^T_{\theta_{LR}}} & \boldsymbol{G_{\theta_R}} \end{bmatrix} \begin{bmatrix} \ddot{\tilde{\Phi}}_L \\ \ddot{\tilde{\Phi}}_R \\ \ddot{\tilde{\theta}}_L \\ \ddot{\tilde{\theta}}_R \end{bmatrix} + \begin{bmatrix} \tilde{\boldsymbol{c}}_L \\ \tilde{\boldsymbol{c}}_R \\ \tilde{\boldsymbol{c}}_{\boldsymbol{\theta_L}} \\ \tilde{\boldsymbol{c}}_{\boldsymbol{\theta_R}} \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{M}}_L \\ \widetilde{\boldsymbol{M}}_R \\ \tilde{\boldsymbol{\tau}}_{\boldsymbol{\theta_L}} \\ \tilde{\boldsymbol{\tau}}_{\boldsymbol{\theta_R}} \end{bmatrix} \tag{17}$$

where $\underline{G}$ is the mass / inertia matrix (a positive definite matrix) and, $\ddot{\tilde{\boldsymbol{\Phi}}}_{\boldsymbol{L}}$, , $\ddot{\tilde{\boldsymbol{\Phi}}}_{\boldsymbol{R}}$ are the wheels' angular accelerations for the two rovers *i=L* and *R*, and $\ddot{\tilde{\boldsymbol{\theta}}}_{\boldsymbol{L}}$, $\ddot{\tilde{\boldsymbol{\theta}}}_{\boldsymbol{R}}$ are the joint rotational accelerations for the two manipulators, *i=L* and *R*, respectively. The indices L and R are referred to the first and second rover and robotics arm, respectively.

The non-linear Coriolis and centrifugal terms are represented by $\tilde{\boldsymbol{c}}_{\boldsymbol{L_L}}$, $\tilde{\boldsymbol{c}}_{\boldsymbol{R}}$, $\tilde{\boldsymbol{c}}_{\boldsymbol{\theta_L}}$, and $\tilde{\boldsymbol{c}}_{\boldsymbol{\theta_R}}$. and $\tilde{\boldsymbol{\tau}}_{\boldsymbol{\theta_L}}$ $\tilde{\boldsymbol{\tau}}_{\boldsymbol{\theta_R}}$are the joint control torques for the two robot manipulators. Finally, $\widetilde{\boldsymbol{M}}_{\boldsymbol{L}}, \widetilde{\boldsymbol{M}}_{\boldsymbol{R}}$are the wheel control moments for the two rovers.

$\dot{\tilde{\Phi}}_i^T = [\,\omega_{i1}, \omega_{i2}, \omega_{i3}\,]$ includes the wheels angular rates of the i[th] rover, *i=L* and *R* for two rovers. If there is no slip, the fourth wheel angular rate can be calculated using (8).

2.4. Optimal Control Allocation (OCA) Technique

The robotics system composed of two rovers and two redundant arms is an undetermined because of the excessive number of sensors and actuators used to control the motions of the links and rovers.

A novel two-stage optimal control technique is derived in this section and the control system block diagram is provided in Figure 2a.

The first stage of the diagram generates the reference trajectories for the end-effectors corresponding to a given payload trajectory. The Impedance control equations representing this first stage are provided in (18). These equations are developed in [2].

$$\ddot{\tilde{X}}_i = \underline{M_i}^{-1}\underline{B}\left\{\dot{\tilde{X}}_{di} - \dot{\tilde{X}}_i\right\} + \underline{M_i}^{-1}\underline{K}\{\tilde{X}_{di} - \tilde{X}_i\}, \qquad i = L, R$$

(18)

where, $\underline{M_i}, \underline{K}\ \underline{B}$, are 6x6 positive definite matrices and are chosen in accordance with the tracking performance requirements. $\tilde{X}_i$. (i varies between L and R for each arm) are the end-effector trajectories while $\tilde{X}_{di}$.correspond to the reference trajectories of the attachment points on the common load.

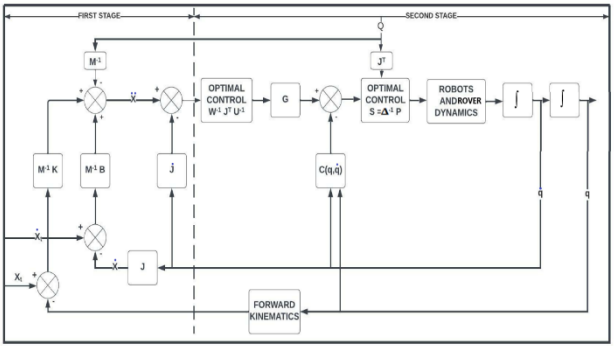


Figure 2: a. Optimal control system block diagram – Two Stage Control

The expressions for $\tilde{X}_i$ and $\dot{\tilde{X}}_i$ can be written as follows:

$$\tilde{X}_i = \begin{bmatrix} \tilde{\Omega}_k \\ \tilde{V}_k \end{bmatrix}^i \tag{19}$$

$$\dot{\tilde{X}}_i = \begin{bmatrix} \dot{\tilde{\Omega}}_k \\ \dot{\tilde{V}}_k \end{bmatrix}^i \tag{20}$$

where k point is the end-effector for the i[th] arm. Performing the least-square minimization of joint rates, the inverse kinematics of the robotics system can be solved as the following:

$$\begin{bmatrix} \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{ci}} \\ \dot{\tilde{\theta}}_i \end{bmatrix} = \underline{W_i}^{-1} \left(\underline{J_c^k}\right)_i^T \ \ \underline{U_i}^{-1} \dot{\tilde{X}}_i \tag{21}$$

$$\begin{bmatrix} \dot{\tilde{\Omega}}_{ci} \\ \ddot{\tilde{\theta}}_i \end{bmatrix} = \underline{W_i}^{-1} \left(\underline{J_c^k}\right)_i^T \ \ \underline{U_i}^{-1} \left\{ \ddot{\tilde{X}}_i - \left(\dot{\underline{J}}_c^k\right)_i \begin{bmatrix} \widetilde{\boldsymbol{\Omega}}_{\boldsymbol{ci}} \\ \dot{\tilde{\theta}}_i \end{bmatrix} \right\} \tag{22}$$

$$\underline{U_i} = \left(\underline{J_c^k}\right)_i \underline{W_i}^{-1} \left(\underline{J_c^k}\right)_i^T \tag{23}$$

where $\underline{W_i}$ is a square positive definite weighting matrix with the dimensions of *(n+3)* by *(n+3)*.

The second stage in the block diagram is predicated on optimal control allocation (OCA). The mathematical model is provided below.

The performance index (a cost function) C is formulated to minimize the wheel moments $\widetilde{\boldsymbol{M}}_{\boldsymbol{L}}, \widetilde{\boldsymbol{M}}_{\boldsymbol{R}}$ and the joint torques $\tilde{\boldsymbol{\tau}}_{\boldsymbol{\theta_L}}\ \tilde{\boldsymbol{\tau}}_{\boldsymbol{\theta_R}}$, and the contact moments and forces $\widetilde{\boldsymbol{N}}_{\boldsymbol{i}}$ and $\widetilde{\boldsymbol{F}}_{\boldsymbol{i}}$ applied by the end-effectors on the common load

The performance index C can be expressed as:

$$C = \frac{1}{2}\ \tilde{S}^T\ \underline{H}\tilde{S} + \tilde{\lambda}^T\ \tilde{E} \tag{24}$$

$\underline{H}$ is a *(2n+18, 2n+18)* positive definite weighting matrix, $\tilde{\lambda}$ is the *((2n+12), 1)* Lagrangian multiplier and $\tilde{E}$ vector includes the equations of constraints and can be calculated as shown in (26).

The $\tilde{S}$ vector contains the contact forces / moments, the wheel moments as well as the joint torques for the two arms and rovers as described below:

$$\tilde{S}^T = \left[\tilde{Q}_L, \tilde{Q}_R, \widetilde{M}_L, \widetilde{M}_R\ , \tilde{\tau}_{\theta_L}\ \tilde{\tau}_{\theta_R}\ \right]$$

$$\tilde{Q}_i = \begin{bmatrix} \widetilde{N}_i \\ \tilde{F}_i \end{bmatrix} \tag{25}$$

The $\tilde{E}$ vector is provided below:

$$\tilde{E} = \begin{bmatrix} m_t\ddot{x}_t - \tilde{F}_L - \tilde{F}_R \\ \left[\underline{I_t}\tilde{\Omega}_t + \tilde{\Omega}_t \times \underline{I_t}\tilde{\Omega}_t\right] - \left[\ \widetilde{N}_L + \widetilde{N}_R - \tilde{d}_L\ \times\ \tilde{F}_L - \tilde{d}_R\ \times\ \tilde{F}_R\right] \\ \begin{bmatrix} G_{W_L} & G_{W_{LR}} & G_{W\theta_L} & G_{W\theta_R} \\ G_{W_{LR}}^T & G_{W_R} & G_{W\theta_L} & G_{W\theta_R} \\ G_{W\theta_L}^T & G_{W\theta_L}^T & G_{\theta_L} & G_{\theta_{LR}} \\ G_{W\theta_R}^T & G_{W\theta_R}^T & G_{\theta_{LR}}^T & G_{\theta_R} \end{bmatrix} \begin{bmatrix} \ddot{\widetilde{\Phi}}_L \\ \ddot{\widetilde{\Phi}}_R \\ \ddot{\tilde{\theta}}_L \\ \ddot{\tilde{\theta}}_R \end{bmatrix} - \begin{bmatrix} \tilde{c}_L \\ \tilde{c}_R \\ \tilde{c}_{\theta_L} \\ \tilde{c}_{\theta_R} \end{bmatrix} - \begin{bmatrix} \widetilde{M}_L \\ \widetilde{M}_R \\ \tilde{\tau}_{\theta_L} \\ \tilde{\tau}_{\theta_R} \end{bmatrix} \end{bmatrix} \tag{26}$$

where $\ddot{\tilde{x}}_t\ , m_t$ are the translational acceleration and the mass of the common load, respectively and. $\tilde{d}_i^T = (\ x_i,\ \ y_i,\ \ z_i\ )$ is the position vector measured from the i[th] arm's contact point to the load's mass center, while $\tilde{\Omega}_t$ and $\underline{I_t}$ are the angular rate and the inertia matrix of the common load around its center of mass.

Once can minimize the performance index C by taking the derivative of C with respect to $\tilde{\lambda}_i$ and $\tilde{S}$ to obtain the minimum norm of wheel moments, joint torques, as well as the contact moments /force exerted by the end-effectors on the common load.

$$\frac{\partial C}{\partial \tilde{S}} = \tilde{0} \tag{27}$$

and

$$\frac{\partial C}{\partial \tilde{\lambda}} = \tilde{0} \tag{28}$$

One can obtain the minimum norm of $\tilde{S}$ containing the wheel moments, joint torques, as well as the contact force and moments by making use of the equations (27) and (28).

$$\tilde{S} = \underline{\Delta}^{-1}\,\tilde{P} \tag{29}$$

where $\underline{\Delta}$ is a *((2n + 18), (2n + 18))* square matrix and $\underline{\Delta}$ and $\tilde{P}$ are presented as follows:

$$\underline{\Delta} = \begin{bmatrix} \underline{H_{NL}} & \underline{0} & -\underline{H_{NR}} & \underline{0} & -\underline{H_{TL}}\left(\underline{J_c^k}\right)_L^T & \underline{W_{TR}}\left(\underline{J_c^k}\right)_R^T \\ (\underline{D_L} - \underline{D_R})\underline{H_{NL}} & \underline{H_{FL}} & \underline{0} & -\underline{H_{FR}} & (\underline{D_R} - \underline{D_L} - \underline{1})\left(\underline{J_c^k}\right)_L^T & \underline{W_{TR}}\left(\underline{J_c^k}\right)_R^T \\ \underline{0} & \underline{1} & \underline{0} & \underline{1} & \underline{0} & \underline{0} \\ \underline{1} & -\underline{D_L} & 1 & \underline{D_R} & \underline{0} & \underline{0} \\ \left(\underline{J_c^k}\right)_{L1}^T & \left(\underline{J_c^k}\right)_{L2}^T & \underline{0} & \underline{0} & \underline{1} & \underline{0} \\ \underline{0} & \underline{0} & \left(\underline{J_c^k}\right)_{R1}^T & \left(\underline{J_c^k}\right)_{R2}^T & \underline{0} & \underline{1} \end{bmatrix} \tag{30}$$

$$\tilde{P} = \begin{bmatrix} \tilde{0} \\ \tilde{0} \\ \tilde{0} \\ \tilde{0} \\ m_t \ddot{\tilde{x}}_t \\ \left[\underline{I_t}\dot{\tilde{\Omega}}_t + \tilde{\Omega}_t \times \underline{I_t}\tilde{\Omega}_t\right] \\ \begin{bmatrix} G_{W_L} & G_{W_{LR}} & G_{W\theta_L} & G_{W\theta_R} \\ G_{W_{LR}}^T & G_{W_R} & G_{W\theta_L} & G_{W\theta_R} \\ G_{W\theta_L}^T & G_{W\theta_L}^T & G_{\theta_L} & G_{\theta_{LR}} \\ G_{W\theta_R}^T & G_{W\theta_R}^T & G_{\theta_{LR}}^T & G_{\theta_R} \end{bmatrix} \begin{bmatrix} \ddot{\tilde{\Phi}}_L \\ \ddot{\tilde{\Phi}}_R \\ \ddot{\tilde{\theta}}_L \\ \ddot{\tilde{\theta}}_R \end{bmatrix} + \begin{bmatrix} \tilde{c}_L \\ \tilde{c}_R \\ \tilde{c}_{\theta_L} \\ \tilde{c}_{\theta_R} \end{bmatrix} \end{bmatrix} \tag{31}$$

$$\underline{D_i} = \begin{bmatrix} 0 & -z_i & y_i \\ z_i & 0 & -x_i \\ -y_i & x_i & 0 \end{bmatrix} \tag{32}$$

2.5. Non-linear Model Predictive Control (NMPC) Technique

The control block diagram of the NMPC is illustrated in Figure.2b. It replaces the second stage of the model in Figure.2a. The reference trajectory shown in this diagram is the output of the first stage, i.e., the impedance control trajectory generation. However, in this case, the future state estimates are also taken into account to estimate the future reference trajectory values.

A robust NMPC algorithm is implemented by optimizing a performance index of the system which considers the predictions of the output signal and the constraints on the states, outputs and inputs as illustrated in Figure 2b.

The main difference between the Optimal Control Allocation (OCA) and the Non-linear Model Predictive Control (NMPC) is that the latter utilizes a model to predict and control future behavior, while the former only takes into account the current and the past.

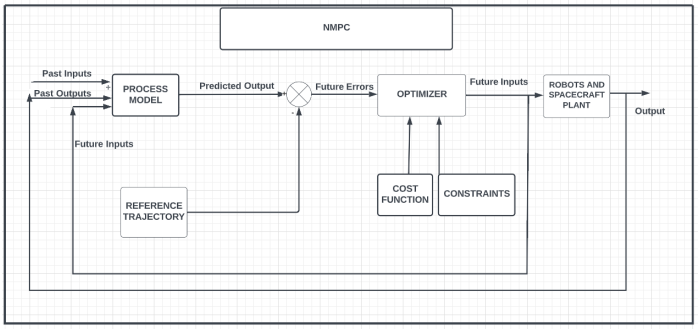


Figure 2: b Nonlinear Model Predictive Control (NMPC) block diagram

The optimization process carried out through the NMPC algorithm is performed at each control interval to predict the system's future behavior. It involves implementing various optimization problems related to the cost functions and constraints. The cost function is a type of scalar which needs to be minimized at intervals to assess the system's performance.

Besides the cost functions, the system also has to perform under various constraints to check its performance. These include the plant output and states. The modified states are adjusted depending on the constraints that are applied to the system.

The conventional MPC formulation for the multi-rover nonlinear system can be written as:

$$\boldsymbol{C} = \int_0^{T_p} \Big[\left(\tilde{\boldsymbol{y}}(\boldsymbol{t}) - \tilde{\boldsymbol{y}}_{\boldsymbol{r}}(\boldsymbol{t})\right)^{\boldsymbol{T}} \underline{\boldsymbol{K}} \left(\tilde{\boldsymbol{y}}(\boldsymbol{t}) - \tilde{\boldsymbol{y}}_{\boldsymbol{r}}(\boldsymbol{t})\right) + \tilde{\boldsymbol{S}}^T(t)\,\underline{\boldsymbol{H}}\,\tilde{\boldsymbol{S}}\,(\mathrm{t}) \Big]\, \boldsymbol{dt}$$

subject to:

$$\begin{aligned} \dot{\tilde{\boldsymbol{y}}} &= \tilde{\boldsymbol{g}}(\tilde{\boldsymbol{y}}) + \underline{\boldsymbol{L}}\,\tilde{\boldsymbol{S}} \\ \tilde{\boldsymbol{z}} &= \tilde{\boldsymbol{g}}_{\boldsymbol{z}}(\tilde{\boldsymbol{y}}) + \underline{\boldsymbol{H}}\,\tilde{\boldsymbol{S}} \\ \tilde{\boldsymbol{y}}(0) &= \tilde{\boldsymbol{y}}(t_0) \end{aligned} \tag{33}$$

where T_p is the prediction horizon; $\underline{\boldsymbol{K}}$ and $\underline{\boldsymbol{H}}$ are *(2(2n+6) x 2(2n+6))* and *((2n+18) x (2n+18))* positive definite square weighting matrices, respectively; $\tilde{\boldsymbol{g}}(\tilde{\boldsymbol{y}}), \tilde{\boldsymbol{g}}_{\boldsymbol{z}}(\tilde{\boldsymbol{y}}), \underline{\boldsymbol{L}}, \underline{\boldsymbol{H}}$ are part of the nonlinear system equations.

Also, $\tilde{\boldsymbol{y}}(\mathrm{t})^{\mathbf{T}} = [q^T, \dot{q}^T]$, is a (1 x 2(6+2n)) state vector, q^T vector is previously defined in Eq.(16), and $\tilde{\boldsymbol{y}}_{\boldsymbol{r}}(\mathrm{t})$ is the reference / desired states.

The non-linear system can now be described with an exact quasi linear parameter varying realization:

$$\begin{aligned} \tilde{\boldsymbol{y}}(k_t + 1) &= \underline{\hat{\boldsymbol{A}}}\big(\hat{\boldsymbol{g}}(k_t)\big)\tilde{\boldsymbol{y}}(k_t) + \underline{\hat{\boldsymbol{B}}}\big(\hat{\boldsymbol{g}}(k_t)\big)\,\tilde{\boldsymbol{S}}(k_t) \\ \tilde{\boldsymbol{z}}(k_t) &= \underline{\hat{\boldsymbol{C}}}\big(\hat{\boldsymbol{g}}(k_t)\big)\tilde{\boldsymbol{y}}(k_t) + \underline{\hat{\boldsymbol{D}}}\big(\hat{\boldsymbol{g}}(k_t)\big)\,\tilde{\boldsymbol{S}}(k_t) \\ \hat{\boldsymbol{g}}(k_t) &= \boldsymbol{f}_{\boldsymbol{g}}\big(\tilde{\boldsymbol{y}}(k_t)\big) \end{aligned} \tag{34}$$

where k_t is the sampling instant and $\tilde{\boldsymbol{z}}(k_t)$ is a vector of the measured outputs at instant k_t.

The NMPC is employed at each sampling instantk_t.and the discrete states $\tilde{\boldsymbol{y}}(k_t.)$ and control inputs $\tilde{\boldsymbol{S}}(k_t.)$ are obtained

minimizing the following performance index i.e., the Cost Function:

$$C = \frac{1}{2}\sum_{j=1}^{N_P}\Big[\big(\tilde{\boldsymbol{y}}(k_t.+j)-\tilde{\boldsymbol{y}}_r(k_t.+j)\big)^T\,\underline{K}\,\big(\tilde{\boldsymbol{y}}(k_t.+j)-\tilde{\boldsymbol{y}}_r(k_t.+j)\big) + \tilde{\boldsymbol{S}}(k_t.+j-1)^T\,\underline{H}\,\tilde{\boldsymbol{S}}(k_t.+j-1)\Big]$$

subject to

$$\tilde{\boldsymbol{y}}(k_t+j+1) = \underline{\hat{\boldsymbol{A}}}\big(\hat{\boldsymbol{g}}(k_t+j)\big)\tilde{\boldsymbol{y}}(k_t+j) + \underline{\hat{\boldsymbol{B}}}\big(\hat{\boldsymbol{g}}(k_t+j)\big)\,\tilde{\boldsymbol{S}}(k_t+j)$$

$$\tilde{\boldsymbol{z}}(k_t+j) = \underline{\hat{\boldsymbol{C}}}\big(\hat{\boldsymbol{g}}(k_t+j)\big)\tilde{\boldsymbol{y}}(k_t+j) + \underline{\hat{\boldsymbol{D}}}\big(\hat{\boldsymbol{g}}(k_t+j)\big)\,\tilde{\boldsymbol{S}}(k_t+j) \quad (35)$$

3. Computer Simulation Results and Discussion

The results of the computer simulations and their discussions are presented in this section. First, a prescribed trajectory for the common payload's center mass is generated. Then, the desired (reference) trajectories for the two end-effectors are obtained using a method known as the impedance control technique (shown as the first stage in the block diagram).

The goal of the simulation is to obtain the minimum joint and rover wheel torques and contact forces while simultaneously tracking the desired end-effector pose using the developed two different control algorithms (i) OCA and (ii) NMPC.

The parameters for the robotic arms and the rovers employed in the computer simulations are presented in Table 1. A mini version of the SSRMS is utilized.

Table 1: The System Parameters Utilized in the Computer Simulation

Description of Hardware Configuration Items	Dimensions (m)	Mass (kg)
Rovers-(#1 and #2)	(0.5x0.5x0.3)	40
Common Load	(0.4x1x0.4)	10
Link #1	(0.1x0.1x0.1)	1
Link #2	(0.1x0.1x0.1)	1
Link #3	(1x0.1x0.1)	3
Link #4	(1x0.1x0.1)	5
Link #5	(0.1x0.1x0.1)	3
Link #6	(0.1x0.1x0.1)	1
Link #7	(1x0.1x0.1)	3

The desired trajectories for the rotational and translational motions of the common load are presented with time in Figure 3.

The minimum norm of the contact moments /forces, the joint torques, as well as the control forces and moments on Rovers 1 and 2 are plotted in Figure 4a-m using OCA and NMPC algorithms. The non-optimum joint torques (in blue), the joint torques realized by application of OCA algorithm (in red) and by NMPC algorithm (in yellow) are plotted for comparison purposes. The comparison of the plots illustrates that the NMPC is superior and then followed by OCA.

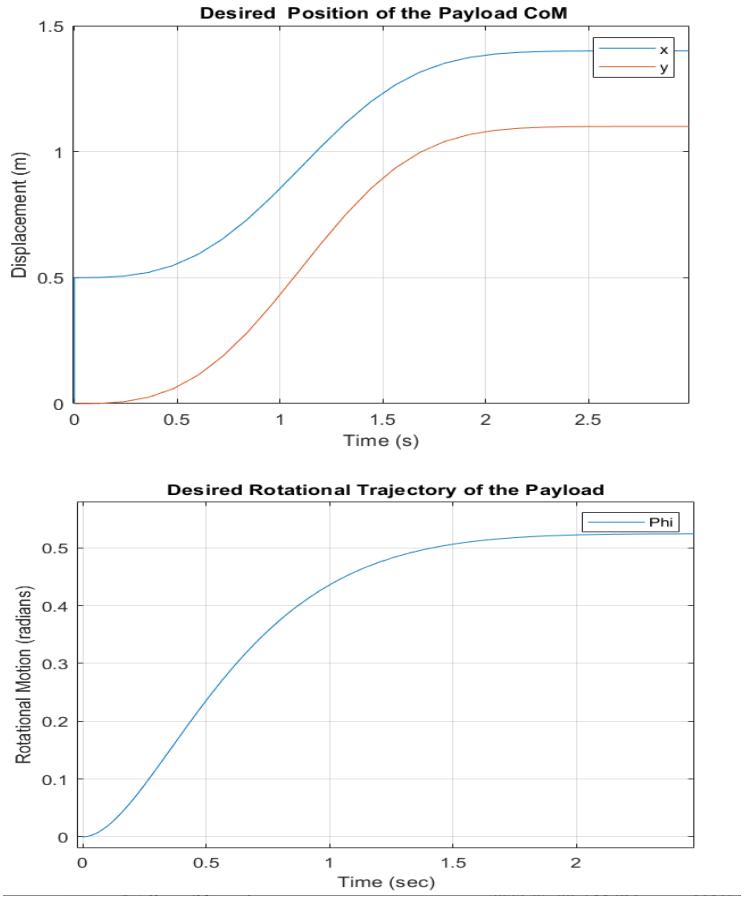


Figure 3: Variation of the reference trajectory for the common load

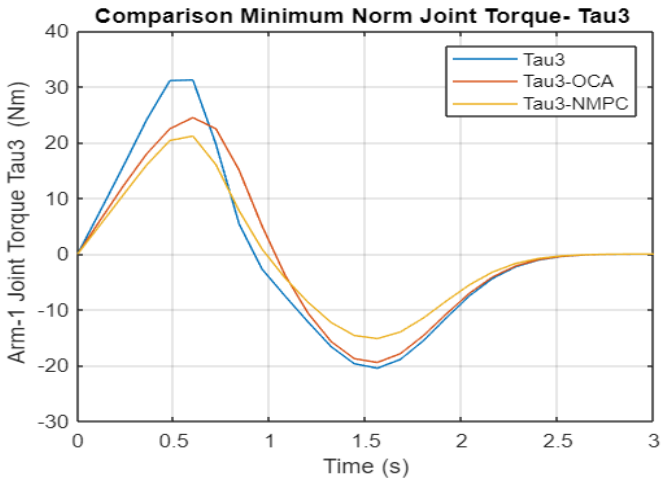


Figure 4: a Variation of the Joint 3 Torque obtained by Non-optimal, OCA, NMPC Algorithms (Arm 1)

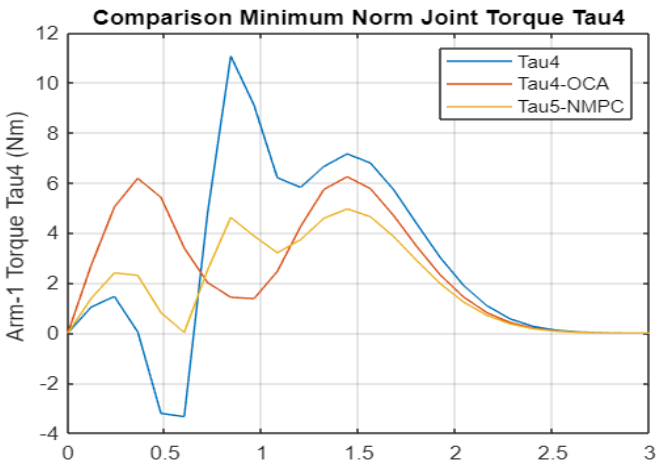


Figure 4: b-Variation of the Joint 4 Torque obtained by Non-optimal, OCA, NMPC Algorithms (Arm 1)

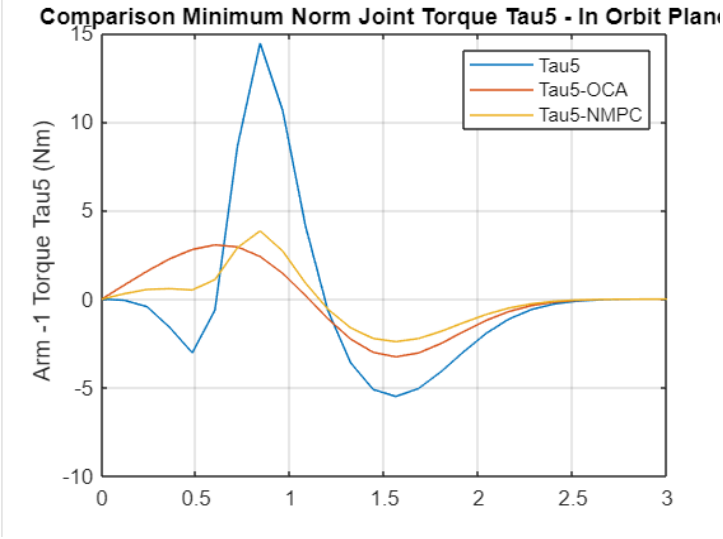


Figure 4: c Variation of the Joint 5 Torque obtained by Non-optimal, OCA, NMPC Algorithms (Arm 1)

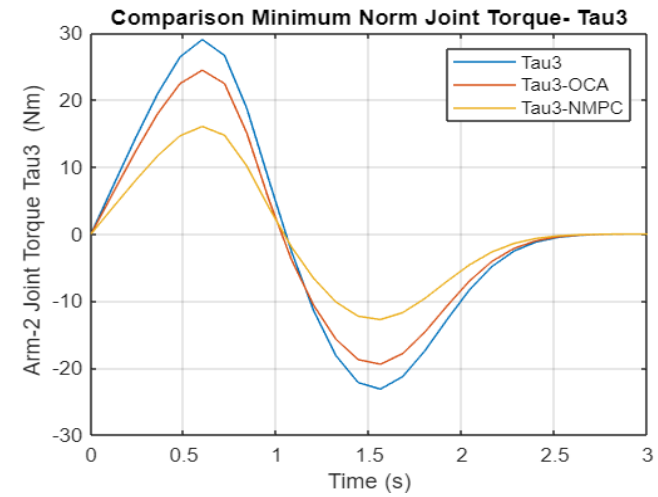


Figure 4: d Variation of the Joint 3 Torque obtained by Non-optimal, OCA, NMPC Algorithms (Arm 2)

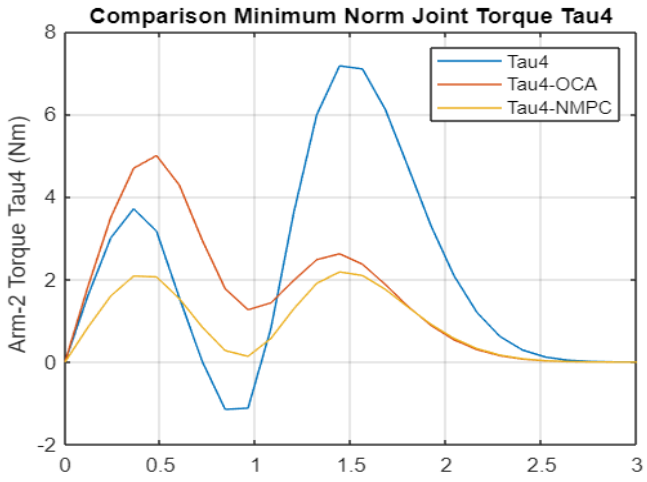


Figure 4: e Variation of the Joint 4 Torque obtained by Non-optimal, OCA, NMPC Algorithms (Arm 2)

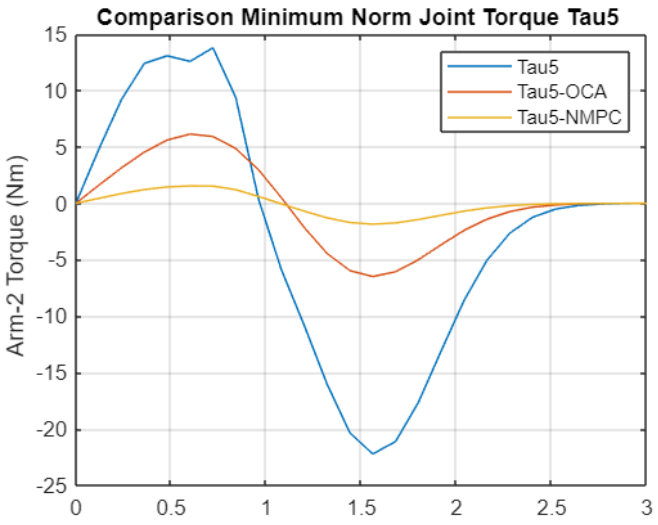


Figure 4: f Variation of the Joint 5 Torque obtained by Non-optimal, OCA, NMPC Algorithms (Arm 2)

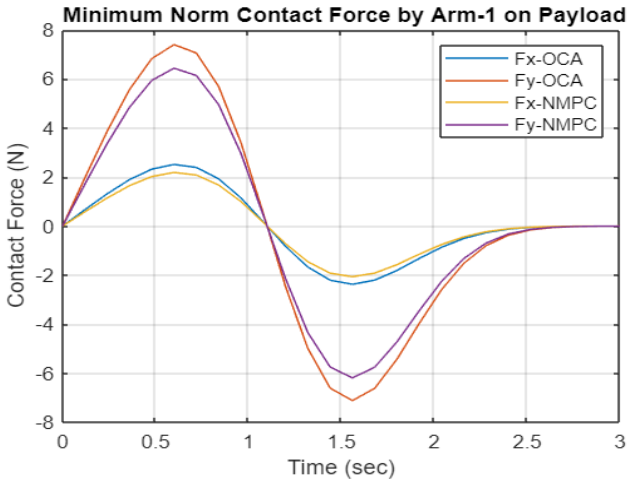


Figure 4: g Variation of the Contact Forces on Payload by Non-optimal, OCA, NMPC Algorithms (Arm 1)

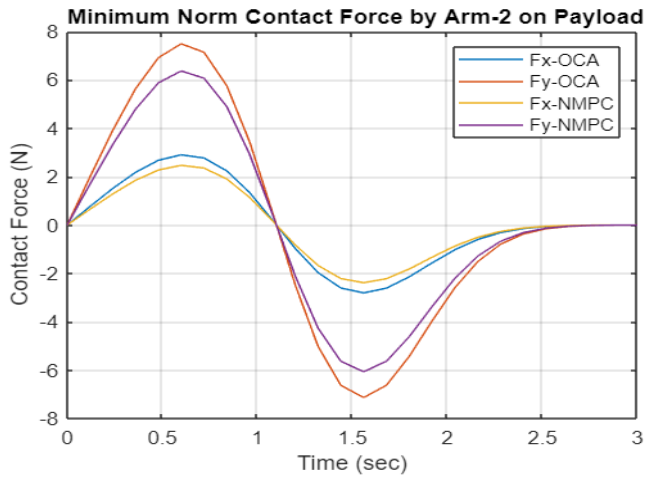


Figure 4: h Variation of the Contact Forces on Payload by Non-optimal, OCA, NMPC Algorithms (Arm 2)

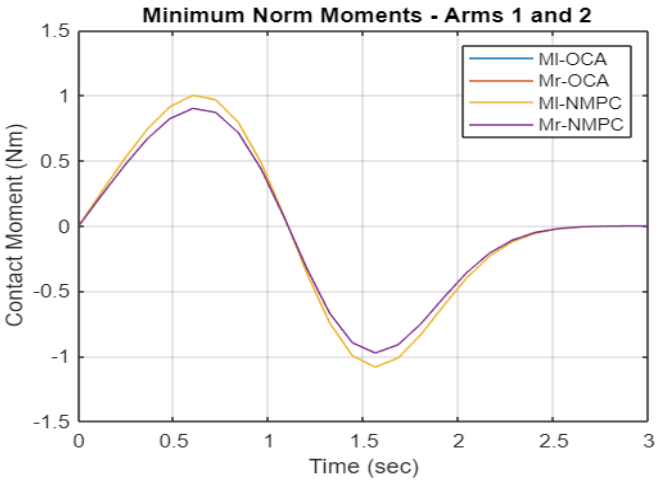


Figure 4: i Variation of the Contact Moments on Payload by Non-optimal, OCA, NMPC Algorithms (Arms 1 and 2)

Again, the NMPC is superior to OCA in obtaining minimum contact moments / forces applied to the common load while the two end-effectors are carrying a common load.

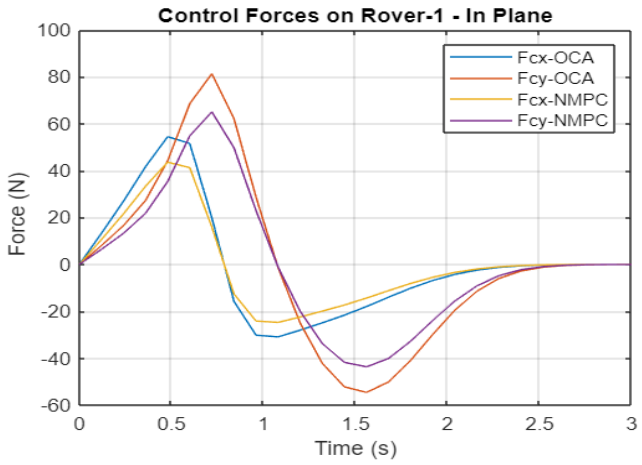


Figure 4j Variation of the Control Forces on Rover 1 by Non-optimal, OCA, NMPC Algorithms (Rover 1)

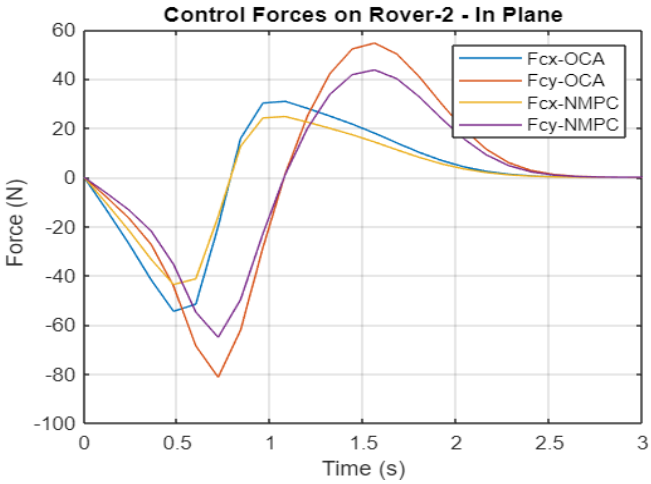


Figure 4: k Variation of the Control Forces on Rover 2 by Non-optimal, OCA, NMPC Algorithms (Rover 2)

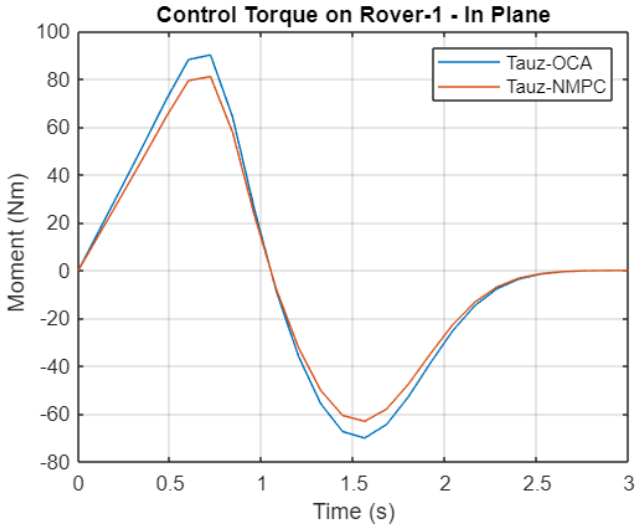


Figure 4: l Variation of the Control Moment on Rover 1 by Non-optimal, OCA, NMPC Algorithms (Rover 1)

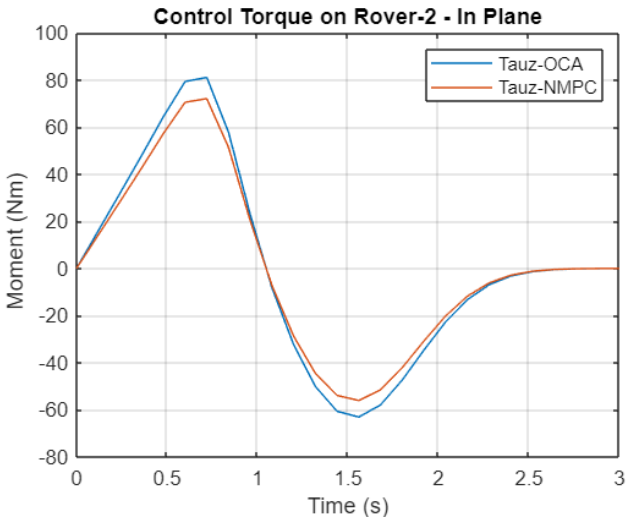


Figure 4: m Variation of the Control Moment on Rover 2 by Non-optimal, OCA, NMPC Algorithms (Rover 2)

A comparative analysis shows that again NMCP is superior to OCA in obtaining minimum norm of control moments and forces for Rovers 1 and 2.

The time variations of joint angular accelerations are shown in Figure 5.

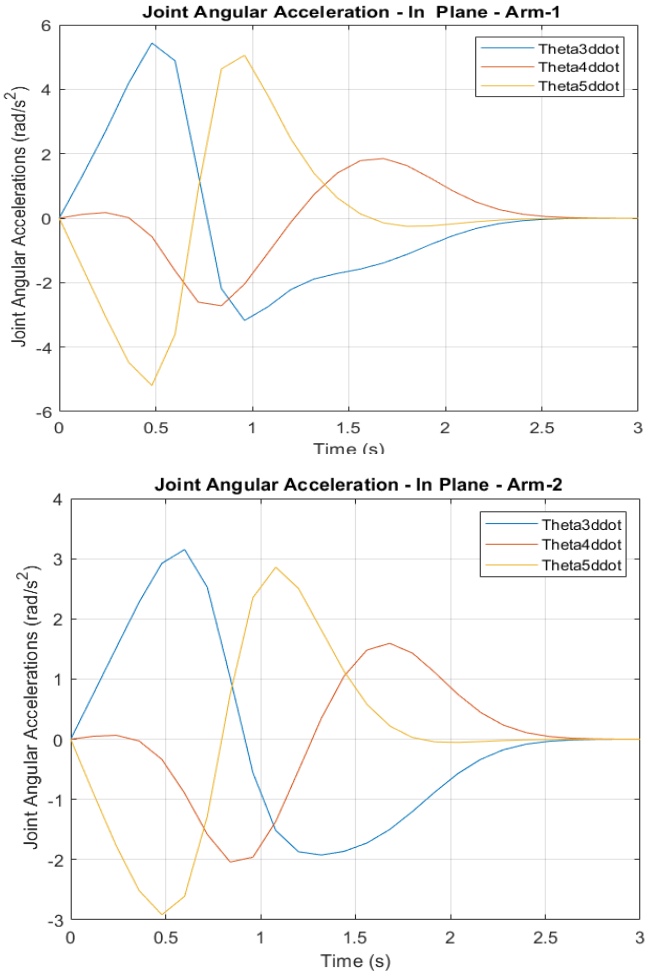
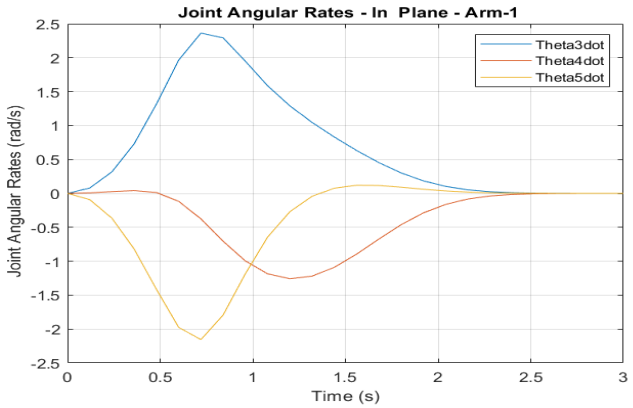


Figure 5: Variation of joint angular accelerations for the first and second arm

The joint accelerations are integrated to calculate rotational rates and angles using (13) and are presented in Figure 6.



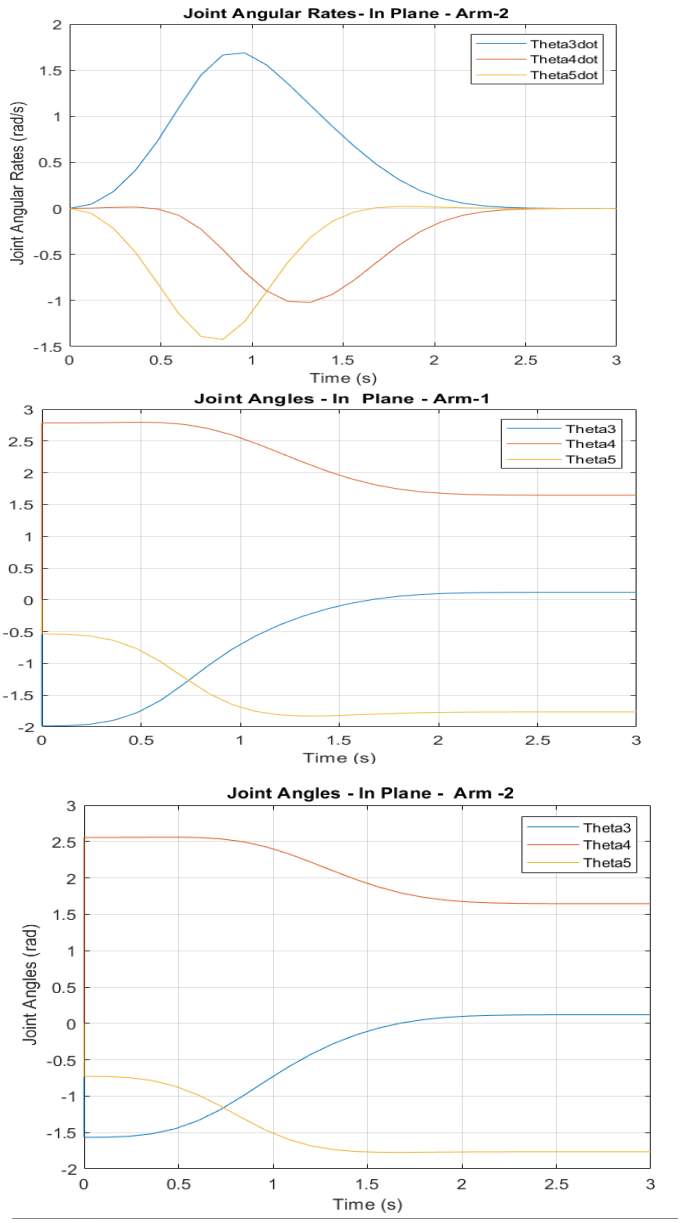
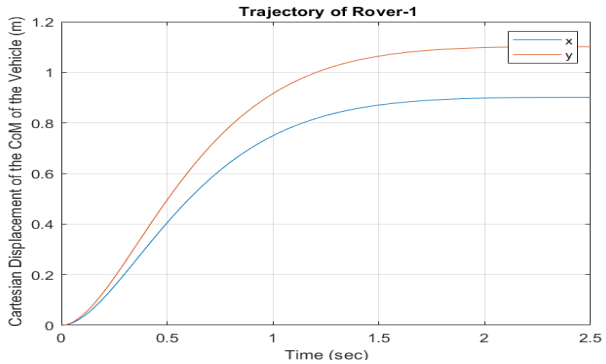


Figure 6: Variation of joint angular rates and angles for the first and second arm

The trajectories of the point *C*, the center of mass of the two rovers are determined by (22) and (8) and are shown in Figure 7.



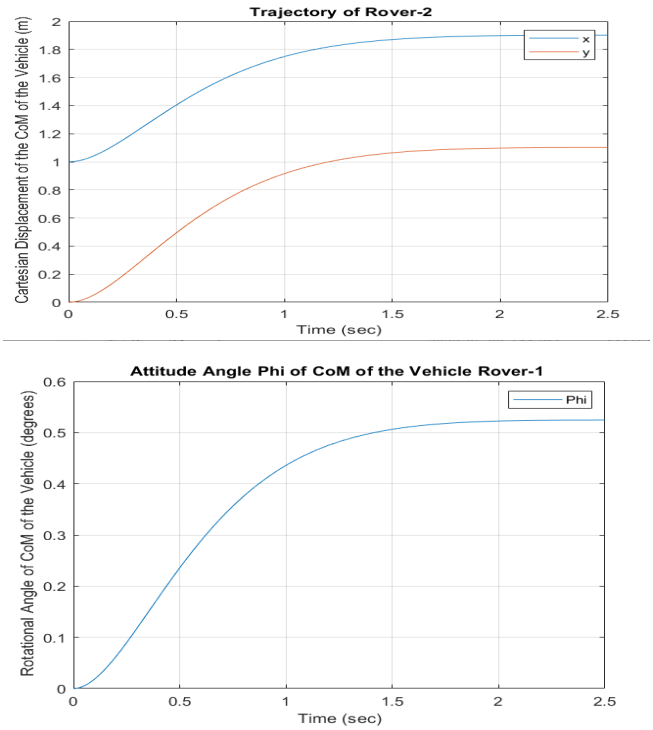


Figure 7: Variation of Rover 1 and 2 positions and orientations with time

The wheel angles of the two rovers are calculated utilizing (22) and are presented in Figure 8.

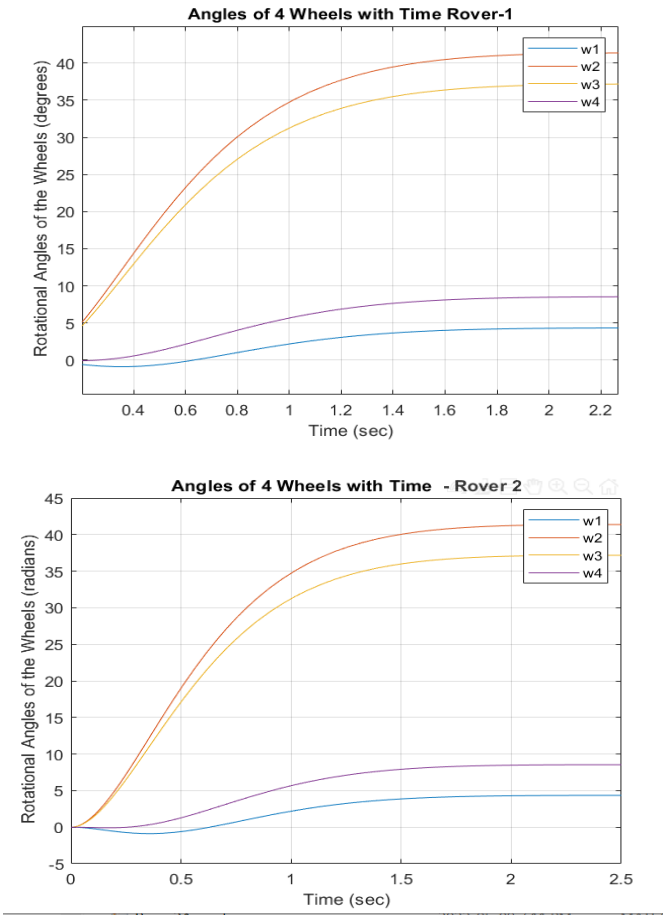


Figure 8 Variation of angles of rotations for rover wheels - rovers 1 and 2

Conclusions and Future Work

The paper presented two novel control algorithms for motion and force control of a multi-rover robotics system when the two end-effectors carrying a common load. One algorithm is predicated on Optimal Control Allocation (OCA) and the other is a discretized (ii) Nonlinear Model Predictive Control (NMPC) algorithm.

The paper focused on developing robust and computationally efficient real-time control algorithms that can minimize the performance index consisting of the norm of the rovers control moments / forces, the joint torques, , as well as the contact moments / forces applied to the common load by two end-effectors.

The norm of wheel moments, joint torques, and the contact moments and forces were minimized to resolve the torque / moment saturation problem often seen while carrying a common load. The paper also presented a minimum norm solution for an underdetermined system subject to non-holonomic constraints Moreover, the developed control algorithm also provided a real-time capability of trajectory for both the rovers and the arms while carrying a common load.

The system consisting of multi-rover with a dual arm was highly non-linear. The linear MPC technique on which most of the previous studies relied was not adequate. On the other hand, the computational complexity of a generic NMPC algorithm was very demanding. Therefore, in this paper, an elegant discretized technique with exact realization was implemented to take into account the full non-linear model and yet provide a simple real-time solution satisfying a minimum performance index subject to constraints.

The results of the computer simulations illustrated that the two algorithms OCA and NMPC worked efficiently. They were able to realize the minimum contact forces and moments and rover wheel moments and forces, joint torques, while manipulating a common load and tracking a reference load trajectory. In addition, the minimal norm solution also satisfied the non-holonomic constraints.

The results revealed that the optimization scheme used by the NMPC algorithm was the most effective when it came to achieving the lowest joint torques and forces. It was then followed by the OCA algorithm and the conventional least square method, respectively.

The authors are currently working on a research project to build a testbed to experimentally validate the computer simulation results. The comparisons of experimental and simulation results will be part of the future research work. Furthermore, the authors assumed no slippage occurred. However, the maximum driving force of each wheel is limited by the dynamic friction coefficient and the magnitude of the normal force acting on it. If this is exceeded, this assumption will no longer be valid. The normal forces will be incorporated in the dynamics model for the future work.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. Kalaycioglu, A. de Ruiter, "Coordinated Motion and Force Control of Multi-Rover Robotics System with Mecanum Wheels," in 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), IEEE: 1–9, 2022, doi:10.1109/IEMTRONICS55184.2022.9795804.

[2] D.S. Necsulescu, B. Kim, S. Kalaycioglu, FREE MOTION, COLLISION AVOIDANCE AND CONTACT MOTION CONTROL FOR MOBILE ROBOTS, Elsevier: 223–228, 1993, doi:10.1016/B978-0-08-041897-1.50042-0.

[3] N. Necsulescu, B. Kim, S. Kalaycioglu, "Contact motion control for mobile robots," in 7th IFAC Symposium on Information Control Problems, IFAC, Toronto, 1992.

[4] R. Fierro, F.L. Lewis, "Control of a nonholonomic mobile robot: backstepping kinematics into dynamics," in Proceedings of 1995 34th IEEE Conference on Decision and Control, IEEE: 3805–3810, doi:10.1109/CDC.1995.479190.

[5] Yu Tian, N. Sidek, N. Sarkar, "Modeling and control of a nonholonomic Wheeled Mobile Robot with wheel slip dynamics," in 2009 IEEE Symposium on Computational Intelligence in Control and Automation, IEEE: 7–14, 2009, doi:10.1109/CICA.2009.4982776.

[6] Y.H. Amengonu, Y.P. Kakad, "Dynamics and control for Constrained Multibody Systems modeled with Maggi's equation: Application to Differential Mobile Robots Partll," IOP Conference Series: Materials Science and Engineering, **65**, 012018, 2014, doi:10.1088/1757-899X/65/1/012018.

[7] G. Campion, B. d'Andrea-Novel, G. Bastin, Controllability and state feedback stabilizability of non holonomic mechanical systems, Springer-Verlag, Berlin/Heidelberg: 106–124, doi:10.1007/BFb0039268.

[8] A.M. Bloch, N.H. McClamroch, "Control of mechanical systems with classical nonholonomic constraints," in Proceedings of the 28th IEEE Conference on Decision and Control, IEEE: 201–205, doi:10.1109/CDC.1989.70103.

[9] S. Kalaycioglu, "Control of multiple robot manipulators with optimal force distribution," in IEEE Canadian Conference on Electrical and Computer Engineering, 1991.

[10] M. Vukov, S. Gros, G. Horn, G. Frison, K. Geebelen, J.B. Jørgensen, J. Swevers, M. Diehl, "Real-time nonlinear MPC and MHE for a large-scale mechatronic application," Control Engineering Practice, **45**, 64–78, 2015, doi:10.1016/j.conengprac.2015.08.012.

[11] J.B. Rawlings, "Tutorial overview of model predictive control," IEEE Control Systems, **20**(3), 38–52, 2000, doi:10.1109/37.845037.

[12] Y. Shi, K. Zhang, "Advanced model predictive control framework for autonomous intelligent mechatronic systems: A tutorial overview and perspectives," Annual Reviews in Control, **52**, 170–196, 2021, doi:10.1016/j.arcontrol.2021.10.008.

[13] P.D. Christofides, R. Scattolini, D. Muñoz de la Peña, J. Liu, "Distributed model predictive control: A tutorial review and future research directions," Computers & Chemical Engineering, **51**, 21–41, 2013, doi:10.1016/j.compchemeng.2012.05.011.

[14] M. Ellis, H. Durand, P.D. Christofides, "A tutorial review of economic model predictive control methods," Journal of Process Control, **24**(8), 1156–1178, 2014, doi:10.1016/j.jprocont.2014.03.010.

[15] F. Michael, Implementation of Linear Model Predictive Control –Tutorial, 2021.

[16] S. Yu, M. Reble, H. Chen, F. Allgöwer, "Inherent Robustness Properties of Quasi-infinite Horizon MPC," IFAC Proceedings Volumes, **44**(1), 179–184, 2011, doi:10.3182/20110828-6-IT-1002.01969.

[17] H. Wei, C. Shen, Y. Shi, "Distributed Lyapunov-Based Model Predictive Formation Tracking Control for Autonomous Underwater Vehicles Subject to Disturbances," IEEE Transactions on Systems, Man, and Cybernetics: Systems, **51**(8), 5198–5208, 2021, doi:10.1109/TSMC.2019.2946127.

[18] H. Wei, Q. Sun, J. Chen, Y. Shi, "Robust distributed model predictive platooning control for heterogeneous autonomous surface vehicles," Control Engineering Practice, **107**, 104655, 2021, doi:10.1016/j.conengprac.2020.104655.

[19] K. Zhang, Q. Sun, Y. Shi, "Trajectory Tracking Control of Autonomous Ground Vehicles Using Adaptive Learning MPC," IEEE Transactions on Neural Networks and Learning Systems, **32**(12), 5554–5564, 2021, doi:10.1109/TNNLS.2020.3048305.

[20] Y. Zou, X. Su, S. Li, Y. Niu, D. Li, "Event-triggered distributed predictive control for asynchronous coordination of multi-agent systems," Automatica, **99**, 92–98, 2019, doi:10.1016/j.automatica.2018.10.019.

[21] K. Zhang, Y. Shi, "Adaptive model predictive control for a class of constrained linear systems with parametric uncertainties," Automatica, **117**, 108974, 2020, doi:10.1016/j.automatica.2020.108974.

[22] J.S. Ladoiye, D.S. Necsulescu, J. Sasiadek, "Force Control of Surgical Robot with Time Delay using Model Predictive Control," in Proceedings of the 15th International Conference on Informatics in Control, Automation and Robotics, SCITEPRESS - Science and Technology Publications: 202–210, 2018, doi:10.5220/0006908602020210.

[23] R.A. Gangapersaud, G. Liu, A.H.J. de Ruiter, "Detumbling of a non-cooperative target with unknown inertial parameters using a space robot," Advances in Space Research, **63**(12), 3900–3915, 2019, doi:10.1016/j.asr.2019.03.002.

[24] T. Englert, A. Völz, F. Mesmer, S. Rhein, K. Graichen, "A software framework for embedded nonlinear model predictive control using a gradient-based augmented Lagrangian approach (GRAMPC)," Optimization and Engineering, **20**(3), 769–809, 2019, doi:10.1007/s11081-018-9417-2.

[25] K. Rathai, Synthesis and Real-time Implementation of Parameterized NMPC Schemes for Automotive Semi-active Suspension Systems, PhD Thesis, Communaut´e Universit´e Grenoble Alpes, Grenoble, 2020.

[26] R. Quirynen, M. Vukov, M. Zanon, M. Diehl, "Autogenerating microsecond solvers for nonlinear MPC: A tutorial using ACADO integrators," Optimal Control Applications and Methods, **36**(5), 685–704, 2015, doi:10.1002/oca.2152.

[27] F. Aghili, "Optimal control of a space manipulator for detumbling of a target satellite," in IEEE Int. Conf. Robot. Automatica, IEEE, 2009.

[28] T. Rybus, J. Seweryn, J. Sasiadek, "Application of predictive control for manipulator mounted on a satellite," Archives of Control Sciences, **28**(1), 105–118, 2018.

[29] M. Wang, J. Luo, U. Walter, "A non-linear model predictive controller with obstacle avoidance for a space robot," Advances in Space Research, **57**(8), 1737–1746, 2016, doi:10.1016/j.asr.2015.06.012.

[30] M. Morato, J. Normey-Rico, O. Sename, "Model Predictive Control Design for Linear Parameter Varying Systems: A Survey," in Annual Reviews in Control, 64–80, 2020.

[31] E. Psomiadis, E. Papadopoulos, "Model-Based/Model Predictive Control Design for Free Floating Space Manipulator Systems," in 2022 30th Mediterranean Conference on Control and Automation (MED), IEEE: 847–852, 2022, doi:10.1109/MED54222.2022.9837196.

[32] M. Wada, S. Mori, "Holonomic and omnidirectional vehicle with conventional tires," in Proceedings of IEEE International Conference on Robotics and Automation, IEEE: 3671–3676, doi:10.1109/ROBOT.1996.509272.

[33] J. Ostrowski, J. Burdick, "The Geometric Mechanics of Undulatory Robotic Locomotion," The International Journal of Robotics Research, **17**(7), 683–701, 1998, doi:10.1177/027836499801700701.

[34] C. Stöger, A. Müller, H. Gattringer, Parameter Identification and Model-Based Control of Redundantly Actuated, Non-holonomic, Omnidirectional Vehicles, 207–229, 2018, doi:10.1007/978-3-319-55011-4_11.

[35] P.F. Muir, C.P. Neuman, "Kinematic modeling of wheeled mobile robots," Journal of Robotic Systems, **4**(2), 281–340, 1987, doi:10.1002/rob.4620040209.

[36] F.G. Pin, S.M. Killough, "A new family of omnidirectional and holonomic wheeled platforms for mobile robots," IEEE Transactions on Robotics and Automation, **10**(4), 480–489, 1994, doi:10.1109/70.313098.

[37] G. Campion, G. Bastin, B. D'Andrea-Novel, "Structural properties and classification of kinematic and dynamic models of wheeled mobile robots," in [1993] Proceedings IEEE International Conference on Robotics and Automation, IEEE Comput. Soc. Press: 462–469, doi:10.1109/ROBOT.1993.292023.

[38] G. Wampfler, M. Salecker, J. Wittenburg, "Kinematics, Dynamics, and Control of Omnidirectional Vehicles with Mecanum Wheels," Mechanics of Structures and Machines, **17**(2), 165–177, 1989, doi:10.1080/15397738909412814.

[39] A. Gfrerrer, "Geometry and kinematics of the Mecanum wheel," Computer Aided Geometric Design, **25**(9), 784–791, 2008, doi:10.1016/j.cagd.2008.07.008.

[40] L.-C. Lin, H.-Y. Shih, "Modeling and Adaptive Control of an Omni-Mecanum-Wheeled Robot," Intelligent Control and Automation, **04**(02), 166–179, 2013, doi:10.4236/ica.2013.42021.

[41] A. Shimada, S. Yajima, P. Viboonchaicheep, K. Samura, "Mecanum-wheel vehicle systems based on position corrective control," in 31st Annual Conference of IEEE Industrial Electronics Society, 2005. IECON 2005., IEEE: 6 pp., 2005, doi:10.1109/IECON.2005.1569224.

[42] Y. Wang, D. Chang, "Motion performance analysis and layout selection for motion system with four Mecanum wheels," Journal of Mechanical Engineering, **45**(5), 307–316, 2009.

[43] M.O. Tatar, C. Popovici, D. Mandru, I. Ardelean, A. Plesa, "Design and development of an autonomous omni-directional mobile robot with Mecanum wheels," in 2014 IEEE International Conference on Automation, Quality and Testing, Robotics, IEEE: 1–6, 2014, doi:10.1109/AQTR.2014.6857869.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 48-53 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Navigation Aid Device for Visually Impaired using Depth Camera

Hendra Kusuma*, Muhammad Attamimi, Julius Sintara

Department of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia

ARTICLE INFO

Article history:
Received: 07 February, 2023
Accepted: 26 April, 2023
Online: 15 May, 2023

Keywords:
Assitive technology
Disability Inclusion
Depth camera
Navigation aid
Stereo audio
Visual impairment

ABSTRACT

People with visual impairment face daily struggle of navigating through unfamiliar places. This problem mainly caused by their lack of spatial awareness, i.e., the ability to estimate the distance between themselves and their surroundings. In order for visually impaired people to navigate independently, an effective navigation aid is required. The proposed navigation aid device utilizes depth camera to collect visual information of surrounding objects. Then, it represents the obtained visual data into stereophonic sound to notify the user directly through an audio device. The aid device is designed to be portable, comfortable, and easy to use. It can further be developed and upgraded to suit the needs of visually impaired users. Designed to be wearable, this proposed device was tested and received excellent score in portability, comfortability, and ease of use. The subjects were able to detect the position of obstacles in front of them with 92.47% accuracy, and could also estimate the distance of the object with Mean Absolute Error of 0.8. Examination on their navigation ability indicated that the subjects could stop before collision with an object and maneuvers through the gap between two parallel obstacles.

1. Introduction

There are approximately 285 millions visually impaired people in the world with 13.68% of them are totally blind and the rest suffer from low vision [1]. In their daily life, people with visual impairment face many difficulties, especially in navigation due to their inability to observe surrounding environment. The sense of sight is the most fundamental sense to navigate, to perceive the environment, and to identify as well as estimate distance of surrounding objects [2]. Of course, these cannot be achieved by visually impaired so that they have to use other senses to produce spatial perception.

In Indonesia, the infrastructure and public facilities for people with disability are limited. Compared to the others, facilities for visually impaired people are still insufficient. For example, considerable amount of public areas in Indonesia are not equipped with tactile paving [3]. Similarly, for public transportation such as bus and train, assistance for blind people is lacking so that it is very difficult for them to travel independently without the help of others.

In consideration of rapid development in technology, we should be able to help people with visual impairment in overcoming those limitations. With the aid of technology like camera, the visual ability could be conveyed though other senses. Some technologies have been implemented to help people with visual impairments, such as Blind People Guidance System using Stereo Camera [4], as well as facial expression recognition technology using deep learning [5]. However, there are still many other opportunities for application of technology to help people with visual impairments, such as navigate aid device for them to navigate independently .

Of those needs, a wearable navigation aid for visually impaired is necessary. This assistive device needs to be portable, comfortable, and also easy to use. It is also important for this device to have capability for further development, so that the users do not need to change or use more than this navigation aid. In this research, a depth camera is implemented as a sensor to collect visual information of surrounding environment in form of color and depth image. Hereafter, the visual information will be processed and represented though a sound to visually impaired. Visual representation in audio is carried out by various frequencies and amplitudes combination of stereophonic sound to reproduce a spatial perception.

The formatter will need to create these components, incorporating the applicable criteria that follow.

*Corresponding Author: Hendra Kusuma, hendraks@ee.its.ac.id

2. Research Method

In general, the method of this device could be divided into three: data collection, processing, and audio output. First, data is taken using a depth camera, which in this research uses Intel Realsense D435i [6]. This device is chosen because of the active stereo IR technology for depth imaging, equipped with a built-in IMU sensor. Its relatively small shape and low power consumption make this tool suitable for use as a wearable device. The depth camera will capture the user's surroundings for processing. For data processing, NVIDIA Jetson Nano [7] is used as the embedded computing unit. Its compact form, low power, and computability make this device also suitable as a wearable device. After processing, the sound information consisting a combination of tones with certain frequency and pattern, which represents the position and the distance estimation of the obstacle, is output through the stereo headphone or earphone directly to the user.

The block diagram of this navigation aid system is shown in Figure 1.

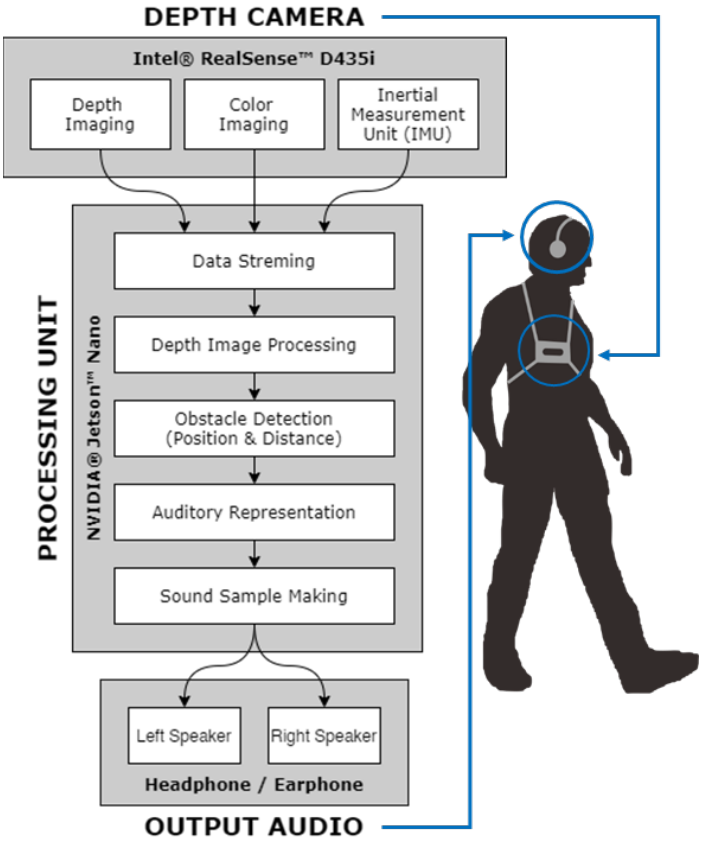


Figure 1: Block Diagram of Navigation Aid Device

2.1. Hardware Design

One of the important points of a wearable device is its design. The hardware design aims to make the device portable, comfortable, and easy to use. First is portability, where all the components used are small in size and light in weight. The power supply used is a battery, so all components must be able to work on a battery, which is why low power consumption is considered. For comfortability, it is necessary for a wearable support to put all the components used in one unit. Therefore, the users do not need to hold anything by hand and the device is integrated into their outfit. Last is ease of use, so that users can use this device independently every day, without the need for help from others or any complicated installation.

The following Figure 2 is the hardware design scheme with all of the components. Thereafter, an illustration of the wearing of this device is shown in Figure 3.

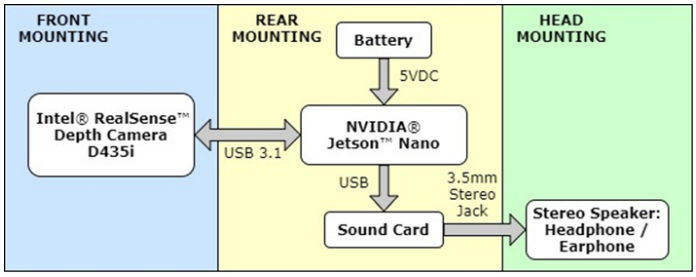


Figure 2: Hardware design scheme

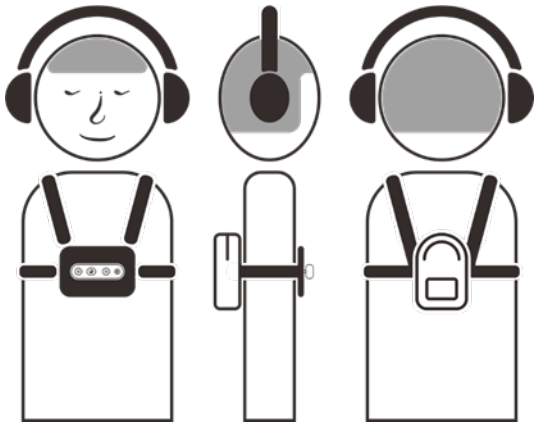


Figure 3: Wearable device installation design

The device is powered by a Lithium-Polymer battery 3 cells with 5200mAH capacity. Two devices that consume most of the power are NVIDIA Jetson Nano and Intel RealSense. NVIDIA Jetson Nano has 7-watt average power, while Intel RealSense consumes 3.5-watt power. The total power consumption is 10.5-watt. With the battery, our device could last about 6 hours of the use.

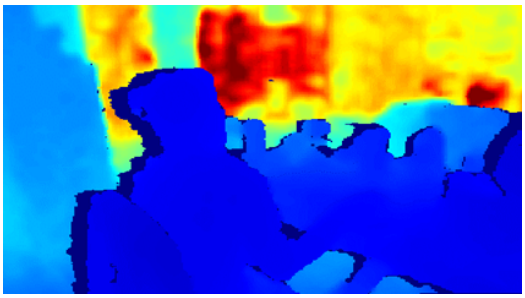
2.2. Software Design

In software design, there are several processes before an image can be represented in audio. First is data retrieval from RealSense™. Information such as depth images, color images, and IMU data can be retrieved from RealSense™ using the SDK provided by Intel® which was developed in open source [8]. The Intel® RealSense™ SDK 2.0, or librealsense, is equipped with a cross-platform library that can be used for various RealSense™ depth camera products. In this study, the Python wrapper from librealsense was used in the Python 3.0 programming environment.

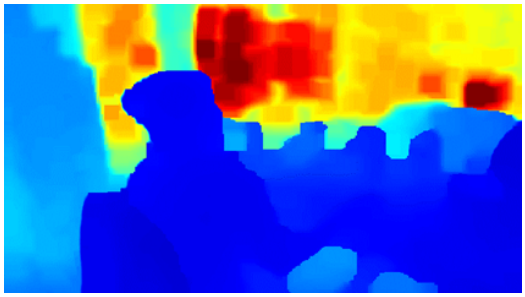
In streaming mode, a callback function will be called every time a new data is available from the sensors. The callback function is run in different thread from the main loop and will store the data from the sensors to variables that could be accessed from

the main loop. Therefore, the information could be obtained simultaneously and the main process could be run at the same time.

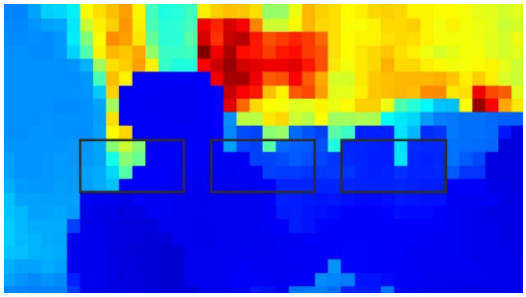
Next is depth image processing. Intel® RealSense ™ products are equipped with an API that is easy to use for various purposes, either with a GUI or in the form of a library, to retrieve data, both depth and color images, in standard units (millimetres for depth images and 8-bit RGB bits for color images). However, the results obtained cannot be used directly. Image processing is required so that the image can be used for the next step. There are several problems in depth images, including depth images and color images that have different viewpoints, unreadable depth values, and how to take values that represent an area of a certain size in the depth image. Depth image processing plays an important role so that the image can be further processed to extract the information. Figure 4 shows the results before and after depth image processing, and also the area that is being used for obstacle detection.



(a)



(b)



(c)

Figure 4. Depth image processing, (a) before processing, (b) after processing, (c) the area used for obstacle detection

The next step is detection of obstacles and their position. From the processed image, it will then be processed to detect existing obstacles. In this study, the position of the obstacle was limited into three parts: front left, front middle, and front right. Each position represents an area with a number of pixels in the depth image. The following Figure 5 is an illustration of the division of the depth image position with α, β, and γ angle for front left, front middle, and front right respectively. For vertical field of view, we used ±20°

For the depth distance in this study, the minimum depth distance is 40 cm, and the maximum depth distance is 160 cm. The meaning of the minimum depth distance is that if the object is closer than the minimum depth distance, the object will be considered very close to the minimum depth measurement of the camera. The meaning of the maximum depth distance is that if the object is farther than the maximum depth distance, then the object will be ignored or considered as no object. This value can be changed and adjusted according to the needs and convenience of each user, but in this study the determined value is used. Distance is also determined based on the accuracy of the depth in this system. In this study, a system depth accuracy of 40 cm was determined, which means that the depth information in the 40 cm range would be considered the same. Thus, there will be 5 categories. First, the undetectable depth beyond the maximum depth limit (160 cm). Second, namely the depth between the maximum limit to the depth accuracy in this case between 120 cm to 160 cm. The third is between 80 cm to 120 cm. The fourth is between 40 cm and 80 cm. The last is a depth that is smaller than the minimum depth limit (40 cm). At this stage the data obtained is in the form of three depth information (front left, front center, and right front) which have been classified into their respective categories (in this case categories 1 to 5).

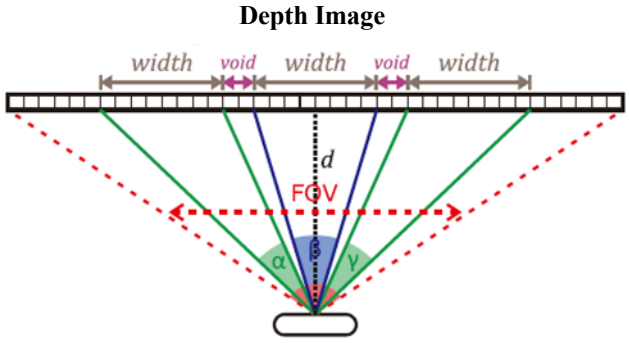


Figure 5: Position division in the depth image

After obtaining depth information to be conveyed in the form of position and depth or distance, it is necessary to represent the information in audio.

The most important thing about this step is to choose the right parameters to represent the depth information that is available. In this experiment, through try and error, the following parameter settings were obtained:

1. For stereo audio, use the volume settings for the right channel and the left channel. When the object is on the front left, the audio will give full volume to the left channel. When the object is in front of the center, the audio will give half volume to the right channel and the left channel. Likewise, when the object is in front of the right, the audio will give full volume to the right channel.
2. For frequency selection, after going through the try and error process, three types of frequencies are used, namely the tones C5, D5, and E5. The respective frequencies are 523.25 Hz,

587.33 Hz, and 659.26 Hz to represent the positions of the front left, front center, and front right, respectively. It is observed that the use of do re mi tones is easier for users to understand and remember.

3. For duration, beeps pattern is used to provide the user with a perception of distance. The farther the object, the slower the beep, as well as the closer the object, the beep will be faster and the pause between beeps will also be faster. Since 5 categories have been determined, and for the first category there are no objects, there are 4 beeps sound patterns, each of which represents the depth distance category. In writing this report, to make it easier to describe the depth category, a simple name is used, namely "no object" for the category of depth further than the maximum limit, then respectively "far", "medium", "close", and finally "very close" for categories where the object is closer than the minimum limit.

Sound information is conveyed through the stereo headphone through periodic cycle. The speed of one cycle will be adjusted according to the user's capability. In one cycle, amplitude of the right audio will be decreased from maximum amplitude, while the left audio will be increased toward maximum amplitude. Three different tones will sound according to the Figure 5 with smooth transition. Therefore, the user could experience surround audio to visualize the spatial information.

3. Results And Discussion

3.1. Installation Test: Portability, Comfortability, and Ease of Use

As a wearable device, the first test is regarding the wear of the device Prototype design of the navigation aid for the visually impaired that has been assembled is installed to the subjects with visual impairment to test the design results. The results of installing a navigation aid on visually impaired subjects are as shown in Figure 6.

(a)

(b)

Figure 6: Installation of navigation aid device on visually impaired subjects, (a) back view, (b) front view

From this installation, we tested the portability, comfort and ease of use of the device in visually impaired subjects. After use for a while, visually impaired subjects were asked to provide an assessment of the value of portability, comfortability, and ease of use with several questions that had a correspondence with the three aspects being measured.

The answers to the questions are classified into positive answers, negative answers, or neutral answers. Answers which are positive answers include comfortable when used, not burdensome, easy to use daily, no difficulty in wearing the device, not limiting movement, not disturbing, etc. For answers that are negative answers include uncomfortable when used, burdensome, difficult to use, cannot be used daily, difficult wear in and / or remove the device, the device limits movement, annoying use of headphones / earphones. Answers that are neutral answers are answers that do not include positive or negative answers including answers with certain reasons or conditions such as a comfortable tool to use but within a certain period, the use of the tool does not limit movement if it is used at certain times, etc.

From the use experiments carried out on two blind subjects, the results are shown in table 1.

Table 1: Result of Qualitative Test On The Device Usage

Question Number	Measured Aspect	Subject 1	Subject 2
I	Comfortability	Positive	Positive
II	Portability	Positive	Positive
III	Portability & Comfortability	Positive	Positive
IV	Ease of use	Positive	Neutral
V	Ease of use	Positive	Neutral
VI	Comfortability	Positive	Positive
VII	Comfortability	Positive	Positive

In addition to qualitative questions, subjects were also asked to provide quantitative assessments for the value of portability, comfortability, and ease of use. Subjects were asked to give an assessment in the form of a number between one and ten (1-10)

with a value of 1 being the lowest and 10 being the highest. The results of the quantitative assessment of two blind subjects are obtained in Table 2.

Table 2 : Result of Quantitative Test On The Device Usage

Measured Aspect	Subject 1	Subject 2
Portability	10	10
Comfortability	10	9
Ease of use	9	9

From the results of the hardware installation testing carried out, the portability, comfortability, and ease of use values were quite good by both subjects. Furthermore, for quantitative assessments with an assessment range of one to ten (1-10) with a value of 1 being the lowest and a value of 10 being the highest, an average value of 10 was obtained for portability, 9.5 for comfortability, and 9 for ease of use.

3.2. Functionality Test: Obstacle Position Detection

The overall system in the form of a navigation aid for the visually impaired is tested on a visually impaired subjects for the functionality of the device.

First, according to the system design, obstacle detection is grouped into three areas, namely obstacles in front of the left, obstacles in front of the middle, and obstacles in front of the right. From these three areas, each obstacle was tested in each area individually and also in combination to find out whether the blind subject could tell whether there were obstacles in that area.

The test was carried out with a combination of laying obstacles according to Table 3. From the tests carried out by two blind subjects, the results obtained in Table 4.

Table 3 : Obstacle Position For Testing

Obstacle Position		
Front Left	**Front Middle**	**Front Right**
None	None	None
Exist	None	None
None	Exist	None
None	None	Exist
Exist	Exist	None
Exist	None	Exist
None	Exist	Exist
Exist	Exist	Exist

Table 4. Result of functionality test on the device usage

Test Subjects	Subject 1	Subject 2
True Positive (TP)	12	9
True Negative (TN)	10	10
False Positive (FP)	2	4
False Negative (FN)	0	3
Accuracy	0.92	0.73
Precision	0.86	0.69
Recall	1	0.75
F1 Score	0.9247	0.7188

The results of this test is maximum accuracy of 92.47%. From the accuracy value obtained, this device can function properly to detect the presence or absence and position of obstacles.

For precision, the maximum value is 86%, and the maximum recall is 100%. From these two values, the F1 score was 92%.

3.3. Functionality Test: Obstacle Distance Estimation

Furthermore, a test is conducted to determine the distance estimation between the subject and the existing obstacles. In accordance with the system design, because the minimum and maximum depth values chosen are 0.4 meters and 1.6 meters with depth accuracy in the system design of 0.4 meters, so the depth is divided into 5 bucketized categories, namely distances above 1.6 meters detected as not obstacles, distances between 1.6 meters and 1.2 meters, the distance between 1.2 meters and 0.8 meters, the distance between 0.8 meters and 0.4 meters, and also the distance that is closer than 0.4 meters.

From the tests carried out on two blind subjects, the results are in Table 5.

From the tests that have been done, the best MAE value or mean absolute error is 0.8 (for bucketized categories). The obtained value is decent for the error rate in distance estimation. A small MAE value indicates that distance estimation errors occur for adjacent category.

Table 5 : Result of Obstacle Distance Estimation Test

Subject		Subject 1	Subject 2
Mean Absolute Error (MAE)	< 0.4 meter	1	1
	0.4 – 0.8 meter	1	2
	0.8 – 1.2 meter	1	3
	1.2 – 1.6 meter	1	1
	> 1.6 meter	0	0
	Average	0.8	1.6

3.4. Functionality Test: Simple Paths

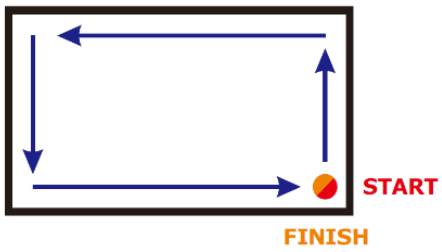
The final test is the application of tools to blind subjects in walking on a predetermined route, to simulate some of the conditions that occur in daily navigation. In this test, there are two routes as illustrated in Figure 7.

In this test, there are two aspects tested. The first is whether visually impaired subjects can avoid collisions with the wall by stopping right before the wall without any other assistive devices. The second is whether the blind subject can spot and manoeuvre through the gap between two parallel obstacles.

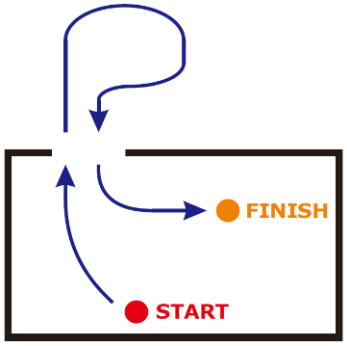
From the conducted test, visually impaired subjects can stop before a collision occurs with the obstacle in front of them, in this case is a wall. Visually imapired subject can also manoeuvre through the gap between two parallel obstacles, in this case is a opened gate.

4. Conclusion

A navigation device is needed by visually impaired people to navigate in their daily life. Therefore, this navigation aid device is design to be portable, comfortable, and ease to use; as evidenced by questionnaire given to blind subjects after wearing this device.



(a)



(b)

Figure 7: Simple paths for navigation aid device testing, (a) first route, (b) second route

The sensor used is Intel® Realsense D435i, which could capture color and depth images sufficiently, and is equipped with IMU. For this research, the information used is only the depth image, but could be developed further using various techniques to maximize the use of the information. The CPU used is Jetson™ Nano with very limited computational capabilities. For additional features and more complex object detection, a CPU that is more powerful with high computational capabilities is required, but still portable in size and could be powered by a battery.

The navigation aid functionality test shows the accuracy of the obstacle detection within three position division is 92.47% and the MAE error (mean absolute error) of the distance estimation to the obstacle is 0.8 for the obstacle distance setting that is less than 1.6 meters from the user. Furthermore, without the help of other tools such as cane, the users can stop before a collision with an obstacle in front, and walk through the gap between two parallel obstacles, according to testing on the simple paths.

This device testing is still limited to a few subjects. Henceforth, for future works, this device can be tested on more subjects with various ages, levels of visual impairment, and backgrounds. Additionally, different method of amplitude and frequency transition can also be explored to observe the effectivity of the device usage to the users.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under project scheme of the Publication Writing and IPR Incentive Program (PPHKI).

References

[1] D. Pascolini, S. P. Mariotti, “Global estimates of visual impairment : 2010,” Br. J. Ophthalmol, **96**(5), 614–618, 2012.

[2] “Daily Life Problems Faced by Blind People.” [Online]. Available: https://wecapable.com/problems-faced-by-blind-people/. [Accessed: Oct. 30, 2019].

[3] E. Khoirunisa, D. Aries Himawanto, “The comparison of guide texture tiles for blind people in public areas between Surakarta and Nagoya city,” Jurnal Kajian Wilayah, **9**(1), 34, 2018.

[4] I. P. Adi, H. Kusuma, M. Attamimi, "Blind People Guidance System using Stereo Camera," in 2019 International Seminar on Intelligent Technology and Its Applications (ISITIA), Surabaya, Indonesia, 298-303, 2019, doi: 10.1109/ISITIA.2019.8937173.

[5] H. Kusuma, M. Attamimi, H. Fahrudin, “Deep learning based facial expressions recognition system for assisting visually impaired persons,” Bulletin of Electrical Engineering and Informatics, **9**(3), 1208-1219, 2020.

[6] Intel, “Intel® RealSense™ Camera D400 series Product Family Datasheet Rev. 01/2019,” 2019.

[7] NVIDIA, “DATA SHEET NVIDIA Jetson Nano System-on-Module Maxwell GPU + ARM Cortex-A57 + 4GB LPDDR4 + 16GB eMMC,” 2020.

[8] Intel, “IntelRealSense/librealsense: Intel® RealSense™ SDK.” [Online]. Available: https://github.com/IntelRealSense/librealsense. [Accessed: Jun. 3, 2020].

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 54-62 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Forecasting the Weather behind Pa Sak Jolasid Dam using Quantum Machine Learning

Chaiyaporn Khemapatapan*, Thammanoon Thepsena

Computer Engineering Program, College of Innovative Technology and Engineering Dhurakij Pundit University Bangkok, Thailand

ARTICLE INFO

Article history:
Received: 27 February, 2023
Accepted: 30 April, 2023
Online: 15 May, 2023

Keywords:
Machine Learning
Quantum Machine Learning
Quantum Circuit
Variational Quantum Classifier
Pa Sak Jolasid Dam

ABSTRACT

This paper extends the idea of creating a Quantum Machine Learning classifier and applying it to real weather data from the weather station behind the Pa Sak Jonlasit Dam. A systematic study of classical features and optimizers with different iterations of parametrized circuits is presented. The study of the weather behind the dam is based on weather data from 2016 to 2022 as a training dataset. Classification is one problem that can be effectively solved with quantum gates. There are several types of classifiers in the quantum domain, such as Quantum Support Vector Machine (QSVM) with kernel approximation, Quantum Neural Networks (QNN), and Variational Quantum Classification (VQC). According to the experiments conducted using Qiskit, an open-source software development kit developed by IBM, Quantum Support Vector Machine (QSVM), Quantum Neural Network (QNN), and Variable Quantum Classification (VQC) achieved accuracy 85.3%, 52.1%, and 70.1% respectively. Testing their performance on a test dataset would be interesting, even in these small examples.

1. Introduction

Programming computers to learn from data is the subfield of artificial intelligence (AI) known as machine learning (ML). In machine learning, support vector machines (SVM) are among the most frequently used classical supervised classification models [1]. The decision boundary and the hyperplane of the data points are divided into two classes by a pair of parallel hyperplanes that are discovered by SVM [2, 3]. However, there is also machine learning at the particle level called quantum computing. Quantum computing is computation using quantum mechanical phenomena such as superposition and entanglement. The difference is from the computer we use today, which is an electronic base on binary state based on transistors. Whereas simple digital computing requires data to be encoded into a binary number where each bit is in a certain state 0 or 1, quantum computing uses quantum bits (qubit). This can be a superposition of state, both 0 and 1 simultaneously. In quantum computing, a new algorithm is required for that problem, i.e., a normal algorithm used in a classical computer cannot be copied and run on a quantum computer at all. The Quantum Computer Algorithm for popular algorithms such as Prime factorization of integers Shor's algorithm is a quantum algorithm that can attack the algorithm RSA and encryption process of 90% of computer systems worldwide in a short period. Quantum computers can also operate on qubits using quantum gates and measurements that change the observed state. Quantum gates and problems encode input variables into quantum states. To facilitate further modeling of the quantum state, quantum algorithms often exhibit probabilities in which they provide guidance for valid only for known probabilities. Quantum machine learning (QML) is an emerging interdisciplinary research field that combines quantum physics and machine learning, using it to help optimize and speed up data processing on the quantum state. In addition to the widespread popularity of QML, there is also the variational quantum classifier (VQC) for solving classification problems. At present, IBM has developed a quantum computer open to researchers or anyone interested in using it called IBM Q Experience, with a set of instructions developed in Python called Qiskit, which has a simulated quantum computer and real 5- and 15-qubit quantum computers to develop and test circuits. In this article, we present an experiment, which is a continuation of previous experiments [4, 5], that studied the forecast of water release from the dam and the weather forecast behind Pa Sak Jolasid Dam, respectively. Both experiments used a classical machine learning process that measures all model results as model accuracy. The model results are satisfactory."

In this paper, an experiment was performed using a Quantum Machine Learning classifier and applying it to real data, which brought information from the weather station located behind the Pa Sak Jonlasit Dam. The Pa Sak Jolasid Dam is an earth dam with a clay core, 4,860 meters long, and 31.50 meters high. The

*Corresponding Author: Chaiyaporn Khemapatapan, chaiyaporn@dpu.ac.th

maximum storage water level is +43.00 MSL, and the water storage capacity is 960 million cubic meters. The total operational budget is 19,230.7900 million baht, and the satellite coordinates are n14.964687, e101.022677 (see Fig.1). The red dot on the map represents the location of the weather station. Studying weather conditions, especially forecasting rainy days, can benefit water inflow management from quantum machine learning classifier techniques applied to actual weather data from Table 1. The total number of data is 1743 samples, divided into 1220 samples of training data and 523 samples of testing data. The number of features has 4 samples or 4 input qubits, and the label has 2 classes. Table 2 shows a sampling of the values of the features, which are average wind, average temperature, average pressure, average humidity, and label values.

A systematic study of the classical feature and optimizer with the different iterations of the parametrized circuits is presented. The study of the weather behind the dam is based on weather data from 2016 to 2022 as a training dataset. Classification is one problem that can be effectively solved with quantum gates. There are several types of classifiers in the quantum domain, such as Quantum Support Vector Machine (QSVM) with kernel approximation, Quantum Neural Networks (QNN), and Variable Quantum Classification (VQC). According to the experiment, Quantum Support Vector Machine (QSVM), Quantum Neural Network (QNN), and Variable Quantum Classification (VQC) achieved 90% accuracy. All of these algorithms were performed using Qiskit, an open-source software development kit (SDK) developed by IBM.

The classification is one problem that can be effectively solved with quantum gates. There are several types of classifiers in the quantum domain such as Quantum Support Vector Machine (QSVM) with kernel approximation [6-8], Quantum Neural Networks (QNN) [9, 10], and Variable Quantum Classification (VQC) [11-14]. The experimental results proved that we can use QML to solve real-world problems that are classically trained and tested before encoding the feature map, evaluating the model, and optimizing it from the algorithm above.

In this article, we will discuss the origin of the theory of quantum applied in section 2, followed by the steps and methods in section 3. Section 4 discusses the experimental results and explains the reasoning. Finally, section 5 provides a summary of the experiments and recommendations.

Regarding the experiment, we applied Quantum Machine Learning classifiers to real data from the weather station located behind the Pa Sak Jolasid Dam. This earth dam has a clay core, and it is 4,860 meters long and 31.50 meters high, with a maximum storage water level of +43.00 MSL, and a water storage capacity of 960 million cubic meters. The total operational budget is 19,230.7900 million baht, with satellite coordinates: n14.964687, e101.022677 (see Fig.1). The red dot on the map represents the location of the weather station, which studies weather conditions, especially forecasting rainy days, and can benefit water inflow management from quantum machine learning classifier techniques applied to actual weather data from Table 1.

The total number of data is 1743 samples, divided into 1220 samples of training data and 523 samples of testing data. The number of features has 4 samples or 4 input Qubits, and the label has 2 classes. Table 2 shows a sampling of the values of the features, which are average wind, average temperature, average pressure, average humidity, and label values. We present a systematic study of the classical feature and optimizer with the different iterations of the parametrized circuits.

In conclusion, the experimental results demonstrate that QML can be used to solve real-world problems, which are classically trained and tested before encoding the feature map, evaluating the model, and optimizing it from the algorithm above. Therefore, the potential applications of quantum machine learning classifiers are promising, and more research in this area should be encouraged

2. Related work

Since we have the weather dataset for the dam, we can make predictions based on the training data. This is a binary classification problem with an input vector x and binary output y in {0, 1}. The goal is to build a quantum circuit that produces a quantum state based on the following study.

2.1. Quantum Computing

What exactly is a quantum computer then? In a nutshell, it could be described as a physical implementation of n qubits with precise state evolution control. A quantum algorithm, according to this definition of quantum computers, is a controlled manipulation of a quantum system followed by a measurement to obtain information from the system. This basically means that a quantum computer can be thought of as a special kind of sampling device. However, because it is a quantum state, the configurations of the experiments are very important. Any quantum evolution can be approximated by a series of elementary manipulations, known as quantum gates, according to a theorem in quantum information [15]. Quantum circuits of these quantum gates are the basis for many quantum algorithms. The idea of a qubit came from upgrading classical bits [16, 17], which are 0 or 1, to a quantum state.

$$0 \to |0\rangle = \begin{pmatrix}1\\0\end{pmatrix}, 1 \to |1\rangle = \begin{pmatrix}0\\1\end{pmatrix}$$

However, what are qubits? Because it is a two-level system defined on $\mathbb{C}^2$, a qubit is frequently referred to as the simplest possible quantum system. This state can be formulated as

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \quad (1)$$

with $\mathbb{C}$ (α, β) and $|\alpha|^2 + |\beta|^2 = 1$, where $|0\rangle$ and are hardware-defined orthonormal states known as computational basis states. The qubit is significant because it is in a superposition—that is, it is a either $|0\rangle$ or $|1\rangle$ at the same time—which means that, in contrast to classical bits, it possesses a mixture of both. Using tensor products, we can generalize this to include *n* untangled qubits.

$$|\psi\rangle \equiv |q_1\rangle \otimes |q_2\rangle \otimes \ldots \otimes |q_n\rangle \quad (2)$$

Figure 1: Pa Sak Jolasid Dam, Coordinates: n14.964687, e101.022677.

Table 1. Dataset Attributions

Datasets	Number of Class	Number of Features	Train Size	Test Size	Total Size
Weather behind Dam	2	4	1220	523	1743

Table 2. Partial Dataset

Date	Avg. Wind	Avg. Temp	Avg. Pressure	Avg. Humid	Rainfall
2016/10/14	2.3	28	999.4	46	Yes
2017/01/04	6.6	26	1003.7	64	No
2018/03/07	3.3	25.9	1002.8	27	Yes
2019/08/01	3.9	25.4	998.4	86	Yes
2020/10/11	3.4	27.8	999.3	73	Yes
2021/02/22	2.1	26.9	1003.2	48	No
2022/05/12	2.7	28	1000.5	82	Yes

where $|q_i\rangle$ stands for qubits. However, the state $|\psi\rangle$ would no longer be separable if the qubits were entangled, and every qubit would either be $|0\rangle$ or $|1\rangle$, resulting in

$$|\psi\rangle = \alpha_1|0\ldots00\rangle + \alpha_2|0\ldots01\rangle + \cdots + \alpha_{2^n-1}|1\ldots11\rangle \quad (3)$$

with $\alpha_i \in \mathbb{C}$, and $\sum_{i=0}^{2^n-1}|\alpha_i|^2 = 1$. Wherever we use the abbreviated notation $|a\rangle\otimes|b\rangle := |ab\rangle$. To make the notation more elegant, we see that the basis states can be written as follows: $|000\rangle\leftrightarrow|0\rangle$,…,$|111\rangle\leftrightarrow|7\rangle$ giving us the straightforward equation. This allows us to translate the notation from binary numbers to integers.

$$|\psi\rangle = \sum_{i=0}^{2^n-1} \alpha_i\,|i\rangle \quad (4)$$

As a result, {$|0\rangle$ …, $|i\rangle$} and n serve as the computational foundation for n qubits. Since there are 2^n distinct strings, one requires 2^n amplitudes α_i to describe the state of n qubits, as we can see. In other words, quantum information is "larger" than classical information because the information stored in a quantum state with n qubits is exponential in n, whereas classical information is linear in n. which suggests quantum advancements thus far.

2.2. Quantum Circuit

We must begin by examining quantum gates in order to construct a quantum algorithm or quantum circuit [18, 19, 20], as mentioned earlier. Unitary transformations are the means by which quantum gates, or rather quantum logic gates, are produced. A straightforward transformation can serve as a quick reminder of what this means.

$$|\emptyset\rangle = U|\psi\rangle$$

where $|\emptyset\rangle$ and $|\psi\rangle$ are two vector spaces in which U is a unitary operator. By "unitary," mean that the hermitian conjugate of the operator is the inverse, $U^\dagger = U^{-1}$, and that the operator is linear. This is important because we can use it to, for example, display

$$\langle\emptyset|\emptyset\rangle = \langle\psi|U^\dagger U\,|\psi\rangle = \langle\psi|\psi\rangle = 1$$

where, if $|\psi\rangle$ is normalized, then by construction it is $|\emptyset\rangle$.

2.2.1 A single Qubit

If we return to the subject of quantum gates, equation (1), the state would either be in the state $|0\rangle$, which has a probability of $|\alpha|^2$ or in the state $|1\rangle$, which has a probability of $|\beta|^2$. Formally, 2 x 2 unitary transformations are used to describe single-qubit gates. We can begin by considering the X gate, which functions as the quantum equivalent of the classical NOT gate.

$$|0\rangle \longmapsto |1\rangle$$

and the reverse This matrix is easily identifiable as one of the Pauli matrices, which are unitary by definition. As a result, we know that we can have at least X, Y, and Z gates with the unitary operators Pauli matrices.

$$\sigma_x = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}, \qquad \sigma_y = \begin{pmatrix}0 & -i\\ i & 0\end{pmatrix}, \qquad \sigma_z = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix} \quad (5)$$

The Pauli rotations are yet another useful set of gates. which are expressed as Pauli gates that are exponential

$$R_j(\theta) = e^{-i\frac{\theta}{2}\sigma_j} \quad (6)$$

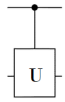
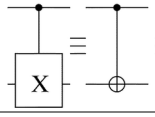
where j is (x, y, z). Since the global phase ($e^{i\gamma}$), the azimuthal (θ) and polar ($\emptyset$) angles can be written into any quantum state,

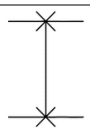
$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$$
$$= e^{i\gamma}\left(cos\frac{\theta}{2}|0\rangle + e^{i\emptyset}sin\frac{\theta}{2}\,|1\rangle\right) \quad (7)$$

2.2.2 Multi Qubit

The controlled U gate is introduced because that work on multiple qubits simultaneously. where U can be any unitary gate with one qubit. For instance, the CNOT gate is obtained by setting $U = x$, and the NOT (X) operation is carried out when the first qubit is in state$|1\rangle$; otherwise, nothing changes in the first qubit. A variety of quantum gates, their circuit, and how they are represented in a matrix show table 3. In a controlled gate, the U is a general unitary operator. We refer to j as (x, y, z) and σ_j denotes the appropriate Pauli matrix Eq. (5)

Table 3. Summary of all the gates in circuit and matrix representation.

Gate	Circuit representation	Matrix representation
H, Hadamard		$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$
Z, Phase Flip		$\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}$
U, Unitary		$\begin{pmatrix}1 & 0\\ 0 & U\end{pmatrix}$
Controlled Not Controlled X CNot		$\begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 1 & 0\end{pmatrix}$

Swap		$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

2.3. Validation and Measurement

The measurement process is the final step in the theory for quantum computers regarding the quantum circuits that make up a quantum evolution [20]. From quantum mechanics, projectors of the Eigen spaces provide the probability of measuring a state. The probability of measuring i = {0, 1} is

$$p(i) = Tr(P_i|\psi\rangle\langle\psi|) = \langle\psi|P_i|\psi\rangle = |\alpha_i|^2 \qquad (9)$$

The qubit's state changes to

$$|\psi\rangle \rightarrow \frac{P_i|\psi\rangle}{\sqrt{\langle\psi|P_i|\psi\rangle}} = |i\rangle$$

The qubits that are able to write the observables as a spectral decomposition of the computational basis are used to estimate the expectation value.

$$\hat{\mathcal{O}} = \sum_{i=1} \lambda_i \, |i\rangle\langle i|$$

where P_i is present. Using a Z gate, the observable yields an eigenvalue of +1 for state $|0\rangle$ and -1 for state $|1\rangle$ so that we can computationally determine which state it is in (9).

$$\langle\psi|\hat{\mathcal{O}}|\psi\rangle = \sum_i \lambda_i \, |\alpha_i|^2 \qquad (10)$$

Since all that is required to determine the eigenvalues, λ_i is an estimation of the state's amplitudes. Since statistics can be used to measure states' amplitudes directly. They introduce a random Bernoulli variable called y_{ij}, where $P(y_{ij} = 0) = 1 - |\alpha_i|^2$ and $P(y_{ij} = 1) = |\alpha_i|^2$ [21]. If repeatedly prepare the state $|\psi\rangle$ and measure it in the computational basis and collect S samples (y_{i1}, …, y_iS), additionally, be aware that $|\alpha_i|^2$ the frequents estimator $\hat{p}_i$ can estimate $|\alpha_i|^2$ by

$$|\alpha_i|^2 \approx \hat{p}_i = \frac{1}{S}\sum_{j=1}^{S} y_{ij}$$

where $\hat{p}_i$'s standard deviation can be found

$$\sigma(\hat{p}_i) = \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{S}}$$

where $\mathcal{O}(S^{-1/2})$ represents the error. Can now approximate (10) to

$$\langle\psi|\hat{\mathcal{O}}|\psi\rangle \approx \sum_i \lambda_i \, \hat{p}_i \pm \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{S}} \qquad (11)$$

2.4. Quantum Support Vector Machine (QSVM)

To efficiently compute kernel inputs, the quantum support vector machine QSVM and the quantum kernel estimator (QSVM-Kernel) [20, 22] make use of the quantum state space as a feature space. By applying a quantum circuit $\Gamma_{\phi(\vec{x})}$ to the initial state $|0^{\otimes n}\rangle$, this algorithm nonlinearly maps the classical data x to the quantum state of n qubits:

$$|\phi(\vec{x})\rangle = \Gamma_{\phi(\vec{x})}|0^{\otimes n}\rangle \qquad (12)$$

The 2^n-dimensional feature space created by the quantum circuit $\Gamma_{\phi(\vec{x})}$ (where n is the number of qubits) is challenging to classically estimate. There are two consecutive layers in this circuit.

$$\Gamma_{\phi(\vec{x})} = U_{\phi(\vec{x})}H^{\otimes n}U_{\phi(\vec{x})}H^{\otimes n} \qquad (13)$$

where $U_{Phi(\vec{x})}$ is a unitary operator that encodes the traditional input data, and H is a Hadamard gate Fig. 2.

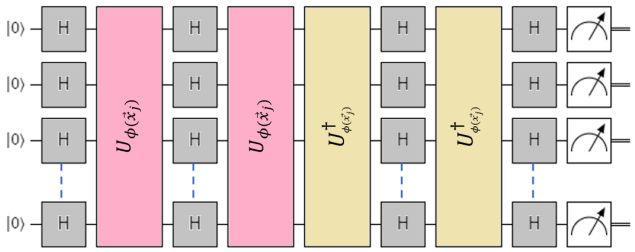


Figure 2: The Circuit of QSVM

During the training phase, the kernel entries are evaluated for the training data and used to locate a separation hyperplane. After that, during the test phase, the new data x and the training data, which are used to classify the new data x according to the separation hyperplane, are used to evaluate the kernel inputs. Quantum computers evaluate the kernel inputs, while classical computers, like those used in a traditional SVM, are used for data classification and separation hyperplane optimization.

2.5. Variational Quantum Circuit

A variational circuit with four features is proposed in [22] to classify the dataset Fig. 3. The variational circuit performs the following operations. The $|0\rangle$ state is used to initialize the circuit's four qubits. The qubits are then placed in a superposition of $|0\rangle$ and $|1\rangle$ by applying the Hadamard gate one at a time. Then, a unitary square matrix designed for state preparation is used to perform a unitary operation on each qubit. The classical data (bits) are encoded into qubits in this method.

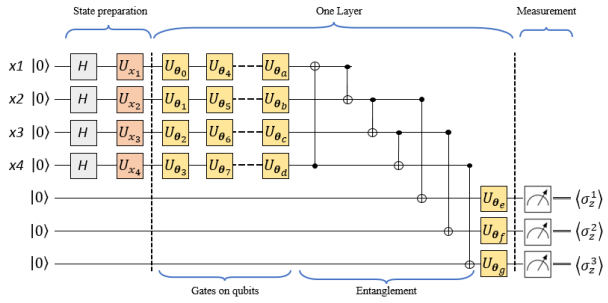


Figure 3. Variational Quantum Circuit

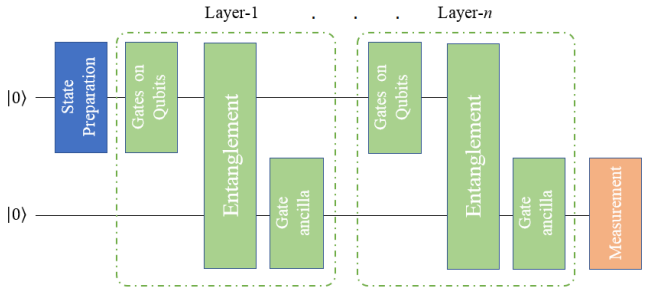


Figure 4. The Variational Quantum Circuit Architecture

Using multiple layers of interleaved rotational gates in data and auxiliary qubits, the variational circuit is designed following state preparation. Optimization is used to adjust the parameters. Fig. 4. shows the seven-layer initial implementation of the circuit as well as the architecture of the variational circuit model. The class label is obtained by processing the resulting qubits and measuring the auxiliary qubits.

2.6. Quantum Amplitude Estimation (QAE)

In [20, 23], a hybrid quantum autoencoder (HQA) variant of the Quantum Amplitude Estimation (QAE) algorithm was proposed [24, 25]. Quantum neural networks (QNNs) based on parameterized quantum circuits (PQC) were utilized in this model, which incorporates both classical and quantum machine learning. The model's overall structure consists of an encoder and a subset of real vector space V of dimension $v = dim(V)$, that transports a quantum state from Hilbert space $H^{\otimes n}$, as well as a decoder that does the opposite of that. The encoder and decoder's functional forms are specified, but the models themselves are not specified. As depicted in Fig. 5, the ϵ encoder is a vector α controlled quantum circuit. The circuit applies the unitary $U_l(\alpha)$ after receiving some $|\psi_{in}\rangle$ states. On the system that combines the input state (v-n) with auxiliary qubits.

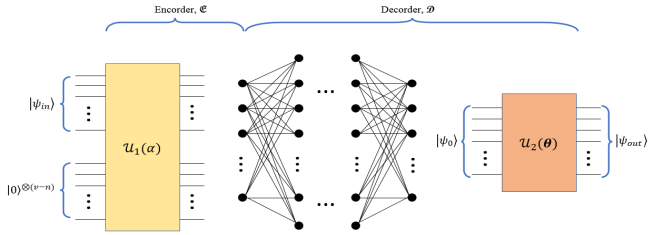


Figure 5. The Quantum Amplitude Estimation Architecture

3. Methodology

The weather dataset from the Thai Meteorological Department was used as the comparative dataset in the research. The dataset consists of 1773 data sets collected from 2016 to 2022, with 1220 of them being designated for training and 523 for testing. The data is divided into 4 variables for the features and 2 classes for the labels, as shown in Tables 2 and 3. However, there is a fairly standard approach to preprocessing. These strategies are not generally reasonable for planning adequate information for quantum classifiers while working with genuine informational collections. It has proposed a preprocessing strategy in this study, as depicted in Fig. 6, which encrypts the data before the QML algorithm uses it. Two QML classifiers are used in this article:

- A quantum support vector machine
- Build a quantum neural network (also known as Variational Classifier)

Both of the QML classifiers utilized preparation of feature maps, implementation of variational circuits, and measurement. The study analyzed the optimizer's feature map, the depth of the variational circuit [26], and the depth of the feature map to understand why these models perform optimally, and attempted to determine if the new information can be effectively condensed.

We have using the Qiskit framework for quantum computing. A typical quantum machine learning model consists of two parts, as shown in Fig. 6, A classical part for pre- and post-processing data and a quantum part for leveraging the power of quantum mechanics to simplify certain calculations.

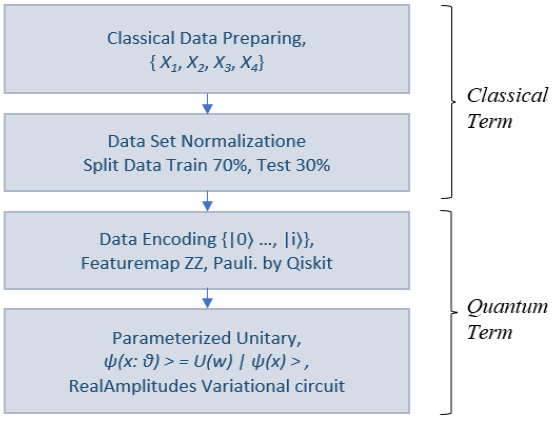


Figure 6: Experimental procedures

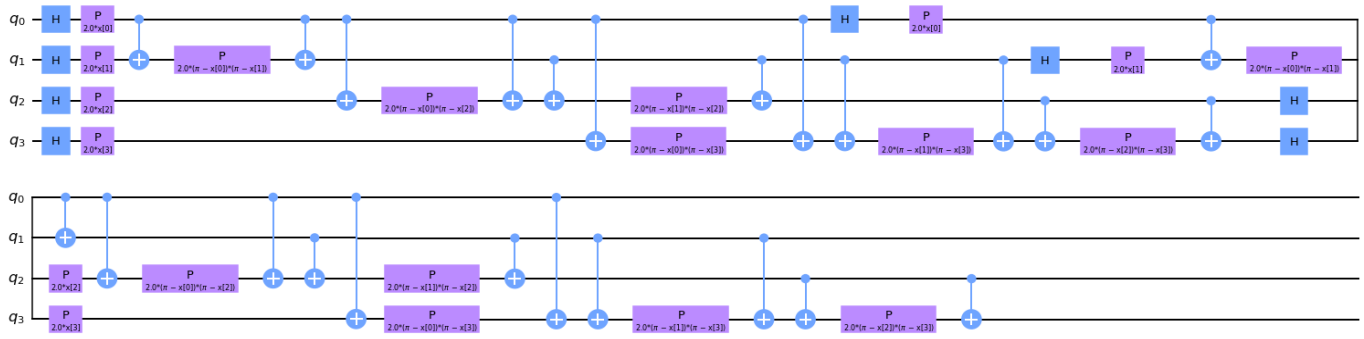


Figure 7: QSVM Featuremaps depths (2).

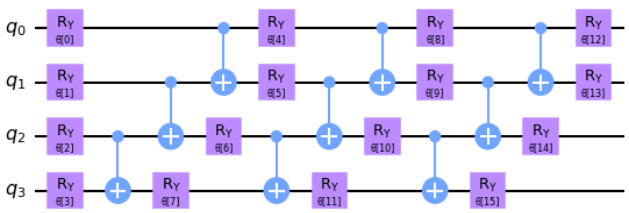


Figure 8: QSVM RealAmplitudes Variational circuit depths (3).

Figure 9: QNN Featuremaps and RealAmplitudes Variational

This study's experimental design is depicted in Fig. 7. The challenge of training very complex machine learning models on large data sets is one reason for utilizing quantum machine learning.

3.1. Data preparing and normalization

We shuffle the data to ensure randomness, remove less relevant features, and normalize the information between the ranges of 0 and 2π, and 0 and 1 to properly use the Hilbert space. The data is divided into a training set for model building and a testset for model testing, with the testset size being kept at 30% of the total dataset. This is a common practice in traditional machine learning such as neural networks and support vector machines.

3.2. Data encoding

Data encoding or state preparation in quantum feature mapping is similar to a classical feature map in that it helps translate data into a different space. In the case of quantum feature mapping, the data is translated into quantum states to be input into an algorithm. The result is a quantum circuit where the parameters depend on the input data, which in our case is the classical weather behind the dam.

It's worth noting that variational quantum circuits are unique in that their parameters can be optimized using classical methods. We utilized two types of feature maps pre-coded in the Qiskit circuit library, namely the ZZFeaturemap and PauliFeaturemap. To evaluate the performance of different models [27, 28], we varied the depths of these feature maps.

3.3. Variational quantum circuit

The model circuit is constructed using gates that evolve the input state. It is based on unitary operations and depends on external parameters that can be adjusted. Given a prepared state, $|\psi_i\rangle$, the model circuit *U(w) maps* $|\psi_i\rangle$ it to another vector,

$$|\psi_i\rangle = U(w)\ |\psi_i\rangle.$$

U(w) is comprised of a series of unitary gates.

4. Results & Discussion

In this research, we make use of the ZZFeaturemap and PauliFeaturemap precoded featuremaps from the Qiskit circuit library. To test the effectiveness of the various models, we changed the featuremaps depths (2). We incorporate more entangle-ment into the model and repeat the encoding circuit by increasing the depth of a feature map. After we used our feature map, a classifier may locate a hyperplane to divide the input data, and a quantum computer can evaluate the data in this feature space as Fig.7. Then we utilized the RealAmplitudes variational circuit from Qiskit. By increasing the depth of the variational circuit, more trainable parameters are introduced into the model that show in Fig. 8. The variational Featuremaps and RealAmplitudes reduced form was applied to write the QNN in Fig. 9. In order to determine the experimental target value in each cycle, the objective function value per iteration of the test was shown in Fig. 10, i.e. QSVM gave less objective value than QNN and VQC in Fig. 11 and Fig. 12, respectively.

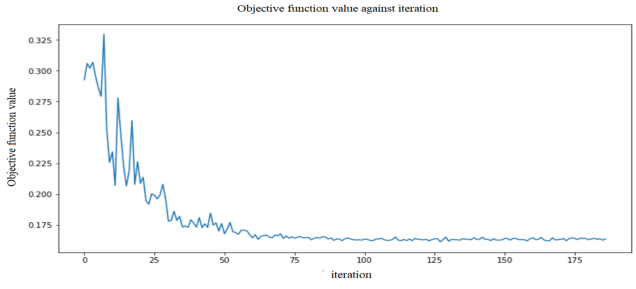


Figure 10: QSVM opjective function value per iteration.

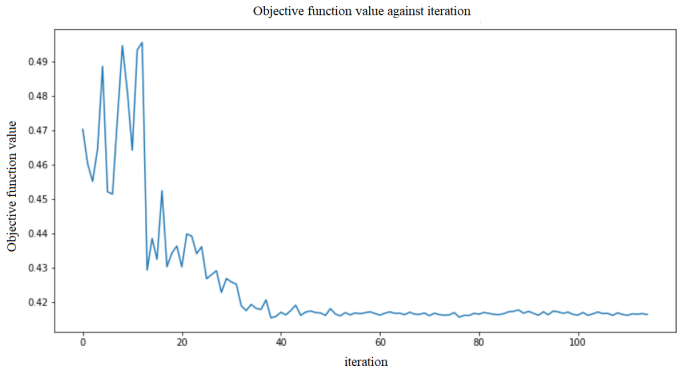


Figure 11: QNN opjective function value per iteration.

Table 4 presents the performance of our models: QSVM, QNN, and VQC. The QSVM obtained an accuracy of 85.3%, while the quantum models QNN and VQC recorded 52.1% and 70.1% accuracy, respectively. The ZZFeaturemap encoding with RealAmplitudes technique was implemented on the model using the weather dataset, with a depth of 3 layers and 300 epochs. The validation accuracy achieved is depicted in Figure 8. Despite the use of three separate attention processes in conjunction with the VQC model, the results of this investigation were satisfactory.

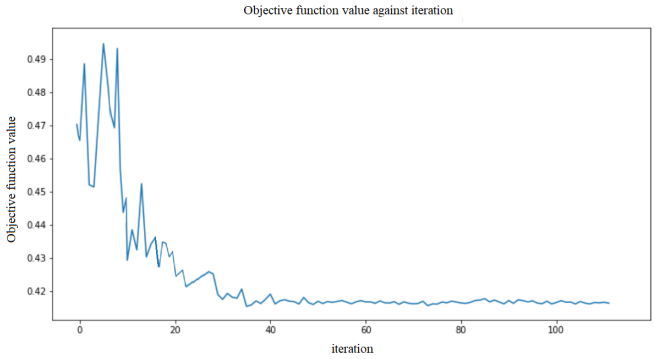


Figure 12: VQC opjective function value per iteration.

Table 4. Classifier test score

Classifier	Score
QSVM	0.853
QNN	0.521
VQC	0.701

5. Conclusion

In this article, we implemented three quantum models using RealAmplitudes techniques. We used ZZFeaturemap encoding as an evaluation optimization, but we acknowledge that this should not be the only optimization used to improve a quantum framework. Furthermore, state preparation is just one aspect of QML algorithms to benefit from when implemented into quantum machine learning. We suggested a pre-processing approach to improve the quantum state preparation for VQC. Our results showed achieved efficiencies of 85.3%, 52.1%, and 70.1%. According to our findings, the QSVM optimizer had the best performance, followed by VQC and QNN. We used ZZFeatureMap with a depth of two and the RealAmplitudes variational form with a depth of three. Moving forward, we plan to explore the use of different data encoding techniques such as RealAmplitudes, amplitude encoding, angle encoding, or other encoding methods to enhance the QML models and increase the number of features to improve performance relative to the established models and cutting-edge techniques. The study was based on a relatively small data set. Therefore, it may influence the assessment of model effectiveness and not discuss data pre-processing techniques because we are primarily interested in the efficiency of quantum models.

Abbreviation

QML	Quantum Machine Learning
QSVM	Quantum Support Vector Machine
QNN	Quantum Neural Networks
VQC	Variational Quantum Classifier
SDK	Software Development Kit
HQA	Hybrid Quantum Autoencoder
QAE	Quantum Amplitude Estimation
PQC	Parameterized Quantum Circuits

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgements

The authors would like to acknowledge the support from the Computer Engineering Program, College of Innovative Technology and Engineering Dhurakij Pundit University Bangkok, and the support for the information data from the Meteorological Department Thailand.

References

[1] J.C. Christopher, Tutorial on Support Vector Machines for Pattern Recognition Data Mining and Knowledge Discovery **2**, 121–167 1998.
[2] C. Khemapatapan, "Service Oriented Classifying of SMS Message", 2011 Eighth International Joint Conference on Computer Science and Software Engineering (JCSSE), May 2011, Thailand, 101-106.
[3] C. Khemapatapan, "2-Stage Soft Defending Scheme Against DDOS Attack Over SDN Based On NB and SVM", Proceeding of 8th International Conference from Scientific Computing to Computational Engineering, Jul 4-7, 2018, Athens Greece, 1-8, 2018.
[4] T. Thepsena, "Reservoir Release Forecasting by Artificial Neural Network at Pa Sak Jolasid Dam" International STEM Education Conference (iSTEM-Ed 2022), July 6-8, 2022
[5] T. Thepsena et al., " Rainfall Prediction over Pasak Jolasid Dam By Machine Learning Techniques " National Conference on Wellness Management: Tourism, Technology, and Community (H.E.A.T Congress 2022), August 18-20, 2022
[6] C. Khemapatapan, "A Classifiers Experimentation with Quantum Machine Lerning" The 2023 International Electrical Engineering Congress (iEECON2023) 2023,
[7] V. Heyraud, Z. Li, Z. Denis, A. Le Boité, and Cristiano Ciuti, "Noisy quantum kernel machines.", Phys. Rev. A 106, 052421 – Published 18 November 2022. DOI 10.1103/PhysRevA.106.052421
[8] S. Omar et al.; "Quantum kernels for electronic health records classification.", APS March Meeting 2022, abstract id.S37.006
[9] W. Li, Z. Lu and D. Deng1,Quantum neural network classifiers: A tutoria, SciPost Phys. Lect.Notes 61 (2022).
[10] S. Laokondee, P. Chongstitvatana, Quantum Neural Network model for Token alloocation for Course Bidding, Computer Science, Physics 2021(ICSEC).

[11] Elham Torabian, Roman V. Krems,"Optimal quantum kernels for small data classification.", Quantum Physics[Submitted on 25 Mar 2022] 14
[12] S. Aaronson and A. Ambainis, Forrelation: “A problem that optimally separates quantum from classical computing,”, SIAMJ. Comput. **47**, 982 2018.
[13] L.Zhou, S.T.Wang, S.Choi, H. Pichler and M.D. Lukin, “Quantum Approximate Optimization Algorithm: Performance, Mechanism, and Implementation on near term device, ” Physical Review X, vol.10, June 2020.
[14] E. Farhi, S. Gutmann and J. Goldstone, “A quantum approximate optimization algorithm,”, Nov 2014
[15] S. Nath Pushpak, S. Jain, “An Introduction to Quantum Machine Learning Techniques”, 2021 9th Interational conference on Reliability, Infocom Technologies and Optimization, Amity University, Noida, India, 2021
[16] Valentin Heyraud, Zejian Li, Zakari Denis, Alexandre Le Boité, and Cristiano Ciuti, "Noisy quantum kernel machines.", Phys. Rev. A 106, 052421 – Published 18 November 2022, DOI: 10.1103/PhysRevA 106.052421
[17] Shehab, Omar ; Krunic, Zoran ; Floether, Frederik ; Seegan, George ; Earnest-Noble, Nate, "Quantum kernels for electronic health records classification.", APS March Meeting 2022, abstract id.S37.006 DOI:10.1109/TQE.2022.3176806
[18] W. Li, Z. Lu and D. Deng1,”Quantum neural network classifiers: A tutoria”, SciPost Phys. Lect.Notes 61 2022, DOI: 10.21468/SciPostPhysLectNote.61
[19] S.Aaronson and A. Ambainis, Forrelation: “A problem that optimally separates quantum from classical computing,” , SIAMJ. Comput. **47**, 982 (2018), DOI:10.1137/15M1050902
[20] L.Zhou, S.T.Wang, S.Choi, H. Pichler and M.D. Lukin, “Quantum Approximate Optimization Algorithm: Performance, Mechanism, and Implementation on near term device, ” Physical Review X, vol.10, June 2020, DOI:10.1103/PhysRevX.10.021067
[21] E. Farhi, S. Gutmann and J. Goldstone, “A quantum approximate optimization algorithm,”, Nov 2014, DOI10.48550/arXiv.1411.4028
[22] Maria Schuld and Nathan Killoran “Quantum Machine Learning in Feature Hilbert Spaces.”, Phys. Rev. Lett. 122, 040504 – Published 1 February 2019, DOI:10.1103/PhysRevLett.122.040504
[23] Vojtech Havlícek, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow & Jay M. Gambetta, “Supervised learning with quantum-enhanced feature spaces.”, Nature **567**, 209–212, 2019, DOI:10.1038/s41586-019-0980-2
[24] M. L. LaBorde, A. C. Rogers, J. P. Dowling, Finding broken gates in quantum circuits: exploiting hybrid machine learning, Quantum Information Processing 19 8,aug 2020, DOI:10.1007/s11128-020-02729-y
[25] S. L. Wu, S. Sun, W. Guan, C. Zhou, J. Chan, C. L. Cheng, T. Pham, Y. Qian, A. Z. Wang, R. Zhang, et al. "Application of quantum machine learning using the quantum kernel algorithm on high energy physics analysis at the lhc 2021, DOI: 10.1103/PhysRevResearch.3.033221
[26] A. Chalumuri, R. Kune, B. S. Manoj, A hybrid classical-quantum approach for multi-class classification, Quantum Information Processing 20, 3 mar 2021, DOI:10.1007/s11128-021-03029-9
[27] G. Brassard, P. Høyer, M. Mosca, A. Tapp, Quantum amplitude amplification and estimation, Quantum Computation and Information 2002, P.53–74, DOI: 10.1090/conm/305/05215
[28] M. Danyai, Daniel S., Begonya G. " Variational Quantum Classifier for Binary Classification: Real vs Synthetic Dataset." ieee Access. DOI: 10.1109/Access.2021.3139323

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 60-65 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN:

Proportional Derivative and Proportional Integral Derivative Controllers for Frequency Support of a Wind Turbine Generator in a Diesel Generation Mix

Abdul Ahad Jhumka[1,2], Robert Tat Fung Ah King[*,1], Chandana Ramasawmy[2], Abdel Khoodaruth[3]

[1]*Department of Electrical and Electronic Engineering, University of Mauritius, Reduit 80837, Mauritius*

[2]*Advanced Mechanical and Electrical Services Ltd., Rose-Hill 71364, Mauritius*

[3]*Department of Mechanical and Production Engineering, University of Mauritius, Reduit, 80837, Mauritius*

ARTICLE INFO

Article history:
Received: 28 February, 2023
Accepted: 28 May, 2023
Online: 25 July, 2023

Keywords:
Wind Turbine Generator
Stability
Proportional Derivative Controller
Proportional Integral Derivative Controller
Inertia

ABSTRACT

The levelized cost of electricity production is highly dependent on the cost of fuel oil on the world market. In order to reduce the dependency on the fuel oil, many countries are adopting an energy transition towards distributed generation. Distributed generation can be described as various means of generating electricity at or near where it will be used. Such generating mode can be a solar PV system, wind turbine generator and other renewable energy sources. However, it entails lots of challenges as it uses power electronics devices as the power grid interface, which causes a reduction in the system inertia and at the same time affecting the frequency, thereby affecting the stability. To enhance this stability, appropriate control measures need to be adopted. This paper brings forward a novel approach for frequency control support of a wind turbine generator (WTG) in a diesel generation mix. The novelty of this research paper explained on the concurrent application of a Proportional derivative (PD) and a Proportional Integral Derivative (PID) for speed and frequency control in a WTG. The analysis of this experimental research was carried out through the modelling of the rate of change of frequency (RoCoF) using MATLAB / Simulink software. The results showed that the use of these controllers in presence of WTG provide frequency support to the system as the frequency varied within the acceptable limit of ±0.5Hz. Additionally, this experimental research work also proved that the use of speed / governor control in form of the PID improved the RoCoF and provided an enhancement in the stability of the test system. Finally, this paper confirmed that the integration of WTG to the grid required the use of appropriate control algorithm for an efficient exploitation of this kind of renewable energy source.

1. Introduction

COP 26 conference on the global climate established the importance on the reduction of the global temperature increase to 1.5°C as a mitigating action against the greenhouse gases [1]. In line with this agreement, many countries are adopting a distributed generation policy as means of reduction of greenhouses gases. Distributed generation (DG) covers the whole spectrum of different power generating technology such as solar PV, wind energy, biomass, etc. [2]. According to [3], the exploitation of distributed generation system is emerging in the global energy market. A direct impact of this alternative form of power generation results in a lower cost of power production from renewable energy sources. In this respect, countries such as Germany and Denmark are making enormous progress in promoting the distributed renewable energy system in their generation mix [4]. Globally, solar PV and wind energy power generation are the most preferred technologies coming out in the light, with wind energy considered as the leading renewable energy source [5]. The increasing use of wind energy in the generation mix brings along a shift towards using power electronics devices as grid interface. The power electronics interface devices have undergone rapid development with semiconductor switches such as insulated gate bipolar transistor (IGBT) are now being used. A direct impact of this transition will

*Corresponding Author: Robert Tat Fung Ah King, , r.ahking@uom.ac.mu

result in reducing the inertia of the grid, which plays a primordial role in the stability of the grid. It is expected that the total inertia of the National Grid in UK will be reduced by up to 70% by 2033/34 [6], [7]. To maintain a stable power grid with wind energy in the generation mix, it is required to avoid unnecessary frequency dip owing to the stochastic nature of the wind energy source. Therefore, the frequency stability of wind turbine provides a hot topic for research.

Previous research works carried on the subject show a replete of control mechanism on the frequency support. In [8], the authors introduced a low order system frequency model with high penetration of wind power plant. This method studies the power system frequency changes during the most critical time, which is ≤ 30 s [9]. It was observed that initial conditions do not have a significant impact on the frequency response. However, the research paper in [8] highlighted on some major limitations of this experiment with regards to the effect of variable speed on the wind turbine generation system (WTGS). In [10], a comparative analysis between wind turbine generation (WTG) and a solar PV system was established. It was observed that wind turbine (WT) requires an extensive control mechanism to be able to provide a stable power as stability was not attained within the first swing. To further substantiate on the research for frequency control support for WT, an efficient control of inertia emulation and frequency support in presence of WTG was proposed in [11]. The experiment proposed a model free control (MFC) strategy for inertia emulation and frequency support of a diesel wind grid system. The MFC employs an approximation-based intelligent proportional integral derivative (PID) controller experiment. The online estimation technique [12] from input-output measurements is a key concept of MFC to approximate the complex system. It was concluded from this experiment that the use of MFC provided a precise inertia emulation and necessary frequency support. Other works on inertia emulation were carried out by [13]-[16].

In [17], the authors adopted a novel approach of using small signal analysis for frequency response of WT. The method depicts a Klein Rogers Kundur (KRK) two area, four generator system. The model was modified to accommodate an additional generator G5 on bus 13. It was concluded from this experiment that integration of wind energy without any frequency control will deteriorate the frequency response of the system. A proper control of the WTG provides necessary frequency support of the system. In [18], the feasibility of using a double fed induction generator (DFIG) to implement frequency regulation was investigated. It was observed that adjustable frequency wind turbines can undertake the frequency regulation responsibility of the power grid. A coordinated primary frequency regulation was considered as essential between the diesel generator (DG) and the WTG. In [19], an assessment of the impact of wind generation on system frequency control was made, where a time series sampling methodology was proposed over a timeframe period for assessing the impact of increased penetration of wind energy. It was concluded that future power system with an increasing penetration level of DFIG and greater levels of High Voltage DC (HVDC) interconnection will present significant frequency control challenges to system operators.

The above expose clearly shows that integrating wind energy to the grid is indeed very challenging due to a reduction of system inertia. It gives a broader picture of the research carried out in the development of the frequency support. Inspired by these obstacles, the rationale of this work was to develop a novel approach of frequency support algorithm, through the concurrent use of a Proportional Derivative (PD) controller and Proportional Integral Derivative (PID) controller.

The remaining of this research paper is structured as follows. Section 2 describes an overview of the system frequency response mechanism, while the methodology is detailed in Section 3. Section 4 showcases the simulations and results based on the principle laid down in Section 3. Section 5 treats about the discussion of the results obtained, whereas Section 6 concludes this research paper.

2. Overview

This section explains the vital role played by the frequency to maintain a stable power with the grid. As per [20], frequency stability is the ability of the power system to maintain a steady frequency following a transient occurrence, which leads to a power mismatch between generation and load. Therefore, it is required to avoid large rate of change of frequency (RoCoF) through application of necessary frequency control. The RoCoF is one of the indicators, which gives the soundness of the system frequency response (SFR).

A frequency control can act in three level steps namely [21]

1. Primary Response
2. Secondary Response
3. Tertiary Response.

Figure 1 shows the frequency response steps in the event of a loss of generation.

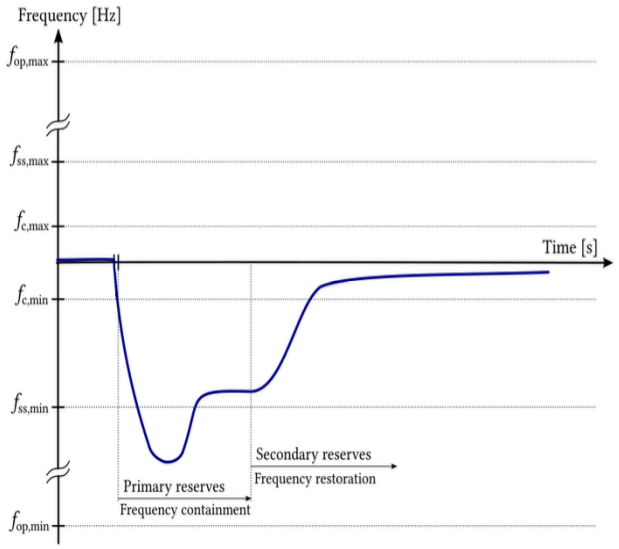


Figure 1: Frequency response in the event of a loss of generation

Another indicator, which can evaluate the system frequency response (SFR) is the frequency nadir. The frequency nadir is the minimum frequency reached during a transient period. It is therefore primordial to understand the role of frequency in maintaining a sound network particularly in presence of

renewable energy sources like WTG. The next section explains the relation between the power imbalance brought in by the frequency disturbance and the system inertia. It also shows that the RoCoF is dependent on the speed of the generator. Hence the importance of having a governor and speed control in the integration of a wind farm.

3. Methodology

Stability of power system is deeply impacted by large frequency deviation. In order to understand the impact of the frequency deviation on the power system, it is required to understand the swing equation, which gives the variation of the system inertia (H) with respect to the change in power ΔP. Equation (1) gives the swing equation.

$$\Delta P = P_m - P_e = \frac{2H}{\omega_s}\frac{d^2\delta}{dt^2} \tag{1}$$

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)

According to [22], the acceleration of the prime mover caused by the unbalanced torque is governed by the equation of motion as in (2)

$$J\frac{d\omega_m}{dt} = T_a = T_m - T_e \tag{2}$$

where

J: Combined moment of inertia of generator and turbine (kg/m^2)

ω_n: Angular velocity of the rotor (mech. rad/s)

t: Time (s)

Given that kinetic energy of the rotating masses is represented by

$$E_k = \frac{1}{2}J{\omega_m}^2 \tag{3}$$

And that power is the rate of change of energy

$$\Delta P = \frac{dE}{dt} = J\frac{d\omega_m}{dt} \tag{4}$$

As

$$H = \frac{\frac{1}{2}J{\omega_m}^2}{VA_{base}} \tag{5}$$

Substituting for J in (5) into (4)

$$\Delta P = \frac{dE}{dt} = \frac{2HxVA_{base}}{{\omega_m}^2} \text{ x } (d\omega_m)/\text{ dt.} \tag{6}$$

As ω = 2πf

$$\Delta P = \frac{dE}{dt} = \frac{2HxVA_{base}}{{\omega_m}^2} \text{ x } 2\pi \text{ (df/dt).} \tag{7}$$

Hence based on (7), the power imbalance brought in by the nature of wind directly impact on the system inertia (H) and the RoCoF (df/dt). Therefore, it is required to have a low RoCoF and additional synthetic inertia to counter the change in power.

As per [23], the maximum power that can be harnessed from a WTG is governed by (8) below

$$P = 0.5\rho AV^3 \tag{8}$$

where

P: Power reaped from the WTG (W)

ρ: Density of air (kg/m^3)

A: Area of blade (m^2)

V: Velocity of wind (m/s)

Equating (8) and (4), the following expression can be obtained

$$J\frac{d\omega}{dt} = 0.5\rho AV^3 \tag{9}$$

Since ω = 2πf

$$\frac{df}{dt} = \frac{1}{4\pi}\frac{\rho AV^3}{J} \tag{10}$$

Based on (10), the wind speed directly affects the rate of change of frequency, RoCoF. Therefore, in order to exploit maximum power from the WTG, it will be necessary to provide a stable power to the grid by developing a governor and inertia control.

4. Simulations & Results

Having established the mathematical models that govern the frequency stability and power quality of a WTG, it is required to

develop a power grid with integration of a WTG. A prototype grid in form of an IEEE 9-bus, with integration of a 2 MW WTG (G3), was used as a test case with a system frequency of 50 Hz. This is shown in Figure 2.

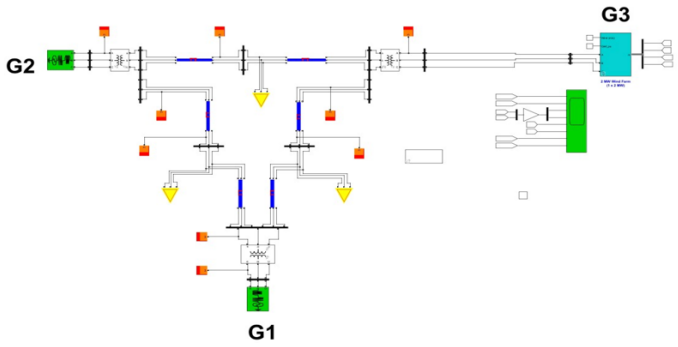


Figure 2: Integration of a Wind Turbine Generator in a generation mix

An initial experiment to assess the impact of integrating WTG, showed that large RoCoF occurred in the mix, which brought instability in the system. Figure 3 shows the unstable characteristics of the generation mix with presence of WTG.

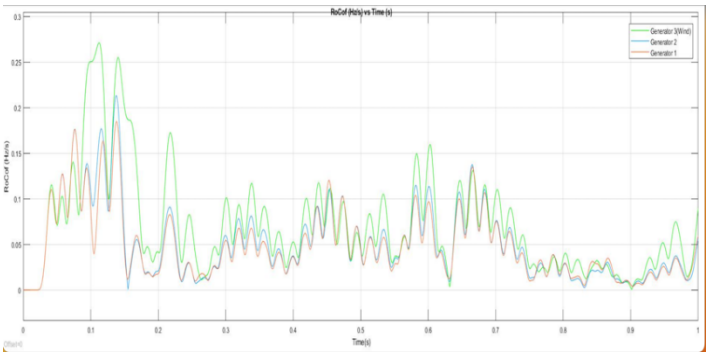


Figure 3: Large RoCoF in presence of WTG

In order to enhance the unstable nature of the system, Proportional Derivative (PD) and Proportional Integral Derivative (PID) controllers were connected in the WTG arrangement for rotor angle and speed control respectively. Such arrangement is shown in Figure 4.

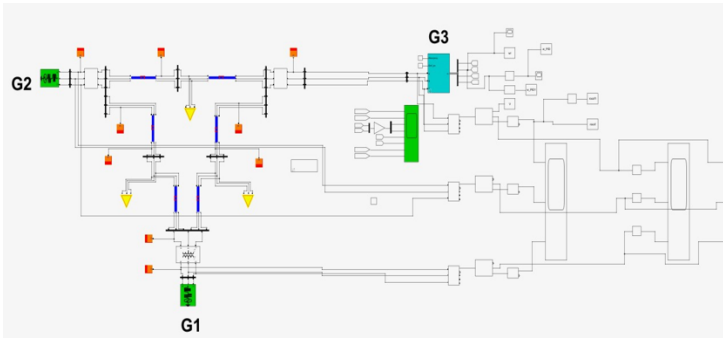


Figure 4: Arrangement of WTG with PD and PID Controllers

Proportional Derivative (PD) Controller

According to [24], a proportional derivative controller is a type of controller used where the output varies in proportion with the input signal. The block diagram of the PD controller used for the rotor angle controller is shown in Figure 5.

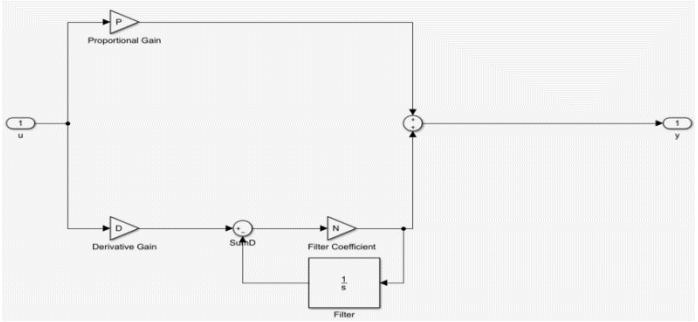


Figure 5: Block diagram of a PD controller used in a WTG

The output of the PD controller can be determined from the transfer function of the controller. The transfer function or gain of the controller gives the ratio of the output signal to the input signal. The gain of the PD controller can be calculated based on the following equation

$$G_s(s) = K_p(1 + T_d(s)). \quad (11)$$

where

K_p: Proportional gain

T_d: Time derivative

Proportional Integral Derivative (PID) Controller

A proportional integral and derivative controller (PID) is a type of controller that uses three different controller types namely the proportional, integral and derivative controllers to set the output function. A typical block diagram of a PID controller is shown in Figure 6.

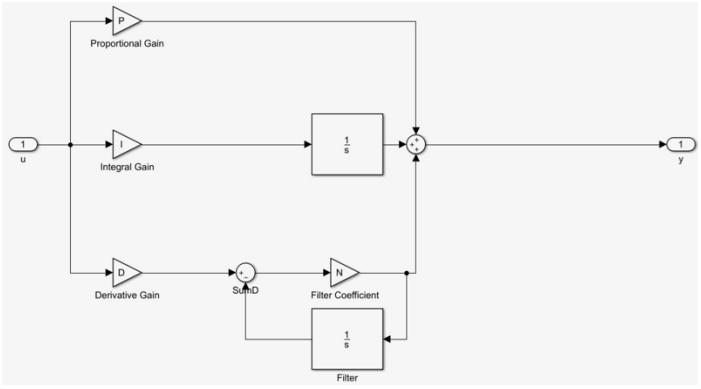


Figure 6: Block diagram of a PID controller

It works by calculating the error margin between the set point and the measured point. Similar to the PD controller, the transfer function of the PID controller will provide the desired output of the controller. The transfer function of the PID controller can be calculated as per (12).

$$G(s) = K_P (1+T_d(s) + 1/(T_d(s)). \quad (12)$$

where

K_p: Proportional gain

T_d: Time derivative

The impact of the PI and the PID controllers was assessed through the temporal variation of the RoCoF over a period of 1s. The result in Figure 7 demonstrates that the contribution of the controllers brought about a reduction in the RoCoF. Hence, an enhancement in the stability of the system.

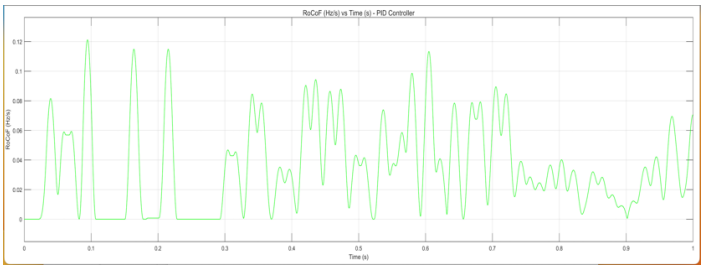


Figure 7: Temporal variation of RoCoF

Additionally, since a PID controller was used for the governor and speed control, the rate of change of frequency with respect to velocity was modelled. The contribution of the PID in the system brought a stabilization in the RoCoF. This is illustrated in Figure 8.

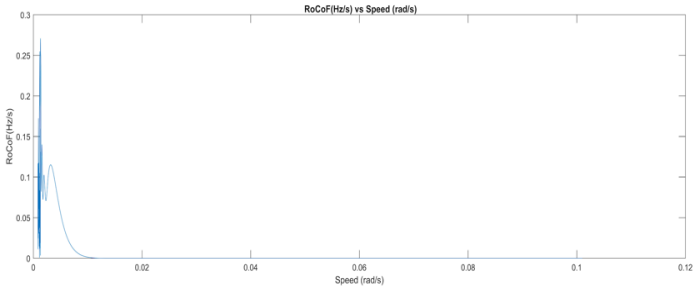


Figure 8: Variation of RoCoF with speed

5. Discussion

The objective of this research work was to develop a novel approach of using concurrently PI and PID controllers for frequency support of a WTG connected in a diesel generation mix. The test case developed in this paper showed that the controllers can provide frequency support to the system by reducing the RoCoF and stabilizing the frequency variation, within ±0.5 Hz as shown in Figure 9.

Additionally, this research work showed that this method proved to be more accurate than [25]. The virtual inertia control developed in [25] as a frequency support resulted in a frequency variation of ±0.8 Hz. Therefore, the concurrent use of PID and PD controllers may be the favoured way in enhancing the stability of a diesel and WTG generation mix.

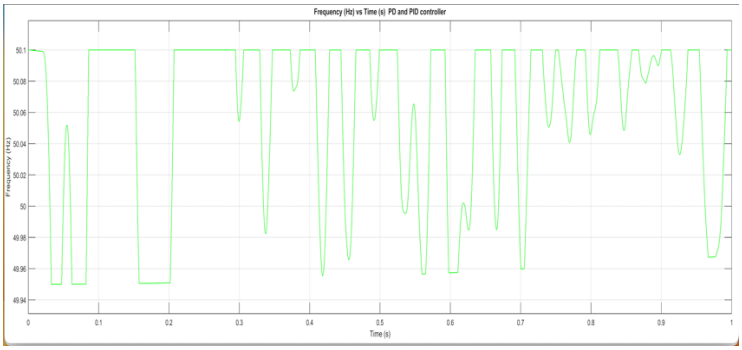


Figure 9: Temporal variation of frequency in presence of PI and PID Controllers

6. Conclusion

This paper discussed about the challenges of integrating a WTG to the grid. It also highlighted the role played by the frequency in arresting the extent of disturbance caused by the penetration of the WTG. Since a WTG is dependent on the external weather factor, any change in speed will result in a high rate of change of frequency, which will lead to power outage.

This research work has also demonstrated that the synchronization of a WTG to a grid is very challenging as it requires extensive control, in the form of PI and PID controllers. A combination of these control strategies constitutes the novelty in the frequency support, through the confirmation of the experimental results obtained. Therefore, it can be concluded that an extensive control concept is mandatory for an efficient exploitation of wind energy. However, the drawback in exploiting such type of renewable energy system will lie in the cost of investment.

This research work can be further extended to the analysis of multi-swing stability.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research work was assisted by Advanced Mechanical and Electrical Services Ltd. as part of the company's research policy.

References

[1] M. Denchak, "Paris Climate Agreement: Everything You Need to Know," NRDC, 2021, [Online], https://www.nrdc.org/stories/paris-climate-agreement-everything-you-need-know#sec-summary.

[2]. Environmental Protection Agency, "Distributed Generation of Electricity and its Environmental Impacts," https://www.epa.gov/energy/distributed-generation-electricity-and-its-environmental-impacts (Accessed: December 28, 2022).

[3]. REN21, "Renewable Energy Policy Network for the 21st Century," Renewables—Global Status Report; REN21 Secretariat: Paris, France, 2017.

[4] C.D. Iweh, S. Gyamfi, E. Tanyi, E. Effah-Donyina, "Distributed generation and renewable energy integration into the grid: Prerequisites, push factors, practical options, issues and merits," Energies, **14**(17), 5375, 2021 https://doi.org/10.3390/en14175375.

[5] H. Ibrahim, M. Ghandour, M. Dimitrova, A. Ilinca, J. Perron, "Integration of wind energy into electricity systems: Technical challenges and actual

solutions," Energy Procedia, **6**, 815–824, 2011, https://doi.org/10.1016/j.egypro.2011.05.092.
[6] National Grid, "UK Future Energy Scenarios," 2018, accessed: 2019- 05-15, [Online], http://fes.nationalgrid.com/media/1363/ fes- interactive- version-final.pdf
[7] E. Rakhshani, D. Gusain, V. Sewdien, J. L. Rueda Torres, M. A. M. M. Van Der Meijden, "A key performance indicator to assess the frequency stability of wind generation dominated power system," IEEE Access, **7**, 130 957–130 969, 2019.
[8] M. Krpan, I. Kuzle, "Introducing low-order system frequency response modelling of a future power system with high penetration of wind power plants with frequency support capabilities," IET Renewable Power Generation, **12**(13), 1453–1461, 2018, https://doi.org/10.1049/iet-rpg.2017.0811.
[9] P.M. Anderson, M. Mirheydar, "A low-order system frequency response model," *IEEE Trans. Power Syst.*, **5**(3), 720–729, 1990, doi: 10.1109/59.65898.
[10] A.A. Jhumka, R.T.F. Ah King, A. Khoodaruth, C. Ramasawmy, "Comparative Performance Analysis of Solar Energy and Wind Energy Systems using Rotor Angle Stability," 2022 IEEE 7th International Energy Conference (ENERGYCON), Riga, Latvia, 2022, 1-5, doi: 10.1109/ENERGYCON53164.2022.9830452.
[11] U. Datta, J. Shi, A. Kalam, "Primary frequency control of a microgrid with integrated dynamic sectional droop and fuzzy based pitch angle control," International Journal of Electrical Power & Energy Systems, **111**, 248–259, 2019, https://doi.org/10.1016/j.ijepes.2019.04.001.
[12] M. Fliess, C. Join, H, Sira-Ramirez, "Non-linear estimation is easy", Int. J. Modell. Identification Control, **4**(1), 12-27, 2008, https://doi.org/10.1504/IJMIC.2008.020996.
[13] J Morren, S.W. Haan, L. Kling, J. Ferreira, "Wind turbines emulating inertia and supporting primary frequency control," IEEE Trans. Power Syst. **21**(1), 433-434, 2006, doi: 10.1109/TPWRS.2005.861956.
[14] J.M. Mauricio, A. Marano, A. Gomez-Exposito, J.L. Martinez-Ramos, "Frequency regulation contribution through variable-speed wind energy conversion systems," IEEE Trans. Power Syst. **24**(1), 173-180, 2009, doi: 10.1109/TPWRS.2008.2009398.
[15] B. Wang, Y. Zhang, K. Sun, K. Tomsovic, "Quantifying the synthetic inertia and load-damping effect of a converter-interfaced power source," Proc. IEEE Int. Energy Conf. (ENERGYCON), Limassol, Cyprus, 1-6, 2018, doi: 10.1109/ENERGYCON.2018.8398838.
[16] S. Wang, K. Tomsovic, "A novel active power control framework for wind turbine generators to improve frequency response," IEEE Trans. Power Syst. **33**(6), 6579-6589, 2018, doi: 10.1109/TPWRS.2018.2829748.
[17] B. Park, Y. Zhang, M. Olama, T. Kuruganti, "Model-free control for frequency response support in microgrids utilizing wind turbines," Electric Power Systems Research, **194**, 107080, 2021, https://doi.org/10.1016/j.epsr.2021.107080.
[18] C. Guo, D. Wang, "Frequency regulation and coordinated control for Complex Wind Power Systems," Complexity, 1–12, 2019, https://doi.org/10.1155/2019/8525397.
[19] R. Doherty, A. Mullane, G. Nolan, D. J. Burke, A. Bryson, M. O'Malley, "An assessment of the impact of wind generation on system frequency control," IEEE Transactions on Power Systems, **25**(1), 452–460, 2010, doi: 10.1109/TPWRS.2009.2030348.
[20] P. Kundur, J. Paserba, V. Ajjarapu, G. Andersson, A. Bose, C. Canizares, N. Hatziargyriou, D. Hill, A. Stankovic, C. Taylor, T. Van Cutsem, V. Vittal, "Definition and Classification of Power System Stability," IEEE Trans. Power Syst., **19**(3), 1387–1401, 2004, doi: 10.1109/TPWRS.2004.825981.
[21] P. Fernández-Bustamante, O. Barambones, I. Calvo, C. Napole, M. Derbeli, "Provision of frequency response from Wind Farms: A Review," Energies, **14**(20), 6689, 2021, https://doi.org/10.3390/en14206689.
[22] P. Kundur, N.J. Balu, M.G. Lauby, Power system stability and control, McGraw-Hill. 1994.
[23] J.F. Manwell, J.G. McGowan, A.L. Rogers, Wind Energy Explained: Theory, Design and Application, Wiley: Hoboken, NJ, USA, 2010.
[24] S. Bashetty, J.I. Guillamon, S.S. Mutnuri, S. Ozcelik, "Design of a robust adaptive controller for the pitch and torque control of wind turbines," Energies, **13**(5), 1195, 2020, https://doi.org/10.3390/en13051195.
[25] D. Yang, E. Jin, J. You, L. Hua, "Dynamic Frequency Support from a DFIG-Based Wind Turbine Generator via Virtual Inertia Control," Appl. Sci., **10**(10), 3376, 2020, https://doi.org/10.3390/app10103376.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 2, 64-70 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Omni-directional Multi-view Image Measurement System in the Co-sphere Framework

Yung-Hsiang Chen[1,2], Jin H. Huang[*,3]

[1]*Ph.D. Program of Mechanical and Aeronautical Engineering, Feng Chia University, Taichung, 407102, Taiwan*

[2]*Aeronautical Systems Research Division, National Chung-Shan Institute of Science and Technology, Taichung, 407102, Taiwan*

[3]*Department of Mechanical Engineering, Feng Chia University, Taichung, 407102, Taiwan*

ARTICLE INFO

Article history:
Received: 26 December, 2022
Accepted: 22 February, 2023
Online: 11 March, 2023

Keywords:
3D Reconstruction
Camera Calibration
Multi-view Image
Stereo camera

ABSTRACT

This study presents an "Omnidirectional multi-view image measurement system", which can be used to provide multi-camera 3D reconstruction and multi-view image information. Its characteristic is that four cameras take images from multiple perspectives in the co-sphere framework. The C_0 is the middle camera fixed as the geometric center point of measurement, and provides a front image. The other three cameras C_1~C_3 provide side images, and the co-circular spheres are separated by 120 degrees to extend the circle. The arc rod adjusts the multi-angle imaging. Place the multi-view camera in the arc track and move to the specified position in the sphere to position and capture images. By changing the angle between the cameras, the range of images captured by the cameras can be changed. If the multi-view images of four cameras C_0, C_1, C_2 and C_3 are captured at the same time, a stereo camera pair can be formed by any two cameras. The stereo camera pair C_0-C_1, C_0-C_2 and C_0-C_3 can be compiled by using the parallax principle of left and right images matching. Finally, through the demonstration and verification of camera calibration and 3D reconstruction, it can be used for all-round multi-view image measurement.

1. Introduction

The optical measurement method has the characteristics of global, non-contact, non-destructive, high measurement accuracy and real-time measurement, and is a very important part in the field of experimental mechanics. Among them, the advantage of non-contact is that it will not cause damage to the object to be measured, and high precision and real-time measurement are very important measurement characteristics for the industry that pursues light, thin and small for research and development. Therefore, combining these advantages, Optical metrology has gradually become one of the current research focuses. The advantage of non-contact is that it will not cause damage to the object to be measured, and high precision and real-time measurement are very important. Combined these advantages, the optical quantity Measurement has gradually become one of the current research focuses. There are many optical measurement methods, such as Electronic Speckle Pattern Interferometry (ESPI), Photoelastic Method, Shadow Moiré Method and Digital Image Correlation (DIC).

When using these optical measurement methods, different cameras are usually used together with image processing technology for research and analysis. There are many stereo vision systems for 3D reconstruction applications [2-6]. For example, in [7], the author presented the 3D-DIC method with stereo vision. The combination of DIC method and 3D stereo vision, that can be applied to 3D deformation measurement and becomes the basic framework of 3D-DIC method. The 3D-DIC method is applied to the surface profile measurement research. Whether it is applied to 3D-DIC measurement or other non-contact image measurement systems, the camera structure is mostly composed of two cameras connected by a horizontal bracket. The distance and angle between them make the images overlap within a certain range. When the measurement range exceeds the visible range of the two image capture devices, only partial images of the object can be captured. The occlusion area is the camera cannot directly observe the measurement object. The camera posture must be adjusted according to the different shapes of the object, and the image capture device must be re-installed. In camera calibration

[*] Corresponding Author: Jin H. Huang, Department of Mechanical Engineering, Feng Chia University, Taichung, Taiwan, email: jhhuang@fcu.edu.tw

" This paper is an extension of work originally presented in 2022 IEEE International Conference on Consumer Electronics - Taiwan [1]"

procedure, the range of camera angle of view and the need for continuous calibration of the adjusted camera are application constraints. It is necessary to simplify and repeat the definition of the relationship between the coordinates of the calibration measurement system and the object to be measured.

In order to solve the visible range of the two image problems, some scholars proposed a dual-axis parallel motion mechanism with two degrees of freedom of vertical motion and horizontal motion [8]. The device uses a servo motor as a positioning control, and is mounted on a camera device, which can precisely control the camera's imaging angle. Each servo motor of the X and Y axes operates independently, and will not become the load of another servo motor for image capture and analysis. Those three cameras to build a co-circular geometric multi-camera imaging platform system [9], connect the three cameras in series with a semi-circular measuring rod, make the optical axes of the multi-cameras co-intersect at a point in space, and the distance between the cameras. The mechanism can be used to adjust the spacing along the semi-circular measuring rod. With the superimposition feature of multiple cameras, the system can increase the range of overlapping areas of 3D reconstruction feature points and the correction parameters established inside, and simplify the complex correction procedures required for on-site measurement and reconstruction of 3D information, expanding the original 3D-DIC observation. The scope of field of view, and simplify the procedure of repeatedly defining and determining the relationship between the coordinates of the calibration measurement system and the object to be measured. According to the application of displacement and strain measurement, when observing objects of different sizes, as long as the camera distance is adjusted, it can be applied to reconstruct the three-dimensional space information of the object, and explore and increase the image data of the three-dimensional scene object in the overlapping area.

2. Research methods

The development of image measurement is bound to develop towards global, large-scale and rapid measurement. In order to meet practical purposes, multi-view images cannot be measured by only one or two cameras. It is necessary to develop an all-round multi-view image measurement system. The development of a multi-view large-scale measurement system is a very important topic. Based on this, this study proposes omnidirectional multi-camera 3D reconstruction and provide multi-view images.

In the application of 3D vision, two images taken at different positions are generally used, and the relative depth of the entire scene can be reconstructed from the 2D images. Developed 3D reconstruction products for non-contact image measurement of structural deformation. It is a system that uses two cameras to form a stereo vision measurement entity structure that can be used to measure the 3D global surface of an object. This technology uses the characteristics of the object surface as a surface comparison target to obtain 3D reconstruction measurements. When the measurement range exceeds the visible range of the two cameras, only partial images can be observed, and the area covered by the object cannot be measured.

In order to overcome the limitation of image occlusion, the "Omni-directional Multi-view Image Measurement System" with the co-sphere to solve the problem of object occlusion. The intersecting area can be increased by multiple cameras for 3D reconstruction. Provides a camera imaging device, which is convenient for shooting multiple sets of images. This device includes three arc measuring rods, so that the multi-camera circular motion can be captured by changing the movement of the multi-camera. That adjust the angle of the camera through the rotating seat. It is characterized in that multiple cameras are fixed on the arc frame, and the observation angle of the cameras is adjusted so that it has the characteristic of a co-sphere.

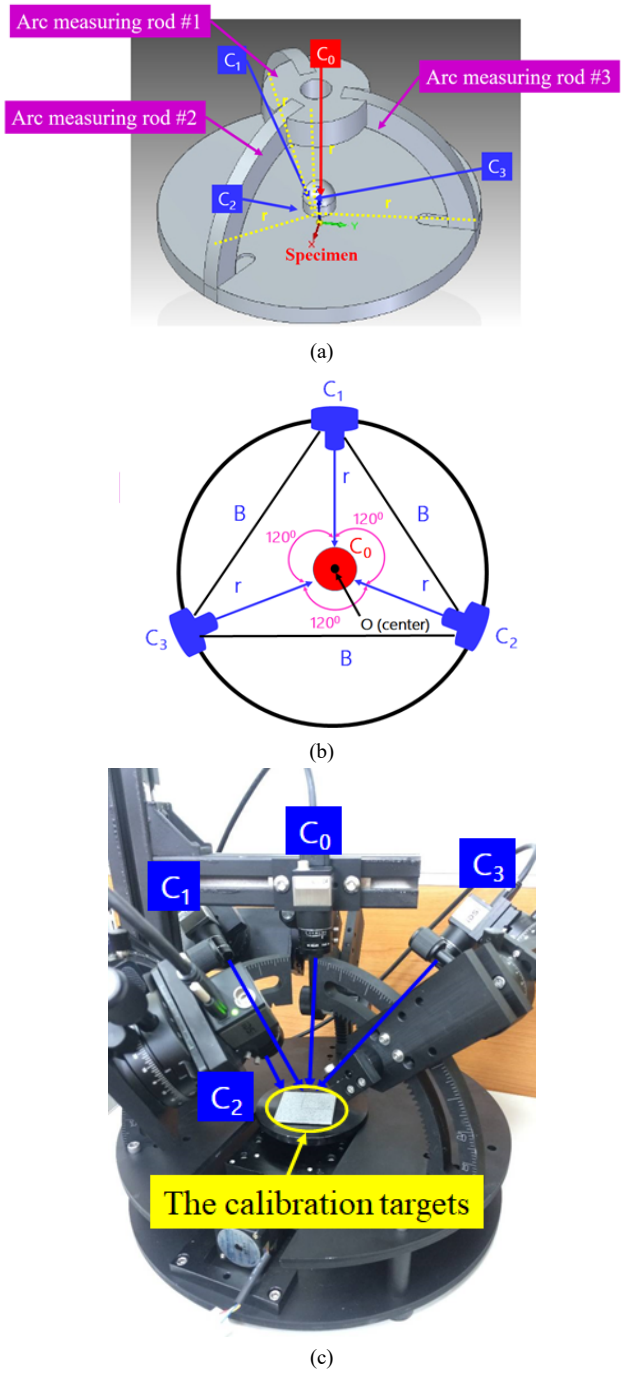
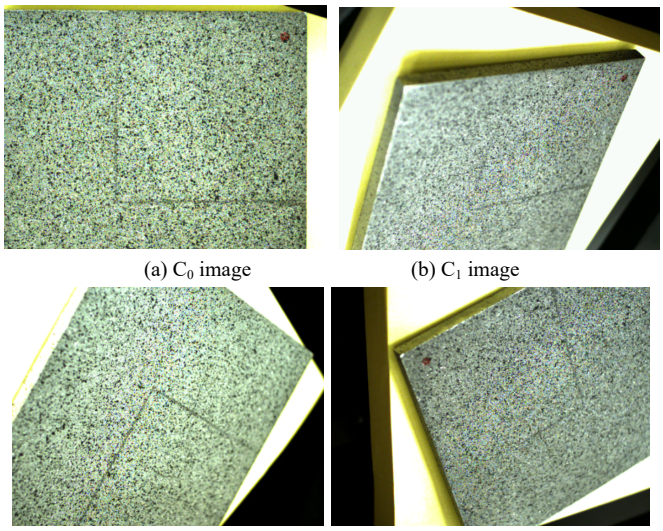


Figure 1: Omni-directional multi-view image measurement system

Figure 1 is the proposed omni-directional multi-view image measurement system. The C_0 is the middle camera fixed as the geometric center point of measurement, and provides a front image. The other three cameras C_1~C_3 provide side images, and the co-circular

spheres are separated by 120 degrees to extend the circle. The arc rod adjusts the multi-angle imaging. Place the multi-view camera in the arc track and move to the specified position in the sphere to position and capture images. By changing the angle between the cameras, the range of images captured by the cameras can be changed. If the multi-view images of four cameras C_0, C_1, C_2 and C_3 are captured at the same time, a stereo camera pair can be formed by any two cameras. The stereo camera pair C_0-C_1, C_0-C_2 and C_0-C_3 can be compiled by using the parallax principle of left and right images matching. Table 1 is omni-directional multi-view image measurement system specifications. Fig. 2 is four cameras captured the multi-angle images.



(a) C_0 image (b) C_1 image

(c) C_2 image (d) C_3 image

Figure 2: Four cameras captured the multi-angle images.

The internal camera calibration parameter is the camera mapping relationship between the 3D calibration target point and the 2D image point in the camera coordinate system. The external camera calibration parameter is the camera refer to the rotation and translation relationship between the world coordinate system, where the target point is located and the camera coordinate system. Camera calibration is the process of obtaining the internal and external parameters of the camera, that are defined the corresponding relationship between the three-dimensional coordinates and the image pixels through the calibration points marked on the calibration target.

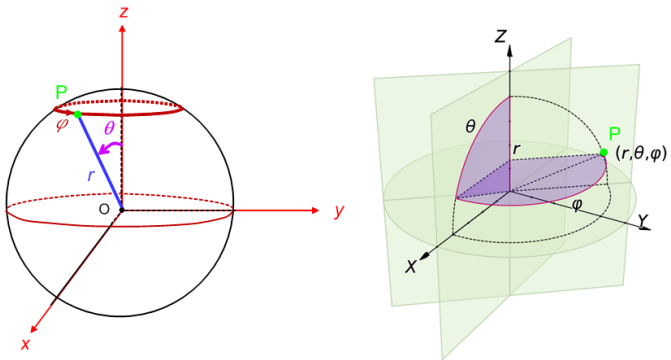
Table 1: Omni-directional multi-view image measurement system specifications

Specifications	Characteristic
Number of cameras	Four-camera multi-view image. The C_0 is the middle camera fixed as the geometric center point of measurement, and provides a front image. The other three cameras C_1~C_3 provide side images, and the co-circular spheres are separated by 120 degrees to extend the circle.
Camera	UI-3130CP Rev. 2 - IDS Imaging Development Systems GmbH, USB 3.0, CMOS, 575.0 fps, 800 × 600, 0.48 MPix, 1/3.6", ON Semiconductor, Global Shutter.
Lens	UH1220-10M, Focal Length: 12 mm, Aperture: f/2.0~C,Min. Working Distance: 10 cm, Distortion: <0.1%
Arc measuring rod	Three arc rods have a radius of 15 cm.
Measurable range	Measurable range: 3 cm × 3 cm × 3 cm (L×W×H), Minimum measurable resolution: 0.375 mm.
Dimension size	40 cm × 40 cm × 40 cm (L×W×H).
Power	110 V / 60 Hz.

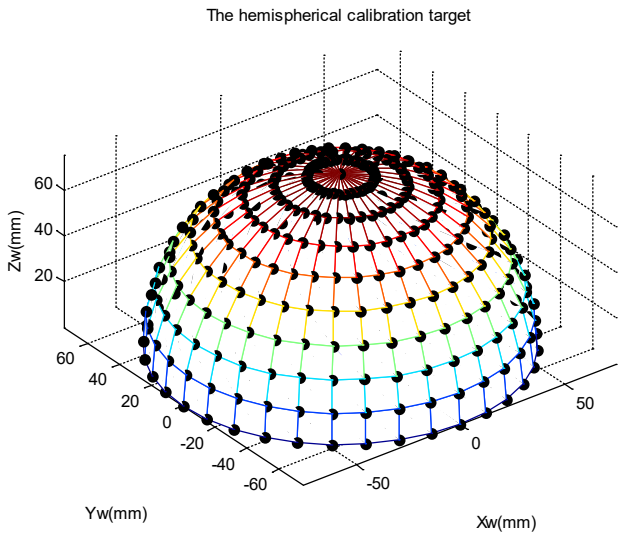
In [10], the author presented a two-step correction method, using radially consistent constraints, so that the correction of camera internal parameters is not affected by radial distortion. The method requires a high-cost three-dimensional high-precision calibration target. In [11], the author proposed the plane calibration method, which is uses image coordinate estimation to obtain a two-dimensional homography matrix. The internal and external parameters of the camera calibration and the radial distortion coefficient can be linearly solved. In [12], the author proposed a camera calibration method based on one-dimensional camera calibration points. This study in order to collect 360° omni-directional image information, the polar coordinate system is used. The advantage of polar coordinates is that it can describe the system directivity of the azimuth angle position of single/ multiple cameras relative to the calibration target on an image plane. In the polar coordinate system are represented by radius and angle, which are included cylindrical coordinate system and spherical coordinate system. Fig. 3 shows the spherical coordinate system and the hemispherical calibration model. A point P in the spherical coordinate system is defined by two angles φ, θ and radius r. The corresponding relationship between spherical coordinates (r,θ,φ) and rectangular coordinates (x,y,z) is as follows:

$$\begin{aligned} x &= r\sin\theta\cos\varphi \\ y &= r\sin\theta\sin\varphi \\ z &= r\cos\theta \end{aligned} \quad , \quad \begin{aligned} r &= \sqrt{x^2+y^2+z^2} \\ \theta &= \cos^{-1} z/r \\ \varphi &= \tan^{-1} y/x \end{aligned} \tag{1}$$

where, O is the origin point. r is the radius length coordinate value, that is the distance from point P to the coordinate center. φ is the coordinate value of the azimuth angle ($0\le\varphi\le2\pi$), that is the counterclockwise rotation angle from the X axis. θ is the elevation angle coordinate value ($0\le\theta\le\pi$), that is the upward rotation angle from the XY plane.



(a) spherical coordinate system



(c) hemispherical calibration model

Figure 3: The spherical coordinate system and the hemispherical calibration model

3. Experimental result

Fig. 4 is the calibration targets. Fig. 4(a) is a 30mm×30mm chess plane calibration target, which is composed of 9×9 square grids intersecting micro-dots with an equidistant distance of 5mm. Figure 4(b) is a hemispherical calibration target with a radius of 15mm and a height of 15mm, which is used to verify the feasibility of 3D reconstruction of the omnidirectional multi-view measurement system. Place the calibration piece in the middle of the measurement system, move the multi-camera in the arc track to the designated position in the sphere to position and take images, and change the range of images captured by the multi-camera by changing the angle between any two cameras .

(a) Chess-board (b) Hemispherical

Figure 4: The calibration targets

3.1. Omni-directional multi-camera calibration results

Calibrate each stereo pair separately using the checkerboard pattern of calibration target. To calibrated multi-camera system for stereo camera pair C_0-C_1, C_0-C_2 and C_0-C_3. The internal and external of multi-camera calibration parameters are showed as Table 2 and Table 3. That are respectively to find out the three-dimensional correspondence in space. Fig. 5 is the experimental result of omni-directional multi-camera calibration.

Fig. 5 (a) is the C_0-C_1 stereo camera as an example. The calibration parameters of the C_0 and C_1 single cameras are obtained respectively, and the C_0-C_1 stereo calibration parameters are obtained by binocular vision to find out the stereo correspondence in space. Similarly, another two sets of stereo cameras can be arranged to perform local three-dimensional reconstruction on C_0-C_2 and C_0-C_3. Finally, the multi-view cameras calibration and pose estimation are shown in Fig. 5(b).

Table 2: The internal camera calibration parameters

(a) C_0-C_1

Parameter	C_0	C_1
Principal point (u_0, v_0)	415.2, 277.1	403.7, 288
Focal length (fx, fy)	2593, 2593.2	2561.4, 2562.7
Skew	0.234	6.753
kappa 1	-0.038	0

(b) C_0-C_2

Parameter	C_0	C_2
Principal point (u_0, v_0)	416.5 , 271	410.3, 292.3
Focal length (fx, fy)	2619.5, 2619.9	2611.7, 2612.6
Skew	-0.012	-0.012
kappa 1	-0.043	-0.043

(c) C_0-C_3

Parameter	C_0	C_3
Principal point (u_0, v_0)	415.4, 265.1	364.2, 270.4
Focal length (fx, fy)	2618.3, 2618.5	2584.9 , 2585.2
Skew	0.006	-2.413
kappa 1	-0.041	-0.034

Table 3: The external of camera calibration parameters

(a) C_0-C_1

Parameter	Rotation [°]	Translation [mm]
X axis	41.38 ± 0.0031	94.36 ± 0.64
Y axis	-34.19 ± 0.0021	110.3 ± 0.76
Z axis	-7.591 ± 0.00014	87.28 ± 7.5
Baseline	169.405 mm	

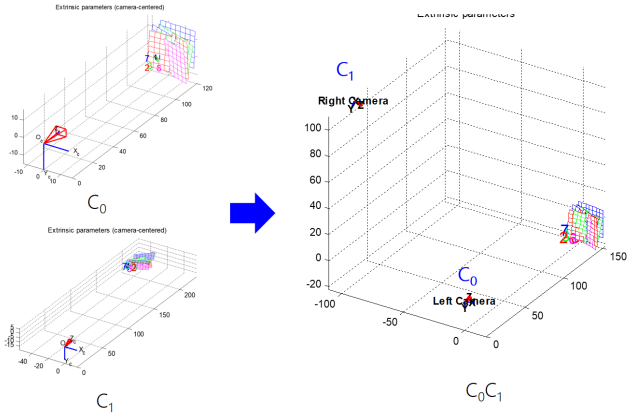
(b) C_0-C_2

Parameter	Rotation [°]	Translation [mm]
X axis	-40.67 ± 0.0026	107 ± 0.35
Y axis	2.595 ± 0.00042	60.76 ± 0.15
Z axis	118.1 ± 0.00048	57.3 ± 3.2

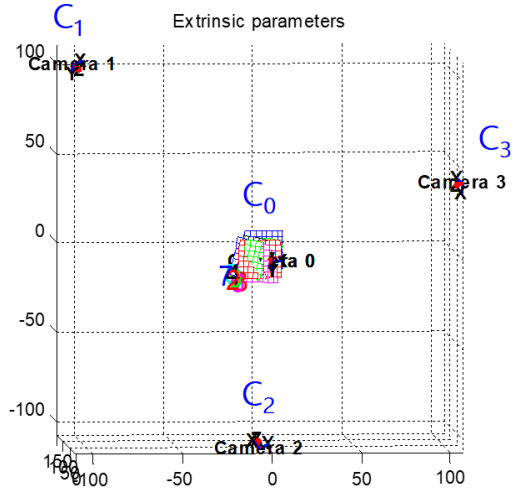
Baseline	135.734 mm

(c) C_0-C_3

Parameter	Rotation [°]	Translation [mm]
X axis	18.32 ± 0.0048	93.07 ± 0.53
Y axis	38.01 ± 0.0064	83.63 ± 0.4
Z axis	-108.4 ± 0.0019	66.92 ± 7.8
Baseline	141.901 mm	



(a) Stereo camera C_0-C_1



(b) Multi-view cameras

Figure 5: The experimental result of omni-directional multi-camera calibration.

3.2. Simulated omni-directional multi-camera image results

This study used the IDS UI-3130CP color camera as an example to discuss the calibration and 3D reconstruction of an omni-directional multi-view image measurement system. The omni-directional multi-camera image system is used to analyze the simulated omni-directional multi-camera image. Fig. 6 is IDS UI-3130CP color camera. Table 4 is IDS UI-3130CP color camera specifications. According to the optical design parameters of the measurement system: The image resolution is 800×600 pixels. The chip size of the image sensor is 4.8 μm×4.8 μm with a 12mm 1:2.0 1/1.8" lens. The best working distance is 15cm, and the simulation results can be measured within the working range of 3cm×3cm×3cm (L×W×H).

Figure 6: IDS UI-3130CP color camera

Table 4. IDS UI-3130CP color camera specifications

Name	**UI-3130CP Rev. 2**
Family	CP
Interface	USB 3.0
Sensor type	CMOS
Manufacturer	ON Semiconductor
Frame rate	396 fps
Resolution (h×v)	800×600 pixels
Optical Area	3.84 mm×2.88 mm
Shutter	Global Shutter
Optical class	1/3.6"
Resolution	0.48 MPix
Pixel size	4.8 μm×4.8 μm

Figure 7 is the simulated the setup of omni-directional multi-view measurement with four cameras. Fig. 8 is the simulated four cameras captured multi-view image. Four-camera multi-view image. The C_0 is the middle camera fixed as the geometric center point of measurement and provides a front image, is showed as Fig. 8(a). The other three cameras C_1~C_3 provide side images and the co-circular spheres are separated by 120 degrees to extend the circle, is showed as Fig. 8(b)-(d).

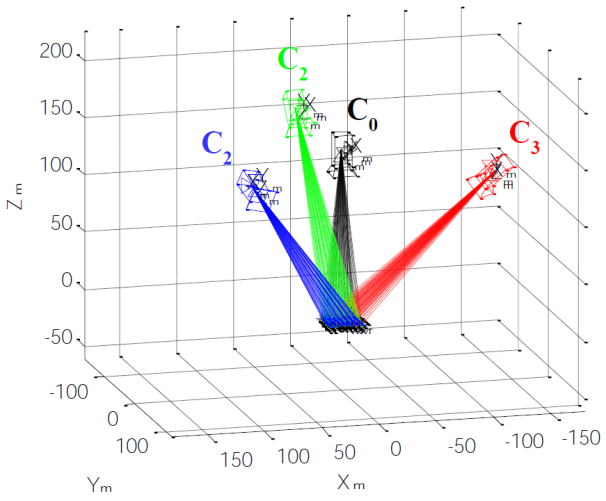
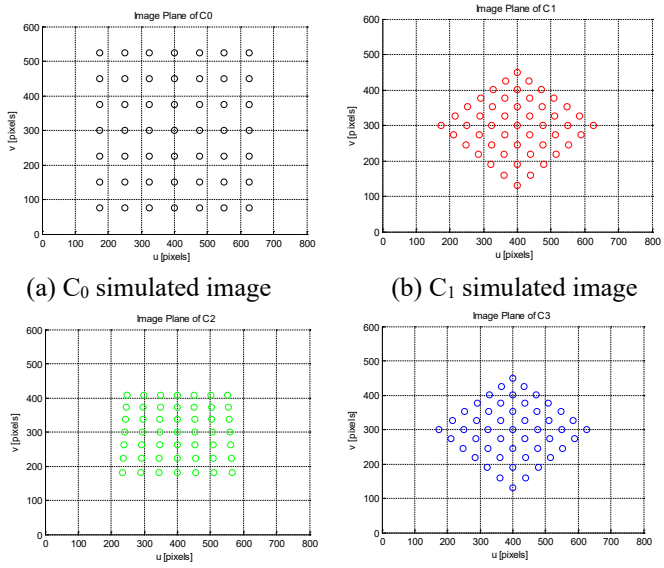


Figure 7: Simulated the setup of omni-directional multi-view measurement with four cameras



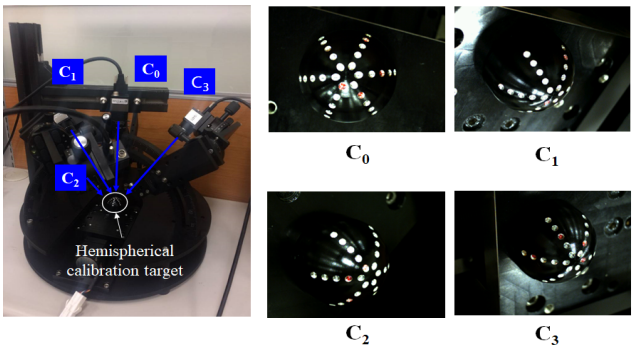
(a) C_0 simulated image (b) C_1 simulated image

(c) C_2 simulated image (d) C_3 simulated image

Figure 8: Simulated four cameras captured multi-view image

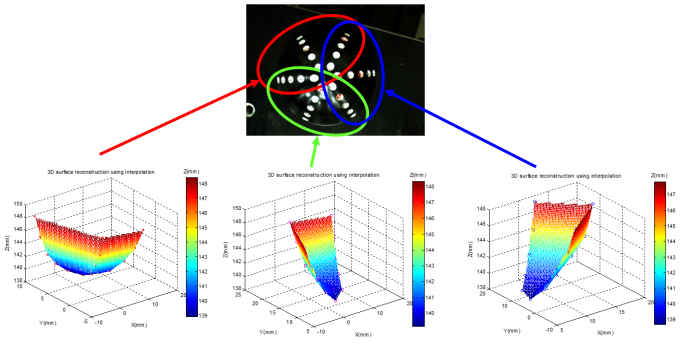
3.3. Omni-directional multi-view 3D reconstruction results

The hemispherical calibration target is a radius of 15mm and a height of 15mm in the middle of the omnidirectional multi-view camera. The multi-view images of four cameras C_0, C_1, C_2, and C_3 through any two cameras to form a stereo camera pair. Using the principle of parallax matching of the left and right images, the stereo camera pair C_0-C_1, C_0-C_2, and C_0-C_3 can be compiled performs three-dimensional reconstruction of each part. We used the hemispherical calibration target for the omnidirectional 3D reconstruction. Fig. 9 is the omni-directional original multi-view images. Fig. 10 is the omni-directional multi-view of 3D reconstruction for hemispherical calibration target.

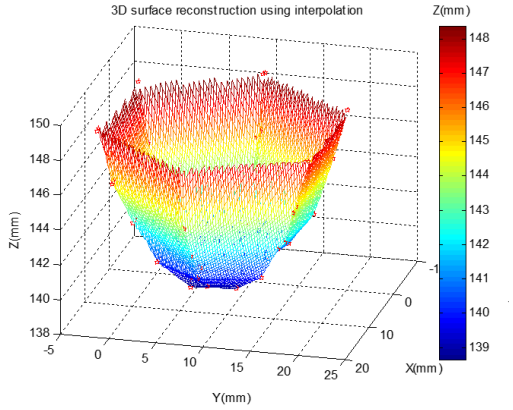


(a) Experimental setup (b) Original image

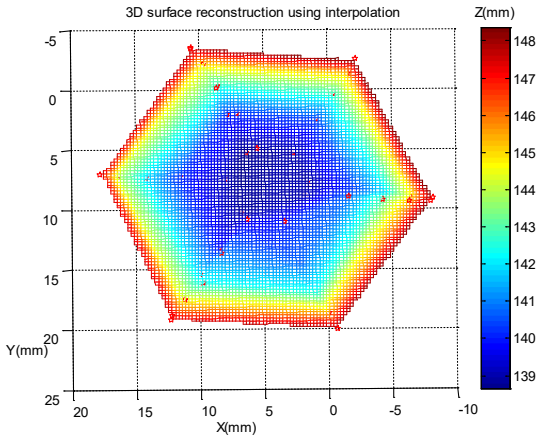
Figure 9: The omni-directional original multi-view images



(a) C_0-C_1 (b) C_0-C_2 (c) C_0-C_3



(d) Right-view of 3D



(e)Top-view of 3D

Figure 10: The omni-directional multi-view of 3D reconstruction for hemispherical calibration target

4. Conclusion

This research successfully developed an "omnidirectional multi-angle image measurement system". Any camera can move along the arc measuring rod to a designated position within the sphere to position and capture images. The three local 3D reconstructions of C_0-C_1, C_0-C_2, and C_0-C_3 are performed through the arrangement of stereo cameras, and the same position superposition calculation is performed, and the optimal 3D reconstruction calculation is performed using the ICP (Iterative Closest Point) iterative closest point algorithm. Finally, the 3D reconstruction experiment results of the 14.8 mm-high semicircular sphere calibration piece were obtained, and the measurement error was 1.3%. With the current best measurement working distance is 15 cm, the multi-angle imaging optical imaging analysis has been completed the measurement range of 3cm×3cm×3cm to 0.375 mm resolution.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Y.H. Chen, J.H. Huang, "Calibration and 3D reconstruction of omni-directional multi-view image measurement system," 2022 IEEE International Conference on Consumer Electronics - Taiwan, 591-592, 2022. DOI: 10.1109/ICCE-Taiwan55306.2022.9869096

[2] D. Murray, J.J. Little, "Using real-time stereo vision for mobile robot navigation," Autonomous Robots, **8**(2), 161-171, 2000. DOI: 10.1023/A:1008987612352
[3] N.G. Oh, J.I. Cho, K. Park, "On performance enhancement of a following tracker using stereo vision," ICCAS 2010, 1259-1262, 2010. DOI: 10.1109/ICCAS.2010.5669716
[4] E. Dandil, K.K. ÇEVİK, "Computer vision based distance meaurement system using stereo camera view," 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 1-4, 2019. DOI: 10.1109/ISMSIT.2019.8932817
[5] J. Liu, J. Pan, N. Bansal, C. Cai, Q. Yan, X. Huang, Y. Xu., "PlaneMVS: 3D plane reconstruction from multi-view stereo," In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
[6] Z. Yang, Z. Ren, Q. Shan, Q. Huang, "MVS2D: Efficient multiview stereo via attention-driven 2D convolutions," Computer Vision and Pattern Recognition (or CVPR) 2022.
[7] P.F. Luo, Y.J. Chao, M.A. Sutton, W. H. Peters, "Accurate measurement of three-dimensional deformations in deformable and rigid bodies using computer vision", Experimental Mechanics, **33**, 123-132, 1993. DOI: 10.1007/BF02322488
[8] Y.H. Chen, Y.S. Shiao, "Two axis independent parallel manipulators of control and stereo vision matching", Journal of Industrial Education and Technology, **33**, 137-148, 2008.
[9] C.H. Hwang, W.C. Wang, Y.H. Chen, "Camera calibration and 3D surface reconstruction for multi-camera semi-circular DIC system", International Conference on Optics in Precision Engineering and Nanotechnology (icOPEN2013), 8769, 123-132, 2013. DOI: 10.1117/12.2021044
[10] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," IEEE Journal of Robotics and Automation, **3**(4), 323-344, 1987. DOI: DOI: 10.1109/JRA.1987.1087109
[11] Z. Zhang, "A flexible new technique for camera calibration," IEEE Transactions on Pattern Analysis and Machine Intelligence, **22**(11), 1330-1334, 2000.
[12] Z. Zhang, "Camera calibration with one-dimensional objects," IEEE Transactions on Pattern Analysis and Machine Intelligence, **26**(7), 892- 899, 2004.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 4, 66-72 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Distribution Management Problem: Heuristic Solution for Vehicle Routing Problem with Time Windows (VRPTW) in the Moroccan Petroleum Sector

Younes Fakhradine El Bahi[1], Latifa Ezzine[2], Zineb Aman[*,1], Imane Moussaoui[3], Miloud Rahmoune[3], Haj El Moussami[1]

[1]*Mechanics & Integrated Engineering, ENSAM School, Meknes, 50000, Morocco*

[2]*Modeling, Control Systems and Telecommunications, EST-Meknes, Moulay Ismail University, Meknes, 50000, Morocco*

[3]*Advanced Materials Studies and Applications, EST-Meknes, Moulay Ismail University, Meknes, 50000, Morocco*

ARTICLE INFO

Article history:
Received:26 February, 2023
Accepted:4 June, 2023
Online: 25 July, 2023

Keywords:
Heuristic
VRPTW
Petroleum Sector

ABSTRACT

The attributes of the vehicle routing problem (VRP) are as many additional characteristics or constraints which aim to better take into account the specificities of real application cs. The variants of the VRP thus formed are the support of an extremely rich literature, comprising an immense variety of heuristics. This article constitutes an industrial application and an objective synthesis of successful and challenging heuristic concepts for time-windowed VRP problems. The purpose will be to minimize transport costs and determining the optimal number of trucks by applying a transport algorithm. The results show that the solution method should help to increase the competitiveness of transportation operations in this important economic sector.

1. Introduction

Oil is one of the most important raw materials in the world. It has been the primary energy source since the mid-1950s. The oil and gas industry plays a vital role as the engine of the global economy. The products produced by this industry support many other vital industries like automotive industry and manufacturing industry.

Changing technologies, markets and customer needs affect the competitiveness of companies, which requires a continuous restructuring of the strategy and positioning tactics of the oil industry. Currently, the main problem facing the oil industry is to minimize transportation costs and production costs. Effective supply chain management can increase overall efficiency, competitiveness, and good sourcing.

The economic importance of the transport sector explains the interest of researchers in the problems of vehicle routing Laporte [1-3]. Much attention has been paid to the development of route definition techniques as they have the potential for major cost savings.

The gas station supply problem consists in determining how to optimally distribute several products (mainly gasoline and diesel fuel) to customers (gas stations) from a depot (refinery or regional depot) according to a planning horizon chosen. It is therefore necessary to determine for each period of the planning horizon, the stations to be visited, the routes and the trucks to be used, the quantity to be delivered of each product and their assignment to the trucks, and this, in ensuring that no client station is out of stock.

The Vehicle Routing Problem with Time Windows (VRPTW) is a major operational research problem. It consists of determining a set of routes to collect or deliver goods to customers within time windows and capacity constraints. Customers are usually visited once and only once by a vehicle; their request must be satisfied by this single visit.

The article is divided into four sections. After the introduction, section 2 presents a review of the literature on vehicle routing problems especially VRPTW. Section 3 offers a presentation of our case study. And Section 4 presents our conclusions and directions for future research.

conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

[*]Corresponding Author: Zineb Aman, ENSAM School, zineb.aman@gmail.com

https://dx.doi.org/10.25046/aj080408

2. Literature review

Vehicle routing problems are much studied because of the growing importance of passenger and freight transportation today. The simplest and best-known problem of this type is the Traveling Salesman Problem (TSP). It consists of determining a route routed by a vehicle, so as to serve a set of customers distributed in a network at minimal cost. This basic model can be enriched by various constraints relating to the number of vehicles, to their loads, or to constraints relating to the nodes, to their time windows, or to dependencies between them.

Several researches are mainly oriented towards solving the vehicle routing problem VRP (Vehicle Routing Problem). The latter is a problem of optimizing vehicle routes to satisfy transport demands. These vehicle routing problems are generally subject to several types of constraints.

The general vehicle routing problem is known as the Vehicle Routing Problem (VRP) and represents a multi-objective combinatorial optimization problem that has been the subject of many works and many variants in the literature. It belongs to the NP-hard category [4-6].

The VRPTW constitutes a generalization of the VRP insofar as we also introduce a temporal constraint on the requested service. Each customer has a window of time within which he wishes to be served. The central depot also has a time window which we commonly refer to as the service horizon or open time of the day. Its role is to set a time slot during which vehicles can make their rounds.

These time constraints will make it necessary to use several vehicles to satisfy all customers over the service horizon. We may want to limit the number of vehicles to be used and in this case customers may not be served [7].

Each customer must be served within a defined time interval, known in advance by the delivery person and any violation of this constraint may result in a penalty. When the time window constraint is not satisfied, either we reject the solution if we consider the rigid case or we construct a penalty function which will be added or combined with the objective function for the case released. In fact, this is a very common problem. Distribution of perishable products (milk, meat, etc.), newspapers, outpatient services, . . . are practical examples of the VRPTW. Within this class of problems, there are two subclasses:

- The rigid PTVFT where the service must imperatively be performed within the time window.
- The released PTVFT or the delay or the advance only generates a penalty [7].

A time window (TW - Time Window) imposes that one or more problem requests be processed respecting an interval for the start of vehicle processing (delivery, collection). This interval is defined by an earliest date and a latest date.

This constraint can apply to pickup or delivery requests [8]. This type of constraint most often involves allowing the vehicle to arrive early, in which case it waits on the vertex in order to start its operations at the start of the time window. This waiting can be authorized on all the vertices of the graph or, only on a part of the vertices. However, the vehicle is not allowed to arrive after the end of the time window. In the classical version of the constraint, if one of the time windows is not respected then the solution is not valid. There are variants in the literature, for example the Soft Time Window (STW) introduced for the first time on a TSP problem [9], then on a VRP problem [10], which allows violation of the time window but penalizes the objective function [8].

The vehicle routing problem (with time windows) is one of the routing problems. Lenstra and Rinnooy Kan proved that VRP is an NP-DUR problem. Its resolution by an exact method turns out to be inappropriate for large instances. It is therefore inevitable to proceed to its resolution by heuristic approaches, which provide feasible and appreciable solutions in a reasonable time [11]. Recently, researchers treat VRPTW by different ways [12-18].

3. Case study

This part evaluates the effectiveness of the solutions proposed at the level of our previous work [19], and this is done by adding more constraints on the model which will make it more real and better solve the problem of distribution of petroleum products of the company.

This work makes a contribution in two ways. First, we deal with an actual case that differs slightly from the scenario that was previously addressed. Trucks, for instance, operate in shifts with predetermined loading and distribution window times. In contrast to earlier language, shift start and finish times varied from truck to truck.

The ultimate goal in this situation is to maximize customer order fulfillment while adhering to a specific priority and using a small, heterogeneous fleet. Additionally, because each journey has a distinct duration and cost, the charges are unaffected by the amount of time that passes between the start of the first trip and its conclusion.

By forcing the execution of the orders, the priority of the consumers may have been managed; however, because the orders can only be concentrated for a limited period of time, the issue is insoluble.

Our second contribution consists of a heuristic suggestion for this particular instance, the major goal of which is to quickly arrive at a workable operational solution so that people in charge of supply chain management can assist it.

This heuristic is an adaptation of Solomon's insertion heuristic for problems involving vehicle scheduling or routing with time windows.

Since each customer has value that is not always reflected in the net benefit of a transaction, looking at it from the perspective of customer compliance offers a more comprehensive perspective than merely cost reduction. In the case of fuels, this is crucial because the profit per transaction is generally low and the gain is based on the volume sold. As a result, it could be crucial to prioritize a customer who makes a lot of orders.

In this section, we'll explicitly characterize the issue at hand and provide some premises and traits that will affect how we approach finding a solution.

3.1. Problem definition

The VRPTW used in this real case can be defined as follows:

Let $G = (\{0, n + 1\} \cup C, A)$ be a directed graph where:

$\{0, n + 1\}$ is the deposit,

$C = \{1, \ldots, n\}$ is the set of customers,

$A = \{(i, j) : i, j \in \{0, n + 1\} \cup C, i \neq j\}$ is the set of arcs. Each arc (i, j) is associated with a travel time ti,j. Similarly, the service time si at depot or station i, where $i \in \{0, n + 1\} \cup C$, is known. The set of trucks associated with the depot is denoted K.

Customer i orders are made up of Pi order lines, where each line specifies a different product type, bucket type to use, and ideal time window. Trucks must arrive at customer i within the proposed time window $[ai, bi]$. Also, each truck must respect their team's time window $[\alpha'k, \beta'k]$.

It is crucial to strike a balance between the aspects of reality that we will model and those that we will simplify through assumptions because of the true nature of the problem, which frequently results in instances where theory and reality diverge.

The model's route generating step contains one of its most significant simplifications. Travel time is directly related to the zone a customer is in because customers are organized by zones. Generally speaking, trucks that visit multiple stations should only go to those that are nearby. Orders may be clubbed together for the purpose of assigning them to one and only one feasible route due to this and other limitations. In relation to capacity, the resulting routes will only have one type of vehicle, allowing us to solve the problem for each type of truck separately. Orders that do not comply with these rules are left out of this issue, as the compliance decision is in the hands of the planners.

- It's interesting to see that less than 1% of all orders are frequently refused. Following that, more crucial presumptions for this specific issue are noted:
- Orders are made up of order lines, which each client can create until a truck is full. More than one product type may be present in these orders, but only one may be present in each compartment.
- Customers are permitted to create multiple orders. Each of these is handled independently because it has its own specifications and due dates.
- All deliveries must be made in a single trip.
- More than one consumer may accompany a journey, but no more than two.
- Trucks may operate many shifts throughout the day. They are viewed differently because they are independent.
- There is a relevance associated with the priority of a command, this will be represented by the constant ρ.

3.2. Resolution process

We divided the problem into two phases in order to solve it in a reasonable amount of computation time:

- Phase I: Creation of requests on a monthly basis (in our example, the month of July).

- Phase II: Scheduling and truck assignments based on the VRPTW mathematical model. The following notations will be used in addition to the parameters that have already been defined:

- Hints:

t, l: Road index

k: Truck index

v : Trip position index

- Sets:

T : Road set

Tk : Set of routes associated with truck k

V : Set of possible path indices. $V = \{1,2, \ldots, N\}$

Vk : Set of possible route clues for truck k

- Parameters:

Di : D$emand$ of $station\ i$

L : Number of $stations$

dij : D$istance$

C : Truck $capaci$ty

Ri: Residual demand < 33, which can be delivered in a second time

$Zt = \{i, j\}$ / $t \in [1: n]$: cluster of station i and j that we will deliver them in one truck

(the clusters must group together a maximum of 2 stations) and n is the total number of clusters

T_s: Total demand of day S

$n_{i/s}$: Customer i should be delivered the day s

n_{ij}: Number of full trucks to deliver for a customer i which will define the frequency / duration in which I can deliver a single truck to a station. Example: If nij =2 truck then the frequency is 1 month, nij = 4 , frequency) 1 truck per week. nij> 4, frequency every 2 days..)

$B_s = Ts/I$: Number of trucks to be loaded at a time

N : Maximum number of possible trips during a shift

ρ_t : Value associated with the priority of path t

α_t : Start time of route time window t

β_t : End time of route time window t

λ_t : Route duration t

α_k : Start date of truck time window k

β_k : End date of truck time window k

Next, we describe the two phases of this problem and how we deal with them.

Phase I : Generation of requests according to a monthly schedule (the month of July in our case)

This first phase is considered an initialization process.

It consists of generating the delivery schedule for monthly requests (the month of July in our case) for all the service stations, as well as the total number of trips during this month and subsequently the number of trucks to rent. The aggregation of orders in the trucks must follow certain rules of the company. This allows us to group commands so that they are assigned to one and only one possible route.

- Depot: working hours (from 5 a.m. to 1 p.m. // 2 p.m. to 4 p.m.) 26 days a month:

It is a common depot between 5 companies located in Mohammadia, it contains 17 truck loading stations (full load). The number of trucks to be loaded at a time varies from one company to another with a loading time of 1 hour.

Shell 17 trucks at a time;

OLA ENERGY 10 trucks simultaneously with Winxo 7 Trucks;

Total 12 simultaneous with Petruom 5 Trucks.

The order of passage is done in turn each week, which then defines the maximum number of possible loadings. Let I be the maximum number of trips to be made in one day for our company:

Week 1: Ola is the first: 4 trips *10 = 40 trucks

Week 2: Ola is second: (I = 3 trips) *10 = 30 trucks

Week 3: Ola is third: (I = 3 trips) *10 = 30 trucks

Week 4: Ola is the first: (I = 4 trips) *10 = 40 trucks.

- Customer storage capacity:

Each customer has a monthly demand and storage capacity. If the capacity is less than 33T, we cannot deliver a full truck so we split the complete order into 2 orders or more. Otherwise, we deliver a complete truck.

- Algorithm :

1- $Di\ /C\ =\ Ni$

2- $Ni - E(Ni)\ =\ Ri,\ 26\ /\ E(Ni)\ =\ nij$ and $j\ \in \{1, \dots, 26\}$

3- $Cluster\ stations\ :\{Ri + Rj = 1(1\ camion\ )$ and $dij \leq 30km\ \rightarrow\ \square Zn$

4- $Ts\ =\ \sum\ n_{i/S}$ for $s\ =\ 1:26$

$n_{i/S} = 1\ if\ s\ modulo\ nij = 0\ \&\ n_{i/S} = 0\ else$

$If\ (n_{i/S} \neq 0\ \&\ n_{j/S} \neq 0)\ Avec\ i, j\ \in Zt;\ \ t \in [1:n]$

$Ts = Ts + 1$ (We add the trucks that will deliver the residual of the cluster) $\&[1:n] = [1:n] - 1$ (We subtract the cluster that we will deliver on the day s)

$Break$

5- $Bs\ =\ Ts\ /\ I$

$If\ Bs > 10$ so $As\ =\ I\ *10$

$Xs = Ts\ - As$

(Xs it's the deliveries that I have to deliver but I don't have the right for the day s, so I will memorize these deliveries and satisfy them the day I have more trucks)

6- $Xs\ =\ XX_{S-1} + Xs$

$For\ p\ =\ 1 : Xs$

$If\ Bs <\ 10$

$Bs\ =\ Bs\ + 1$

$Xs\ =\ Xs\ - 1$

$Nc\ =\ max\ Bs\ for\ \{s = 1, \dots, 26\}$

Phase II : Truck assignment and scheduling from the VRPTW mathematical model

The company's primary goal is to maximize the trucks and trips that stop at the most service stations while adhering to their time restrictions, each station's priority weighted accordingly.

The assignment problem's goal function can be stated as follows: Maximize:

$$\sum_{t \in T} \sum_{v \in V} \sum_{k \in K} \rho t x_{tvk}\ , \qquad (1)$$

The binary variable xtvk has a value of 1 if route t corresponds to the v-th trip of truck k and a value of 0 otherwise. We employed a sequential insertion heuristic, which distributes roads to one truck at a time until the truck can no longer accept another one, because addressing this problem using exact approaches proved challenging with a huge number of trucks and roads.

For each route of the truck, the starting position and time must be defined, which leads to the scheduling problem. The following model deals with route scheduling for each truck k:

- Variables :

x_{tv}: binary variable which takes the value 1 if the route t corresponds to the v-th trip of truck k; 0 otherwise.

d_v: departure date of the truck's v-th trip k

Minimize: $\sum_{v \in Vk}\ dv \qquad (2)$

Under constraints of:

$$\sum_{v \in Vk} xt,v = 1, \quad \forall t \in Tk \quad (3)$$

$$\sum_{t \in Tk} xt,v = 1, \forall v \in Vk \quad (4)$$

$$\sum_{t \in Tk} at\, xt,v \leq dv \leq \sum_{t \in Tk} \beta txt,v, \quad (5)$$

$$\forall v \in Vk$$

$$dv \geq dv - 1 + \sum_{t \in Tk} \lambda txt, v - 1, (6)$$

$$\forall v \in Vk \mid v \neq 1$$

$$\alpha' k \leq d1 \, (7)$$

$$dn(Vk) \leq \beta' k - \sum_{t \in Tk} \lambda txt, n(Vk) \, (8)$$

$$xt,v \in \{0,1\}, \quad \forall t \in Tk, \ \forall v \in Vk \, (9)$$

$$dv \in \mathbb{R}, \forall v \in Vk \, (10)$$

The goal function (2), which relates to a supplementary optimization criterion, seeks to decrease departure times while prioritizing the shortest routes before the longest ones. Truck k will only make one of the trips given to it, thanks to constraint (3). Each travel position will only be able to handle one route, according to constraint (4). Each route must adhere to its time frame if it corresponds to the v-th trip, according to constraint (5). The start time of the v-th trip will begin after the previous trip has ended, according to constraint (6). The first journey is required to adhere to the truck's departure time per constraint (7). The last trip must finish before the truck's end time k, according to constraint (8). Lastly, limitations (9) and (10) show the nature of the variables.

The sequential insertion heuristic we propose can be expressed as follows:

For each truck:

1. Populate a list with all unassigned routes.

2. Sort routes in descending order of priority.

3. Filter routes based on truck time window compatibility.

4. Assign the truck a route from the list that meets a certain departure criteria.

5. For each item in the list:

(A) Verify compatibility with the allocated routes as of right now. If the (a) check is positive, solve the scheduling problem including the current route.

If not, proceed to the next thing on the list.

(b) If the scheduling issue has been correctly resolved, assign the current route to the truck, remove it from the list, and determine if any other possible routes are still available. If not, proceed to the next thing on the list. (c) Go to the next item on the list if there are any suitable paths. If not, proceed to the next truck.

Once all vehicles have been moved or all routes have been assigned, the process is complete.

On real data from the month of July, we tested our heuristic.

Each day's orders and truck involvement varied, allowing a wide range of potential scenarios to be covered.

The Python code for our heuristic was created using Jupyter Anaconda Navigator on a Windows 10 PC with an Intel Core i5-4200U processor clocked at 1.6 to 2.3 GHz and 4GB of RAM.

Table 1 and Figure 1 present the findings.

Our heuristic therefore allowed us to plan the path or the circuit of truck 1 for example in an optimized way.

Table 1: Planning extract for Truck 1

	Truck 1					
	ID	**Day**	**Arrival Time**	**Departure Time**	**Delivery Volume**	**Distance traveled**
Deposit	LOMA0662668	Thursday		07:00		0
Station-service 6	LOMA0672999	Thursday	08:00	09:30:00	33	18
Deposit	LOMA0793001	Thursday	10:53	12:00		104
Station-service 11	LOMA0662997	Thursday	12:59	13:29	16	124
Station-service 28	LOMA0662974	Thursday	13:36	14:06	17	127
Deposit	LOMA0662668	Thursday	15:16		0	304

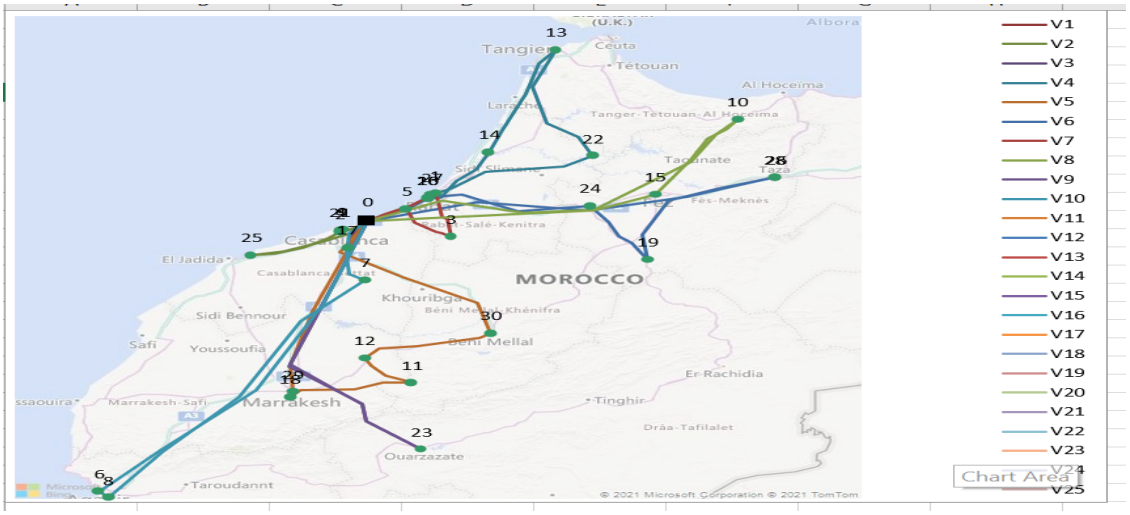


Figure 1: Truck route for the first day of July

4. Conclusion

Our heuristic produces a suggestion that planners typically ask for in the morning or throughout the day, particularly when significant alterations to the current schedule are required. As a result, they must be able to quickly deliver a solid initial plan. The greatest cases of this problem can be handled by the heuristic in a maximum of 10 minutes, which is a respectable amount of time given the compliance attained.

According to the data, it is possible to achieve at least the same level of compliance as the company, thus it is possible to include the truck teams' time slots in the scheduling issue the company confronts.

Future research will concentrate on the impact of more real-world factors in the two phases mentioned, testing the heuristic's various aspects to see if it can solve them, including: the environmental factor, worker safety, road quality, and the efficiency of using secondary optimization criteria on the solution. Finally, testing various route departure criteria may help the heuristic perform better.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

I am grateful to all of those with whom I have had the pleasure to work during this and other related projects.

References

[1] "The vehicle routing problem: An overview of exact and approximate algorithms", European Journal of Operational Research, **59**, 345-358, 1992, doi.org/10.1016/0377-2217(92)90192-C.

[2] G. Laporte, "The traveling salesman problem: An overview of exact and approximate algorithms", European Journal of Operational Research, **59**, 231-247, 1992, doi.org/10.1016/0377-2217(92)90138-Y.

[3] G. Laporte, I. H. Osman, "Routing problems: A Bibliography", Annals of Operations Research, **61**, 227-262, 1995, doi.org/10.1007/BF02098290.

[4] N. Christofides, A. Mingozzi, P. Toth, "The vehicle routing problem", 315-338. Wiley, Chichester, 1979, doi.org/10.1007/978-1-4615-5755-5_1.

[5] M. Desrochers, J.K. Lenstraa et M.W.P. Savelsbergh, "A classication scheme for vehicle routing and scheduling problems", European Journal of Operational Research, **46**(3), 322-332, 1990, doi.org/10.1016/0377-2217(90)90007-X.

[6] M. Sol et M. Savelsbergh, "A branch-and-price algorithm for the pickup and delivery problem with time windows", Memorandum COSOR 94-22, Dept. Of Mathematics and Computin, 1994.

[7] H. Housroum, "Une approche génétique pour la résolution du problème VRPTW dynamique", Thèse Doctorat en informatique, Université d'Artois, 2005.

[8] M. Chassaing, "Problèmes de transport à la demande avec prise en compte de la qualité de service", Thèse Doctorat en informatique, Université Blaise Pascal-Clermont Ferrand II, 2015.

[9] T.R. Sexton, Y.M. Choi, "Pickup and delivery of partial loads with "soft" time windows", American Journal of Mathematical and Management Sciences, **6**(3-4), 369-398, 1986, doi.org/10.1080/01966324.1999.10737484

[10] S.P. Greaves, M.A. Figliozzi, "Collecting commercial vehicle tour data with passive global positioning system technology: Issues and potential applications", Transportation Research Record, **2049**(1), 158-166, 2008, doi.org/10.3141/2049-19.

[11] M. Akil, "Problème de tournées de véhicules avec contraintes et fenêtre de temps", Mémoire Magister en informatique, UMMTO, 2013.

[12] W. Dong, K. Zhou, H. Qi, C. He, J. Zhang, "A tissue P system based evolutionary algorithm for multi-objective VRPTW", Swarm and evolutionary computation, 39, 310-322, 2018, doi.org/10.1016/j.swevo.2017.11.001.

[13] W. Zhang, D. Yang, G. Zhang, M. Gen, "Hybrid multi-objective evolutionary algorithm with fast sampling strategy-based global search and route sequence difference-based local search for VRPTW", Expert Systems with Applications, **145**, 113-151, 2020, doi.org/10.1016/j.eswa.2019.113151.

[14] M. Saint-Guillain, Y. Deville, C. Solnon, "A multistage stochastic programming approach to the dynamic and stochastic VRPTW", In International Conference on Integration of Constraint Programming, Artificial Intelligence, and Operations Research, 357-374, Springer, Cham, 2015, doi.org/10.1007/978-3-319-18008-3_25.

[15] T. Gocken, M. Yaktubay, "Comparison of different clustering algorithms via genetic algorithm for VRPTW", 2019, doi.org/10.2507/IJSIMM18(4)485.

[16] V. Pureza, R. Morabito, M. Reimann, "Vehicle routing with multiple deliverymen: Modeling and heuristic approaches for the VRPTW", European Journal of Operational Research, **218**(3), 636-647, 2012, doi.org/10.1016/j.ejor.2011.12.005.

[17] H.B. Ticha, N. Absi, D. Feillet, A. Quilliot, "Empirical analysis for the VRPTW with a multigraph representation for the road network", Computers & Operations Research, **88**, 103-116, 2017, doi.org/10.1016/j.cor.2017.06.024.

[18] W. Xu, X. Wang, Q. Guo, X. Song, R. Zhao, G. Zhao, D. He, "Gathering Strength, Gathering Storms: Knowledge Transfer via Selection for VRPTW", Mathematics, **10**(16), 28-88, 2022, doi.org/10.3390/math10162888.

[19] Y.F. El Bahi, L. Ezzine, Z. Aman., H. El Moussami, "Distribution management problem: case of vehicle routing problem with capacity constraints "CVRP" in the Moroccan petroleum sector", In 8th International Conference on Control, Decision and Information Technologies (CoDIT) (1), 1149-1154, IEEE, 2022, doi.org/10.1109/CoDIT55151.2022.9803955.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 69-79 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Analysis of Linear and Non-Linear Short-Term Pulse Rate Variability to Evaluate Emotional Changes during the Trier Social Stress Test

Alvin Sahroni[*,1], Isnatin Miladiyah[2], Nur Widiasmara[3], Hendra Setiawan[1]

[1]*Electrical Engineering Department, Universitas Islam Indonesia, Yogyakarta, 55584, Indonesia*

[2]*Pharmacology Department, Universitas Islam Indonesia, Yogyakarta, 55584, Indonesia*

[3]*Psychology Department, Universitas Islam Indonesia, Yogyakarta, 55584, Indonesia*

ARTICLE INFO

Article history:
Received: 31 December, 2022
Accepted: 05 May, 2023
Online: 15 May, 2023

Keywords:
Mental Stress
Emotion
Pulse Rate Variability
Short-term
TSST

ABSTRACT

In conjunction with psychological stress, physiological indicators such as heart rate (HR) and heart rate variability (HRV) are frequently employed. This study uses a substitute for heart rate variability (HRV) known as short-term pulse rate variability (PRV) to evaluate emotional changes. We examined sixteen college students using a low-cost photoplethysmograph and obtained a short-term PRV reading from one of their index fingers. Each PRV's parameters during resting condition were established using a particular Trier Social Stress Test (TSST) and divided into four phases (R1: baseline rest, R2: anticipatory stress, R3: stressful event, and R4: recovery period). The Positive Affect and Negative Affect Schedule (PANAS) questionnaire was used to assess the participants' moods during the experiment. The psychological assessment based on PANAS shows that negative affect tends to increase along the TSST procedure, especially from the baseline rest (R1) to the stressful event (R3), even though it is not statistically significant ($p > 0.05$). The physiological assessment using PRV revealed that between R1 and R3, the short-term SDRR, pNN50, RMSSD, LF, HF, total power, SD1, and elliptical area of PRV tended to rise considerably ($p < 0.01$). The properties of PRV show that the heart rate fluctuation also represents psychological changes. We concluded that applying the TSST procedure to induce stress modulates the features of PRV, particularly the time and frequency domain variability properties. Observed patterns following stressful events (interview sessions) indicated an increase in PRV values, mainly the RMSSD, HF, SD1, and elliptical area ($p < 0.001$). Caution should be applied while perceiving physiological alterations as an immediate sign of mental threats. Nonetheless, these alterations are likely caused by the association between stress and negative emotions.

1. Introduction

Individuals may encounter a variety of stressful situations in their daily lives, and if not managed correctly, these events can have adverse effects [1]. Negative emotions, such as anxiety, fear, and anger, are the primary reasons for certain psychological states and the negative reactions resulting from stress [2]. Negative emotions can also shorten attention spans and reduce concentration [3]. Mental stress is spreading rapidly in modern society, permeating practically all aspects of daily life [4]. Therefore, stress is re ceiving much attention in contemporary society, as most people in developed countries regularly endure physical or psychological symptoms caused by stress [5]. Additional symptoms include persistent illness, fatigue, and headaches [6].

Furthermore, stress is already recognized as a mutually beneficial connection between individuals and their surroundings, that strains or exceeds an individual's ability to cope. Harmful, frightening, or unpleasant circumstances frequently cause stress, leading to anger, embarrassment, and anxiety, showing a connection between stress and emotion [7]. According to researchers, the link between stress and negative emotions is inexorably linked. For instance, some research suggests that depression is a stress reaction [8, 9]. Similarly, stress can foster the onset of disease by compromising the immune system as it worsens [10]. Consequently, it is essential to address the source of

[*]Corresponding Author: Alvin Sahroni, alvinsahroni@uii.ac.id

tension. Additionally, developing a system that can readily and accurately evaluate and detect mental stress and other negative emotions is crucial.

Modern methods have been developed for scientifically evaluating psychological changes. The most popular methods for evaluating psychological traits involve completing surveys, scheduling consultations with psychologists, or evaluating online encounters during the pandemic. However, these procedures are time-consuming, expensive, and not always practical for everyone. Inconsistency in filling out the response sheets and failing to understand the instructions render questionnaires worthless as objective assessments [11-13].

Psychophysiological studies have been conducted to objectively improve the psychological assessment to observe how psychological changes can affect the physiological aspects. Thus, experimental studies were conducted to evaluate physiological responses after inducing stress or other emotions and moods. Most studies use a mental arithmetic test or another task demanding participants to accomplish it while measuring their body's response. The Trier Social Stress Test (TSST) was established and developed more than two decades ago, and researchers have discovered that it has become one of the most prevalent methods for inducing stress under controlled conditions [14].

Another way to assess mental wellness accurately is by utilizing biochemical and hormone details. Objectively evaluating salivary cortisol levels is one of the best ways to measure psychological conditions associated with human physiology. It is predicated on a specific feeling or thought, such as stress, which releases cortisol and adrenaline stress hormones by activating the sympathetic nervous system and the hypothalamic-pituitary-adrenal (HPA) axis [15]. Cortisol is unreliable for quickly estimating mental stress because of the handling period and the possibility of inaccuracy during repeated measurements [16, 17].

Previous electrophysiological studies suggest that electric biosignals, such as trapezius muscle activity, electroencephalogram (EEG), and electrocardiogram (ECG) data, as well as an assortment of these methods, can be used to detect mental tension and a range of emotions [18]. Biosignals are seen as a practical, unbiased way to evaluate psychological changes because there are few psychological tests, and salivary cortisol preparation is technically challenging. Another issue is the wearable feature of evaluating psychological characteristics while engaging in ordinary tasks.

Notably, most of the previously described research used HRV parameters from standard ECG equipment as the objective assessment to extract the heart's bio-signal information to quantify emotions or mental stress [18]. Regardless, the pulse rate variability (PRV) using photoplethysmography as an instrument for examination still has several limitations, particularly given its low-cost development. The fact that PRV cannot entirely substitute HRV in some circumstances is one of the main arguments against using it in place of an ECG [19, 20]. A wearable device for daily use may not be appropriate for ECG's HRV since it requires numerous electrodes to be hooked to the body, which may be discomfiting for certain people. The latest research that utilized PPG to encounter the uncomfortable issue also showed that remote PPG (rPPG) could detect mental health conditions with 96% accuracy using the heart rate parameter during cognitive stress tasks [21]. This report shows that our proposed study is not the first time PPG has been used as a physiological tool to detect stress, and the development cost is also considered low. Therefore, more research is needed in this area to discover if PRV can be utilized to evaluate emotions or other psychological states throughout different settings.

The majority of studies on the TSST procedure still rely on through traditional heart rate measurement devices to quantify the physiological properties. However, the study that measured the pulse transit time (PTT) using the fingertip photoplethysmograph (PPG) and the electrocardiogram (ECG) during TSST sessions uncovered extensive information concerning the cardiovascular system that enabled observation of physiological changes after stress induction [22]. The most recent PPG result have a comparable capacity to identify mental stress as well as emotions, producing results with an accuracy of over 80%. Instead of using the Trier Social Stress Test (TSST) to exploit the susceptibility to the stress response, the majority of studies employed Mental Arithmetic Tests (MAT) and Stroop Tests to induce negative emotions [23].

This study aims to determine if stress inducement can affect negative emotions, as well as which interview events have the greatest impact on an individual's mental state. To better quantify and objectively evaluate this relationship between physiological characteristics and emotions as a sign of mental stress , a more thorough understanding is necessary. Therefore, we recommend employing the Trier Social Stress Test (TSST) to induce psychological stress, assess the subsequent affective changes, and extend the biosignal analysis to measure mental stress. This study may demonstrate the importance of identifying emotional traits in addition to mental stress, which could contribute to incorrect diagnoses of various stress symptoms. In contrast to our previous findings, which only analyzed specific PRV parameters (time domain and linear analysis), we expand the study by applying the common objectives of PRV's linear and non-linear, as well as the frequency domain analysis. The motivation is to discover additional parameters that can identify and better understand psychological changes.

Our previous study reported that a portable photoplethysmograph sensor could measure emotional changes after stress induction [24]. A low-cost device can objectively present emotional or mental changes based on physiological changes. As a result of our analysis of the pulse rate variability (PRV) time series, data revealed that RMSSD was the most crucial metric for capturing negative feelings and moods following the generated stress process. RMSSD was substantially higher after the main task or during the stressful event (R3) compared to before the interview or baseline rest (R1) ($p < 0.01$) [24]. We believe that additional linear and non-linear analysis could further explain the understanding of electrophysiological properties during stress or adverse effects from a psychological aspect in terms of the autonomic nervous system, even though our previous report already found a critical finding demonstrating physiological changes during mental state discrepancy.

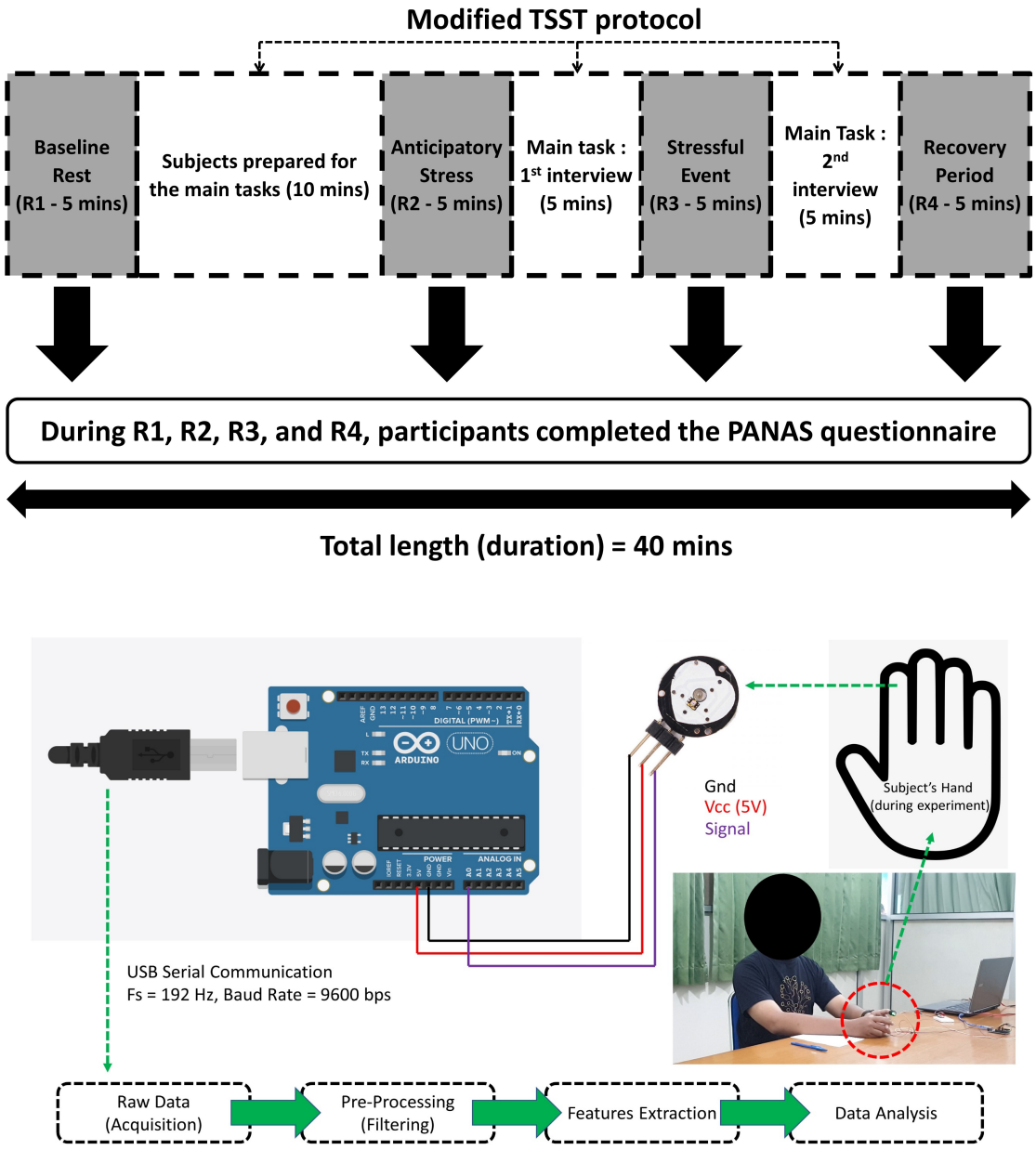


Figure 1: Trier Social Stress Test (TSST) experimental design and apparatus for evaluating psychological factors based on the Positive Affect Negative Affect Scale (PANAS) questionnaire and physiological properties [24]

2. Methods and Materials

2.1. Research Ethics

Prior to the conducting the investigations, the research topic and methodology were disclosed to the participants, and their informed consent was obtained. In pursuance of the principles of the Declaration of Helsinki, the Ethics Committee of the Faculty of Medicine at Universitas Islam Indonesia Yogyakarta approved this study (approval number 80/Ka.Kom.Et/70/KE/II/2019). The information of every subject's was kept secure and confidential.

2.2. Participants and Procedures

This investigation included sixteen healthy male and female students between the ages of 19 and 21 who were participated in this study. The participants were undergraduate or graduate students majoring in social, medical, or engineering fields. Each participant signed an informed consent form. The participants declared in the informed consent document that they were in healthy condition, had not taken any regular medications before the studies, had slept well, had abstained from smoking, had normal weight based on Body Mass Index (BMI), and had no record of cardiovascular or other severe disorders in their families. Prior to involvement, the study was unknown to the participants, and the requirements of the entire procedure were confirmed by the research team during a short interview before signing the consent.

2.3. Experimental Design

The Trier Social Stress Test (TSST) Protocol created by Kirschbaum served as the foundation for the experimental design. We adjusted the protocol to consider regional preferences. The TSST standard protocol includes inducing psychological stress and monitoring the body's reaction [25]. The meta-analysis suggested that TSST is a considerably valuable and relevant standardized protocol for examinations of stress hormone reactivity [26, 27]. The modification related to the language and scope focused on providing opportunities with specific needs and how to ensure the interviewer includes them in the list. The task subjects had to complete was applying for a special opportunity through interviews, and the interview conducted in the local language was required. The purpose of using the modified procedure is based on the significant limitation of TSST, which is considered to have a higher degree of habituation of the HPA axis response with repeated exposures [28, 29]. Thus, modification of TSST procedure is needed. The experiment was divided into seven distinct phases and took around 40 minutes.

Figure 1 depicts the entire experiment in detail. To evaluate the physiological condition, we retrieved the pulse rate signal during the rest condition to avoid any disruptions. We categorized the rest condition into four states: baseline rest (R1), anticipatory stress (R2), stressful event (R3), and recovery period (R4) [30, 31]. Baseline rest (R1) was an initial condition to record the participant's current state before enrolling in the experiment. The anticipatory stress (R2) was a rest condition after each participant prepared for the interviews. The stressful event rest (R3) was a rest situation where the main task was held. Each participant met an interviewer and had a conversation. In the end, the recovery period (R4) was a rest state after completing all phases, and the debrief session was also held during that period. The stressful periods in segments R3 and R4 are crucial components of this procedure, where the duration of these stressful events is around ten minutes during the interviews to elicit psychological stress. Throughout the experiment, biosignals from the individuals were captured. To enhance the stress effect, we hired a professional interviewer who often did the main tasks as the interviewer.

2.4. Data Acquisition

From the commencement of the experiment during baseline rest (R1) to the conclusion of the primary task (R4), approximately 40 minutes of biosignals were collected. We utilized data from four different areas (R1, R2, R3, and R4), as shown in Figure 1. Figure 1 also demonstrates that a photoplethysmography (PPG) sensor was used to record biosignals and estimate short-term pulse rate variability (PRV) properties. During the investigation, the PPG sensor was attached to the index fingers of those who participated. We connected the sensor to an Arduino device, which captured the analog data (via the ADC interface), converted it to digital format, and then transmitted it to a computer at a sampling rate of 192 Hz and a transmission baud rate of 9600 bps. After obtaining the raw data, we extracted the features for subsequent analysis through pre-processing (filtering).

2.5. Data Processing and Feature Extraction

We ensure that there is no powerline interference by applying pre- and post-processing to the PRV signals of the participants. We utilized a digital band-pass filter to accommodate the 50–60 Hz frequency of the powerline. In addition, we used a low-pass filter with a cut-off frequency of 20 Hz to eliminate noise from the PPG's pulse rate data. Each parameter can be found in Table 1. We extracted the short-term PRV parameters (5 minutes) consisting of the mean peak-to-peak (R-R) interval (MeanRR), heart-rate change deviations (SDRR/Standard Deviation of RR), NN50, pNN50, square root of the mean squared difference between adjacent R-R intervals (RMSSD), mean heart rate (MeanHR), and standard deviation of heart rate (STDHR). The analysis of these parameters is known as linear time-domain analysis. In addition, we demonstrated linear frequency-domain analysis, which included very low frequency (VLF), low frequency (LF), high frequency (HF), and total power (TotPower). We also derived the LF normalization units (LFnu), HF normalization units (HFnu), and the ratio of LF and HF (LFHF). Pulse rate variability analysis also serves the purpose of non-linear analysis. We utilized the Poincaré analysis by extracting the SD1 and SD2 parameters and then calculating the SD1 and SD2 ratio (SDrat) and elliptical area.

2.6. Psychological Assessments

We created the PANAS scale, which has positive and negative subscales, to evaluate the effects of psychological stress that has been intentionally caused [32]. We assessed respondents' positive (interested, happy, tough, enthusiastic, proud, alert, inspired, diligent, considerate, active) and negative emotions (stressful, disappointed, guilty, fear, unfriendly, angry, embarrassed, nervous, anxious, worry) after the task using the TSST procedure during baseline rest (R1), anticipatory stress (R2), stressful events (R3), and recovery period (R4), which attempts to assess emotional changes in the negative direction or stress. The PANAS-based questionnaire asked sixteen participants to score their emotional state using twenty words covering positive and negative emotions during the TSST experiences. The twenty words were divided into ten descriptors expressing positive and negative sentiments and emotions in each subscale. The following are the five categories used to grade the questions based on the Likert scale: very irrelevant (1), irrelevant (2), neutral (3), relevant (4), and highly relevant (5). The most significant adverse effect suggests representing great psychological tension, and we calculated the questionnaire scores for positive and negative effects to quantify mental stress. The total score for each positive or negative feeling assessment can range from 10 to 50. Lower scoring means the corresponding emotion is less significant.

2.7. Data Analysis

We used the mean and standard error (SE) to present comprehensive data on the PANAS scale questionnaire's score to analyze psychological effects and PRV's parameters data to evaluate physiological traits. We employed the non-parametric Kruskal-Wallis test to identify the significant changes between states because the data would likely be less regularly distributed.

Table 1: The list of equations for the linear and non-linear analysis of pulse rate variability measurement

Linear Methods

$$MeanRR = \frac{1}{n}\sum_{i=1}^{n} RR_i \quad (1)$$

$$SDRR = \sqrt{\frac{\sum_{i=1}^{n} RR_i - MeanRR}{n}} \quad (2)$$

$$NN50 = count(|RR_{i+1} - RR_i|)_{>50ms} \quad (3)$$

$$pNN50 = \frac{count(|RR_{i+1} - RR_i|)_{>50ms} \times 100\%}{n-1} \quad (4)$$

$$RMSSD = \sqrt{\frac{\sum_{i=1}^{N}(RR_{i+1} - RR_i)}{N}} \quad (5)$$

$$MeanHR = \frac{\sum_{i=1}^{N}(60000/RR_i)}{n} \quad (6)$$

$$STDHR = \sqrt{\frac{\sum_{i=1}^{n}\left(\frac{60000}{RR_i}\right) - MeanHR}{n-1}} \quad (7)$$

$$PF = \sum_{k=1}^{K_f} f_k \quad (8)$$

$$VLF < 0.04\ Hz$$
$$LF = 0.04 - 0.15\ Hz, LFnu = LF/(LF + HF)$$
$$HF = 0.15 - 0.40\ Hz, HFnu = HF/(LF + HF) \quad (9)$$

$$LFHF = LF/HF \quad (10)$$

Non-Linear Methods

$$SD1 = \sqrt{var(x_1)}, SD2 = \sqrt{var(x_2)}$$
$$where, x_1 = \frac{\overrightarrow{RR_1} - \overrightarrow{RR_{\iota+1}}}{\sqrt{2}}, x_2 = \frac{\overrightarrow{RR_\iota} + \overrightarrow{RR_{\iota+1}}}{\sqrt{2}} \quad (11)$$

$$EllipArea = \pi \times SD1 \times SD2$$

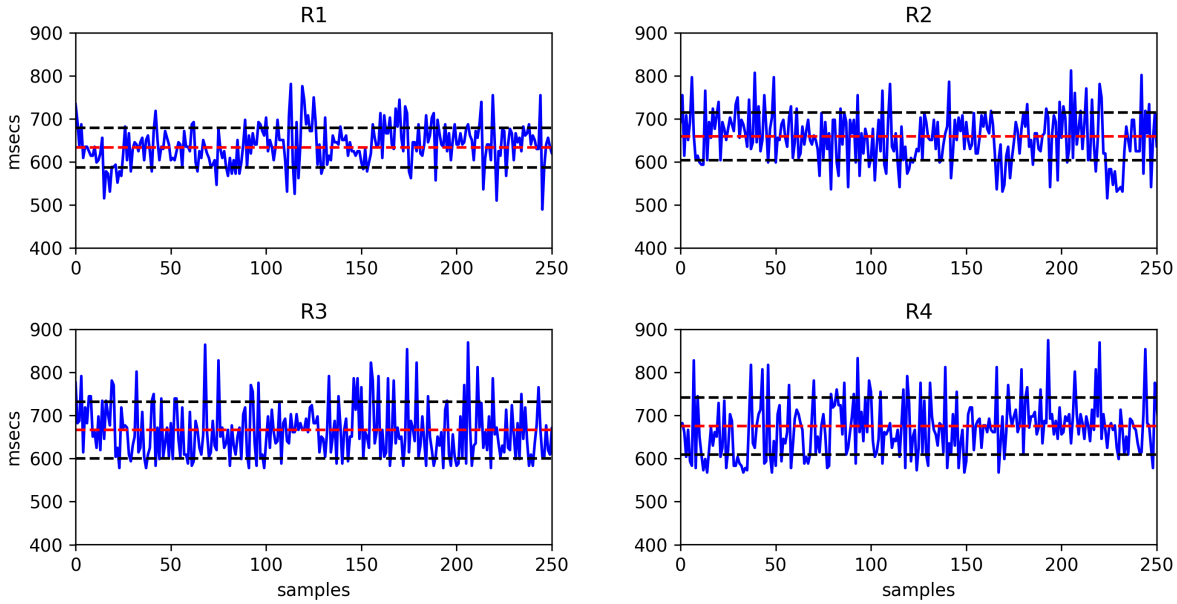


Figure 2: The raw data of peak-to-peak interval (RR) time series data from R1, R2, R3, and R4 on a subject (red dot line : average, black dot line : standard deviation)

statistical test, including the multiple comparisons (posthoc) and correlation test. Python libraries were used to process entire data sets.

3. Results

3.1. Pulse Rate Variability Properties

Firstly, we examined the raw peak-to-peak interval time series to determine how the rest segments during baseline rest (R1), anticipatory stress (R2), stressful event (R3), and recovery period (R4) differed from one another before extracting and analyzing the features from the pulse-rate time series data. The PRV time series for each segment (R1, R2, R3, and R4) from a single subject are displayed in Figure 2. We can see in Figure 2 that there are no significant differences between segments in the peak-to-peak average of PRV values (MeanRR). The MeanRR increases from 633 msec in R1 to 660 msec in R2. During R3 and R4, the MeanRR continued to rise, reaching 666 and 676 msec, respectively. These findings show that the heart beats consistently more slowly after the primary task.

We also found that the gap between intervals expanded from baseline rest (R1) to recovery period (R4) as we observed the standard deviation from the RR array list. The standard deviation for R1 is 46 msec before the primary assignment (interview session), and R2's is 56 msec. R3 and R4 were slightly stable at 66 msec after the preparation phase (R2). According to the data, the heart rate varies more during the stressful events (R3 and R4) than during the preparation portions of baseline rest and anticipatory stress sessions (R1 and R2).

3.2. Psychological Assessment during Stress Induction

Figure 3 illustrates that the positive affect (PA) scores remained constant throughout all rest conditions, with an average score of around 28.59 within the range of 28 to 31 (R1, R2, R3, and R4). These findings indicate that participants were able to continuously regulate their emotions under various experimental conditions. However, individuals' negative effect scores varied amongst the

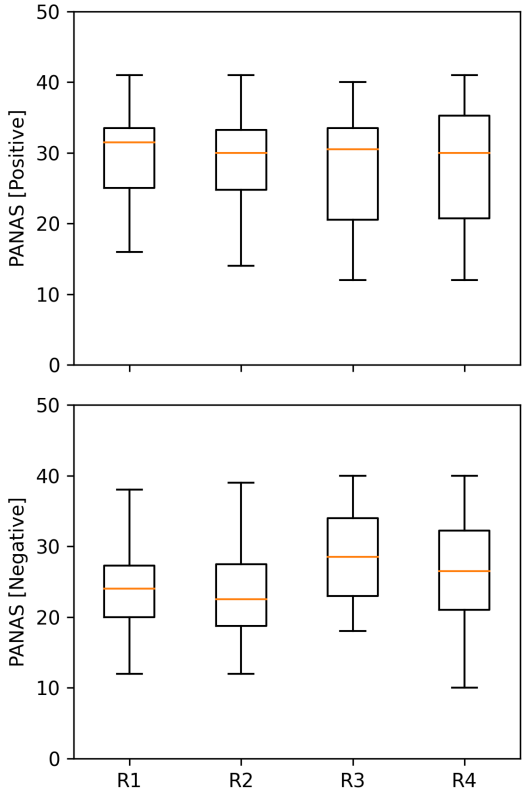


Figure 3: The psychological assessment using the Positive Affect Negative Affect Scale (PANAS) to evaluate the emotional changes PANAS [positive]: positive emotions; PANAS [negative]: negative emotions

Table 2: Summary of feature extraction (linear and non-linear methods) for physiological and psychological metrics

Parameters	R1		R2		R3		R4	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Linear								
MeanRR (msec)	700.86	26.53	708.48	20.32	748.88	20.17	761.67	20.83
SDRR (msec)	56.58	4.35	64.63	3.84	83.20	5.35	82.16	5.47
NN50	108.44	13.89	90.06	12.00	102.00	12.38	126.25	16.28
pNN50 (%)	32.38	4.58	39.18	4.73	55.16	3.86	49.91	3.59
RMSSD (msec)	58.72	6.23	71.68	7.28	109.05	5.75	97.18	6.52
MeanHR (bpm)	88.05	3.31	86.47	2.52	82.02	2.27	80.68	2.35
STDHR (bpm)	7.03	0.53	7.62	0.41	8.85	0.50	8.59	0.54
LF ($msec^2$/Hz)	628.28	107.10	756.24	82.61	1388.12	259.77	1248.16	195.92
HF ($msec^2$/Hz)	596.00	121.83	974.85	182.58	1974.66	281.30	1848.86	328.69
LFHFrat	1.26	0.16	1.37	0.41	0.77	0.09	0.84	0.11
LFnu	52.78	2.86	49.16	3.94	41.08	3.33	42.66	3.31
HFnu	47.22	2.86	50.84	3.94	58.92	3.33	57.34	3.31
TotPow ($msec^2$/Hz)	1685.15	273.24	2293.90	234.91	3974.85	719.27	4016.15	566.44
VLF ($msec^2$/Hz)	460.86	100.28	562.82	93.40	612.08	245.85	919.13	181.71
Non-Linear								
SD1 (msec)	41.52	4.40	50.68	5.15	77.11	4.07	68.71	4.61
SD2 (msec)	67.06	5.37	74.01	4.51	87.83	6.84	92.05	7.12
SDrat	1.86	0.21	1.91	0.39	1.14	0.05	1.38	0.09
EllipArea ($msec^2$)	9293.42	1412.48	12173.47	1761.88	22353.27	3278.50	20738.29	2508.50
Psychological Assessment – Positive Affect Negative Affect Scale (PANAS)								
Positive Affect (PA)	29.25	1.88	28.38	2.06	28.00	2.24	28.75	2.32
Negative Affect (NA)	24.44	1.56	23.19	1.76	28.56	1.70	26.75	2.00

MeanRR: Mean of RR/NN (peak-to-peak) intervals, SDRR: Standard Deviation of RR/NN (peak-to-peak) Intervals, NN50: number of peak-to-peak intervals > 50 msec; pNN50: NN50 over total numbers of peak-to-peak; RMSSD: root mean square of successive differences, MeanHR: Mean of Heart Rate, STDHR: Standard Deviation of Heart Rate, LF: low-frequency, HF: high-frequency, LFHF: ratio of LF over HF; LFnu: low-frequency normalized unit; HFnu: high-frequency normalized unit; TotPow: total power, VLF: very-low-frequency, SD1: standard deviation perpendicular to the line of identity; SD2: standard deviation along the line of identity; SDrat: ratio of SD1 over SD2. EllipArea: elliptical area.

various rest scenarios after stress induction. Although the significance test p-value was higher than 0.05, the scores increased from 23 at the start of the trial to 29 at its completion. There was a difference in the negative effect score between the baseline rest (R1) and stressful event when the assessments were made after the interviews (R3) ($p = 0.72$) and recovery period after the interviews (R4) ($p = 1$). However, no observable differences were found between the stressful period and after interviews (R3, R4), as well as the initial baseline rest (R1) and anticipatory stress period (R2).

These results revealed that the PANAS assessment could reliably identify the increase in negative emotions adhering to stress exposure. The effects of this subjective assessment should be contrasted to physiological data.

3.3. Physiological Assessment during Stress Induction

We performed a short-term PRV analysis for each resting condition (R1, R2, R3, and R4) for 5 minutes. The linear method consists of the time and frequency domain analysis, while the non-linear analysis is based on Poincaré representing the geometrical analysis. We presented the linear and non-linear analysis features in Table 1 and provided the overall parameter summary in Table 2.

We compared the time domain characteristics that we retrieved, such as MeanRR, SDRR, RMSSD, NN50, pNN50, MeanHR, and STDHR, between the baseline rest (R1) and anticipatory stress (R2) to the main task as the stressful event (R3) and recovery period (R4), where the parameters frequently increased. When compared to R1 and R2, where the values were lower, the average of MeanRR during R3 and R4 is higher than 700 msec. As a result, following the interviews, the heart rate decreased to less than 85 beats per minute. The variations in pulse rate variability for SDRR, RMSSD, and STDHR also increased.

We discovered that R3 and R4 had more significant overall frequency components than R1 and R2. In contrast to the low-frequency component, the high-frequency component is more

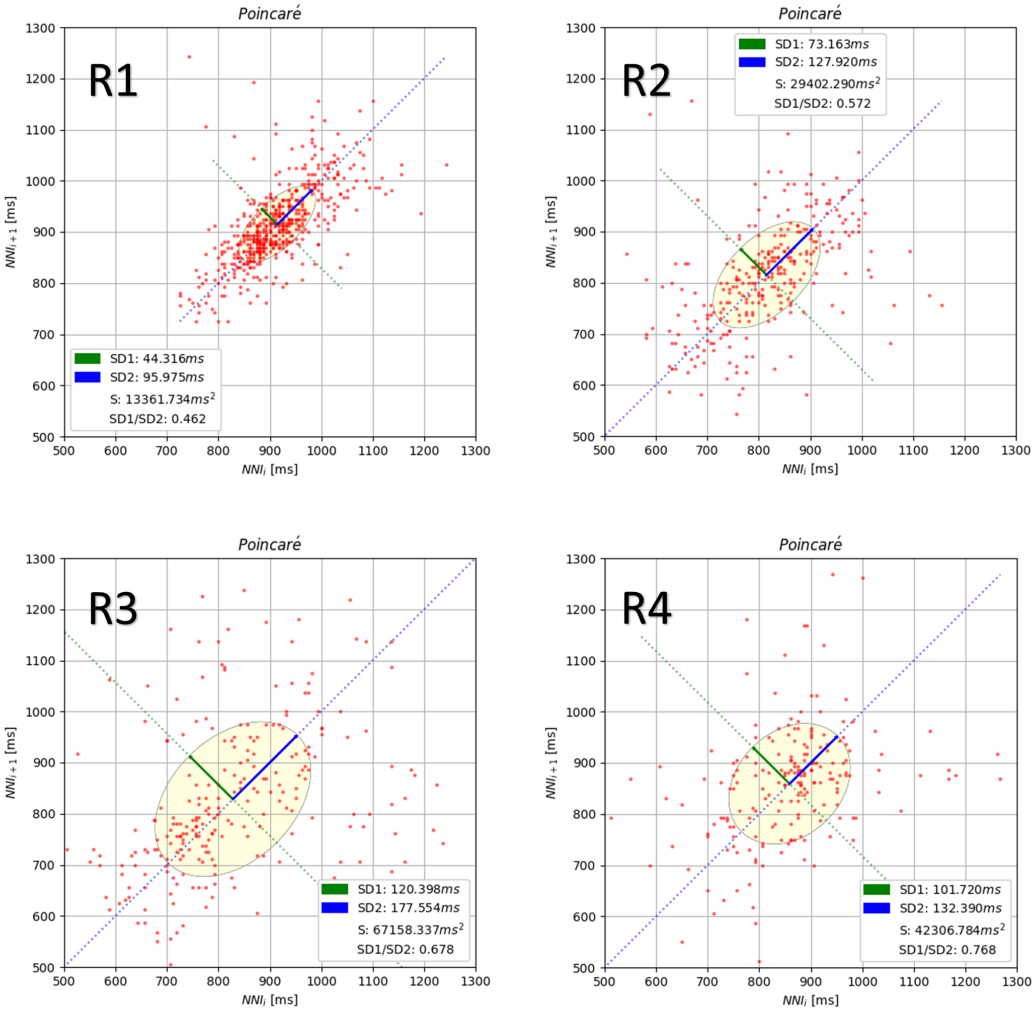


Figure 4: The Poincaré properties (SD1, SD2, and elliptical area) during R1, R2, R3, and R4

interesting. As a result of the interviews, the LF to HF ratio is lower, indicating that during the assignment, sympathetic activity predominated over parasympathetic activity.

Since the parameters, especially SDRR and RMSSD, had comparable characteristics, we discovered that the characteristics of SD1, SD2, and elliptical area are equivalent to the deviation of time series parameters. Figure 4 illustrates Poincaré properties on a topic during R1, R2, R3, and R4. As we can see, following the preparation stages, the distribution of the RR time series is dispersed. The SD1 parameter follows the R2 by more than 100 msec and is followed by a larger elliptical area by more than 30000 msec^2.

To assess the significance of the link between the various resting states (R1, R2, R3, and R4), we evaluated all parameters using both the linear and non-linear methods. Figure 5 demonstrates that the differences between resting state segments are mostly found at the baseline rest (R1) and stressful event (R3), the baseline rest (R1) and recovery period (R4), and the anticipatory stress (R2) and stressful event (R3). From the baseline rest (R1) to the stressful event (R3), the features that distinguish R1-R3 significantly with p-values less than 0.01 are SDRR ($p = 0.0016$), pNN50 ($p = 0.0021$), RMSSD ($p < 0.0001$), LF ($p = 0.0097$), HF ($p = 0.0002$), total power ($p = 0.0029$), SD1 ($p < 0.0001$), SD2 ($p = 0.0420$), SD ratio ($p = 0.0046$), and elliptical area ($p = 0.0005$).

According to the baseline rest (R1) and recovery period (R4) segments, the features that discriminate the two segments (R1-R4) are SDRR ($p = 0.0023$), RMSSD ($p = 0.0026$), LF ($p = 0.0097$), HF ($p = 0.0014$), total power ($p = 0.0015$), SD1 ($p = 0.0026$), SD2 ($p = 0.0077$), and elliptical area ($p = 0.0019$).

Other of the PRV's features that can be utilized to separate between the anticipatory stress segment (R2) and stressful event (R3) are RMSSD ($p = 0.0061$), SD1 ($p = 0.00611$), HF ($p = 0.0471$), and elliptical area ($p = 0.0166$). The elliptical area can also differentiate the anticipatory stress (R2) and recovery period (R4) with a p-value of 0.0471.

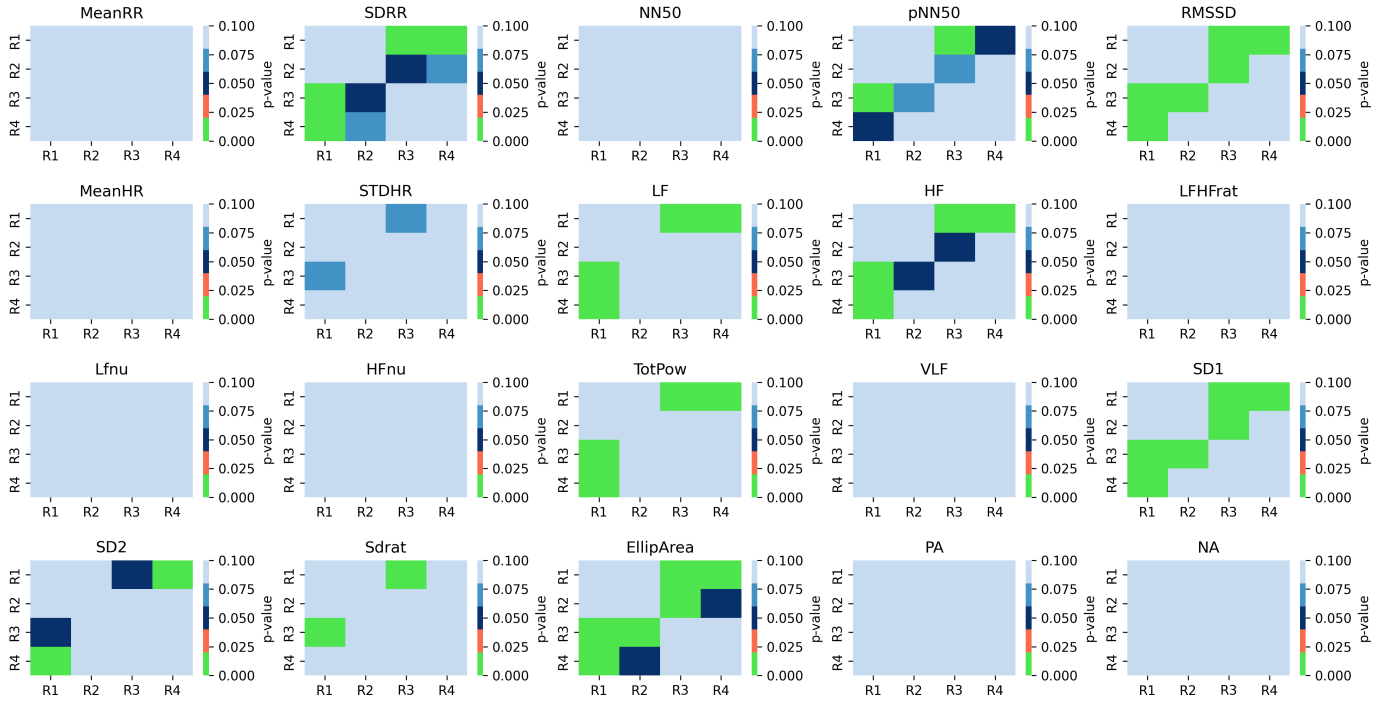


Figure 5: Multiple-comparison statistical test to evaluate the physiological and psychological evaluation under conditions R1, R2, R4, and R4. The color of the heatmap represents the p-value for each pair of comparisons

In general, with various PRV parameters that can distinguish two segments (R1-R3, R1-R4, R2-R3, and R2-R4), the most sensitive physiological parameters are RMSSD, HF, SD1, and elliptical area ($p < 0.001$). Finally, we found that the statistical results point out that the baseline rest (R1) and stressful event (R3) are considered to be able to elicit psychological changes, followed by the baseline rest (R1) and stressful periods (R4).

3.4. Correlation between the physiological and psychological measurement

We performed a correlation analysis to determine the relationship between the negative emotions measured by the PANAS scale and the overall linear and non-linear PRV properties. Since the negative effects of the PANAS questionnaire raise differences between segments, PRV features tend to follow the changes in subjective evaluation in the opposite direction. We established a correlation analysis using all segments and on each segment of R1, R2, R3, and R4. We combined the PRV features values from the overall segment and correlated them to the negative effects of PANAS. We found no correlation between the PRV features and negative scores. We also employed the correlation analysis on each segment and found that only the recovery period (R4) presents a good negative correlation on low-frequency component (LF) (correlation = -0.6; $p = 0.0142$) and total power (correlation = -0.5; $p = 0.0412$). We concluded that an increasing negative affect score represents a lower power of frequency components on physiological measurement during a specific segment (recovery period/R4).

4. Discussion

Our proposed study employs heart rate variability (HRV) to measure physiological changes induced by purposefully inducing mental stress. We used a low-cost portable photoplethysmograph (PPG) sensor to retrieve the pulse rate and extract the pulse rate variability (PRV) properties as the surrogate for HRV, which necessitates a heart rate sensor on the chest. To see how the physiological traits relate to the psychological changes, we fully exploited the features of PRV using linear and non-linear analysis. Instead of having participants perform commonly used particular tasks, we developed a modified Trier Social Stress Test (TSST) protocol to create a naturally stressful situation. To further our understanding, we compared the findings with those from earlier studies and confirmed several results.

Pulse rate variability (PRV) is considered a tool to surrogate heart rate variability (HRV) [33, 34]. HRV changes are already known as part of the autonomic nervous system modulation changes [35]. A psychological stimulus triggers an autonomic response in panic disorder, including flushing, tachycardia, palpitations, hypertension, and gastrointestinal symptoms. The autonomic response may occasionally disappear by repeatedly exposing ourselves to the stimuli in comforting conditions. Salivation, stomach motility, and acid secretion can all be induced by food-related thoughts. When stress activates the sympathetic nervous system, norepinephrine is mainly released at the synaptic junction with the enteric nervous system, which decreases GI motility [36]. Therefore, using PRV's parameters is considered a meaningful way to represent stress and how physiological changes affect the autonomic nervous system.

The RMSSD, SD1, and elliptical area of the PRV were the parameters that were most sensitive to differentiate between before (baseline rest (R1) and anticipatory stress (R2)) and after inducing stress (stressful event (R3) and recovery period (R4)) ($p < 0.01$), followed by SDRR, pNN50, LF, HF, and total power ($p < 0.05$). We discovered that the negative emotions that occurred corresponded to these parameters. The peak-to-peak interval was primarily observed to decrease in stressful circumstances in prior studies [37]. Another study demonstrated that the heart rate of HRV rose in response to feelings of relaxation and fear, but no other measures showed any discernible variations [38]. In our opinion, stress or other negative circumstances are more closely related to changes in HRV than changes in heart rate. Therefore, a declining heart rate (lower peak-to-peak interval) might not indicate psychological adjustments. Otherwise, it could be a good idea to investigate the variance of the peak-to-peak interval in the future.

According to the feature extraction, the PRV variants SDRR, RMSSD, and SD1 better differentiate between psychological states before and after stressful events (R3). The high-frequency component (HF), which reflects sympathetic activity, best represents the SDRR, RMSSD, and SD1 [39]. Before the interview session, as the stressful events, the baseline rest (R1) and anticipatory stress (R2) demonstrate that mental tension and negative feelings were present and extended before the main stress events (R3) and recovery period (R4). Our data demonstrate that the changes of PRV had lower values before the stressful period during the main task (R1, R2) than after the interview (R3, R4). It implies that the primary task (assessing anxious sensations) is completed first, followed by mental tension and negative emotions verified through a questionnaire. As a result, the physiological aspect becomes more crucial to analyze as an objective measurement and takes precedence over the subjective assessment.

The PANAS scale can only measure psychological shifts in positive and negative directions while recognizing positive and negative feelings through a numerical score. Although emotion and stress can be associated, it is essential to consider whether the results of this study represent changes in emotion or stress itself because they showed the same negative sentiments as prior reports when stress was induced [3]. Additionally, the study we suggest offers fresh perspectives on assessing each component of happy and negative emotions to provide a helpful study of mental stress needed to create an objective physiological measuring system. The findings of this study imply that although stress and particular unpleasant emotions are connected and share similar physiological characteristics, alterations in physiological signals during stress induction cannot be utilized alone to quantify stress directly. However, stress symptoms, such as alterations in negative emotions, are the primary driver of physical changes and should be investigated to identify the emotions that lead to mental stress.

Based on our results of proposed study, we can confirm that that the stressors are affecting the changes in HRV properties, even though we are using PRV as a surrogate [21,40]. The variability of PRV's properties represents the changes in both linear and non-linear parameters corresponding to psychological changes. However, we observed that the stressful event (R3) tends to increase the PRV's parameters (RMSSD, pNN50, and elliptical area), whereas previous studies reported that the properties tend to decrease [18]. During the stress-induced procedure using the TSST protocol, the cardiovascular characteristics change before the stressful event (R1/R2), where they have lower feature properties. The features rise after the anticipatory stress (R3/R4), which means that our study does not contradict the previous findings because the lower HRV properties were found before the stressful events and started to increase during the recovery period after inducing the TSST procedure. To strengthen the findings, we need further studies to establish a more extended observation period where we utilize the daily stress recording and assessment.

5. Conclusion

During stress induction, the modified TSST protocol reliably induced negative emotions in subjects, as demonstrated by this study. The majority of short-term PRVs for SDRR, pNN50, RMSSD, LF, HF, total power, SD1, and elliptical area exhibited increased activity in correlation with negative emotion levels ($p < 0.01$). This study implies that negative emotions, believed to be a symptom of mental stress, increased immediately from the baseline condition following the first interview session (after inducing stress). In addition, significant differences were observed only between the onset of the resting state or baseline rest (R1) and after the interview as stressful events (R3). As adverse effects increase, the parameters demonstrate an increase in sympathetic activity based on their overall characteristics. In addition, the undesirable physiological properties observed were identical to the mental stress-related alterations reported in previous studies. Future research should examine the detection of mental tension from the viewpoint of adverse emotions.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The Directorate of Research and Community Service, Universitas Islam Indonesia, Yogyakarta, as an Excellence Interdisciplinary Research Program with grant number 014-Dir-DPPM-70-PUPT-PIII-XII-2018, financially funded this work. We also want to express our highest gratitude to the participants who have willingly joined this study.

References

[1] J. Du, J. Huang, Y. An, and W. Xu, "The Relationship between stress and negative emotion: The Mediating role of rumination," Clinical Research and Trials, **4**(1), 2018, doi: 10.15761/crt.1000208.

[2] J. Lee and S.K. Yoo, "Recognition of Negative Emotion Using Long Short-Term Memory with Bio-Signal Feature Compression," Sensors, **20**(2), 573, 2020. doi: 10.3390/s20020573.

[3] C. Degroote, A. Schwaninger, N. Heimgartner, P. Hedinger, U. Ehlert, and P. H. Wirtz, "Acute stress improves concentration performance: Opposite effects of anxiety and cortisol, "Experimental Psychology, **67**(2), 88-98, 2020. doi: 10.1027/1618-3169/a000481.

[4] H. Selye H, The stress of life. McGraw-Hill, New York, 1956.

[5] American Psychological Association, "Stress in America: The State of Our Nation," Stress in America TM Survey, 2017.

[6] S. Singh, and A. Vats," Impact of physiological and psychological stress on students, "Headache, **9**(4), 111-113, 2020.

[7] R. S. Lazarus, Psychological Stress and the Coping Process. Mcgraw-Hill, 1966.

[8] J. K. Kiecolt-Glaser, L. McGuire, T. F. Robles, and R. Glaser, "Emotions, Morbidity, and Mortality: New Perspectives from Psychoneuroimmunology," Annual Review of Psychology, **53**(1), 83–107, 2002, doi: 10.1146/annurev.psych.53.100901.135217.

[9] E. M. Sternberg, "The Stress Response and the Regulation of Inflammatory Disease," Annals of Internal Medicine, **117**(10), 854, 1992, doi: 10.7326/0003-4819-117-10-854.

[10] A. Seiler, C. P. Fagundes, and L. M. Christian, "The impact of everyday stressors on the immune system and healthm," In Stress Challenges and Immunity in Space, Springer, Cham, 71-92, 2020.

[11] S. Gamonal-Limcaoco, E. Montero-Mateos, M. T. Lozano-López, A. Maciá-Casas, J. Matías-Fernández, and C. Roncero, "Perceived stress in different countries at the beginning of the coronavirus pandemic," The International Journal of Psychiatry in Medicine, 009121742110337, 2021, doi: 10.1177/00912174211033710.

[12] C. Joaquim, T. Luzia, H. A. Michael, P. M. O Fransisco, D. A. Silvia, A. Pedro, S. H. Aaron, S. Berta, "Negative affect and stress-related brain metabolism in patients with metastatic breast cancer," Cancer, **126**(13), 3122-3131, 2020. doi: 10.1002/cncr.32902.

[13] O. Bălan, G. Moise, L. Petrescu, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "Emotion Classification Based on Biophysical Signals and Machine Learning Techniques," Symmetry, **12**(1), 21, 2020. doi: 10.3390/sym12010021.

[14] D. E. Eagle, J. A. Rash, L. Tice, and R. J. Proeschold-Bell, "Evaluation of a remote, internet-delivered version of the Trier Social Stress Test," International Journal of Psychophysiology, 2021, doi: https://doi.org/10.1016/j.ijpsycho.2021.03.009.

[15] C. Hartling, Y. Fan, A. Weigand, I. Trilla, M. Gärtner, M. Bajbouj, and S. Grimm, "Interaction of HPA axis genetics and early life stress shapes emotion recognition in healthy adults," Psychoneuroendocrinology, **99**, 28-37, 2019. doi: 10.1016/j.psyneuen.2018.08.030.

[16] L.M. Glenk, O. D. Kothgassner, A. Felnhofer, J. Gotovina, C. Pranger, A. N. Jensen, and E. Jensen-Jarolim, "Salivary cortisol responses to acute stress vary between allergic and healthy individuals: the role of plasma oxytocin, emotion regulation strategies, reported stress and anxiety," Stress, **23**(3), 275-283, 2020. doi: 10.1080/10253890.2019.1675629.

[17] M. Cohen, and R. Khalaila, "Saliva pH as a biomarker of exam stress and a predictor of exam performance," Journal of psychosomatic research, **77**(5), 420-425, 2014. doi: 10.1016/j.jpsychores.2014.07.003.

[18] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis and M. Tsiknakis, "Review on psychological stress detection using bio-signals," IEEE Transactions on Affective Computing, 2019. doi: 10.1109/TAFFC.2019.2927337.

[19] E. Mejía-Mejía, K. Budidha, T.Y. Abay, J.M. May, and P.A. Kyriacou, "Heart rate variability (HRV) and pulse rate variability (PRV) for the assessment of autonomic responses," Frontiers in physiology, **11**, 779, 2020. doi: 10.3389/fphys.2020.00779.

[20] A. Sch fer, and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability? A review on studies comparing photoplethysmographic technology with an electrocardiogram," Int. J. Cardiol., 166, 15–29, 2013. doi: 10.1016/j.ijcard.2012.03.119.

[21] H. M. Morales-Fajardo et al., "Towards a Non-Contact Method for Identifying Stress Using Remote Photoplethysmography in Academic Environments," Sensors, **22**(10), 3780, 2022, doi: https://doi.org/10.3390/s22103780.

[22] S. Hey, A. Gharbi, B. Von Haaren, K. Walter, N. König, and S. Löffler, "Continuous noninvasive pulse transit time measurement for physiological stress monitoring" In Proc. 2009 International Conferencce on eHealth, Telemedicine, and Social Medicine, Cancun, Mexico, 2009, 113-116. doi: 10.1109/etelemed15088.2009.

[23] M. Zubair and C. Yoon, "Multilevel mental stress detection using ultra-short pulse rate variability series", Biomedical Signal Processing and Control, **57**, 2020. doi: 10.1016/j.bspc.2019.101736.

[24] A. Sahroni, N. Widiasmara, I. Miladiyah, and H. Setiawan, "Short-Term Pulse Rate Variability to Measure Changes in Emotion during Trier Social Stress Test," 2022 9th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), 2022, doi: 10.1109/icitacee55701.2022.9923994.

[25] C. Kirschbaum, KM. Pirke, DH. Hellhammer, "The trier social stress test a tool for investigating psychobiological stress responses in a laboratory setting", Neuropsychobiology, 28, 76-81, 1993.

[26] M. A. Birkett, "The Trier Social Stress Test Protocol for Inducing Psychological Stress," Journal of Visualized Experiments, 56, 2011, doi: https://doi.org/10.3791/3238.

[27] J. A. Seddon et al., "Meta-analysis of the effectiveness of the Trier Social Stress Test in eliciting physiological stress responses in children and adolescents," Psychoneuroendocrinology, **116**, 104582, 2020, doi: https://doi.org/10.1016/j.psyneuen.2020.104582.

[28] J. C. Pruessner, J. Gaab, D. H. Hellhammer, D. Lintz, N. Schommer, and C. Kirschbaum, "Increasing correlations between personality traits and cortisol stress responses obtained by data aggregation," Psychoneuroendocrinology, 22(8), 615–625, 1997, doi: https://doi.org/10.1016/s0306-4530(97)00072-3.

[29] N. C. Schommer, D. H. Hellhammer, and C. Kirschbaum, "Dissociation Between Reactivity of the Hypothalamus-Pituitary-Adrenal Axis and the Sympathetic-Adrenal-Medullary System to Repeated Psychosocial Stress," Psychosomatic Medicine, **65**(3), 450–460, 2003, doi: https://doi.org/10.1097/01.psy.0000035721.12441.17.

[30] S. Schlatter, L. Schmidt, M. Lilot, A. Guillot, and U. Debarnot, "Implementing biofeedback as a proactive coping strategy: Psychological and physiological effects on anticipatory stress," Behaviour Research and Therapy, 140, 103834, 2021, doi: https://doi.org/10.1016/j.brat.2021.103834.

[31] A. P. Allen, P. J. Kennedy, S. Dockray, J. F. Cryan, T. G. Dinan, and G. Clarke, "The Trier Social Stress Test: Principles and practice," Neurobiology of Stress, **6,** 113–126, 2017, doi: https://doi.org/10.1016/j.ynstr.2016.11.001.

[32] Brosschot, J. F., Thayer, J. F.: Heart rate response is longer after negative emotions than after positive emotions. International journal of psychophysiology. **50**(3), 181-187, 2003.

[33] A. K. Verma, P. N. Aarotale, P. Dehkordi, J.-S. Lou, and K. Tavakolian, "Relationship between Ischemic Stroke and Pulse Rate Variability as a Surrogate of Heart Rate Variability," Brain Sciences, 9(7), 162, 2019, doi: https://doi.org/10.3390/brainsci9070162.

[34] E. Gil, M. Orini, R. Bailón, J. M. Vergara, L. Mainardi, and P. Laguna, "Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions," Physiological Measurement, **31**(9), 1271–1290, 2010, doi: https://doi.org/10.1088/0967-3334/31/9/015.

[35] J.-M. Grégoire, C. Gilon, S. Carlier, and H. Bersini, "Autonomic nervous system assessment using heart rate variability," Acta Cardiologica,1–15, 2023, doi: https://doi.org/10.1080/00015385.2023.2177371.

[36] M. G. Ziegler, "Psychological Stress and the Autonomic Nervous System," Primer on the Autonomic Nervous System, 291–293, 2012, doi: 10.1016/b978-0-12-386525-0.00061-5.

[37] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature," Psychiatry Investigation, **15**(3), 235–245, 2018, doi: 10.30773/pi.2017.08.17.

[38] M. T. Valderas, J. Bolea, P. Laguna, M. Vallverdu, and R. Bailon, "Human emotion recognition using heart rate variability analysis with spectral bands based on respiration," 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, doi: 10.1109/embc.2015.7319792.

[39] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," Frontiers in Public Health, **5**(258), 2017, doi: 10.3389/fpubh.2017.00258.

[40] K. Yoo and W. Lee, "Mental stress assessment based on pulse photoplethysmography," 2011 IEEE 15th International Symposium on Consumer Electronics (ISCE), 2011, doi: https://doi.org/10.1109/isce.2011.5973841.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 2, 71-77 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Detecting the Movement of the Pilot's Body During Flight Operations

Yung-Hsiang Chen[1,2], Chen-Chi Fan[1], Jin H. Huang[*,3]

[1]*Ph.D. Program of Mechanical and Aeronautical Engineering, Feng Chia University, Taichung, 407102, Taiwan*

[2]*Aeronautical Systems Research Division, National Chung-Shan Institute of Science and Technology, Taichung, 407102, Taiwan*

[3]*Department of Mechanical Engineering, Feng Chia University, Taichung, 407102, Taiwan*

ARTICLE INFO

Article history:
Received: 03 January, 2023
Accepted: 22 February, 2023
Online: 11 March, 2023

Keywords:
3D Space Trajectory
Aircraft Cockpit Design
Markerless Motion
Multi-view Image

ABSTRACT

This research presents a "Multi-camera for pilot's cockpit measurement system", which uses four multi-view images to eliminate the instrument and human body shielding and record the touched area. That could record the body reaction time (velocity and acceleration) and trajectory of the tested personnel. Real-time conversion of multi-view images corresponding to the 3D skeletal joint coordinate information of the human body, which measure the human-computer interaction human factors engineering integration of limb reaction time and trajectory measurement system. Finally, make prototypes, test and optimize, and achieve the research on the optimal cockpit touch area by conducting multi-view image simulation feasibility experiment framework and measurement process method. Using multiple depth-sensing cameras to perform low-cost, standardized automatic labeling of human skeleton joint dynamic capture.

1. Introduction

Considering human factors engineering design is the foundation of workplace planning and the basis for product design development. Its anthropometric data is very important for work and daily life, and the correct anthropometric data is of great help to workplace design, work planning and human-machine interface safety considerations. Due to the rapid development of camera technology in recent years, the image resolution has also been greatly improved. It can achieve a certain accuracy in experimental measurement applications. It can also be used for non-contact optical measurement with image analysis software. The human motion tracking technologies are included marker/markerless. Markerless motion applications are such as sports science. For example, in [2], the author presented captured of an athlete for sport research. The basic concept of digital photogrammetry is to locate the specific punctuation point of the measurement position in the image, compare the measurement punctuation position at different times, and then obtain the displacement of the measurement point. Camera measurement can get good measurement results as long as the punctuation points in the image are clearly visible. The image analysis method has the advantages of non-contact measurement method, and the influence of the local environment on the measurement is also very small. Therefore, this research chooses to apply the camera to develop the metrology technology.

In [3], the author presented the DIC (Digital Image Correlation) method with 3D stereo vision. The DIC method is applied to 3D deformation measurement. The 3D-DIC method is applied to the surface profile measurement research. Whether 3D-DIC or other non-contact image measurement systems is composed of stereo cameras. The distance and angle between two cameras overlap within a certain range. When the measurement range exceeds the visible range of the two image capture devices, only partial images of the object can be captured. The camera must be adjusted according to the different shapes of the object, and the image capture device must be re-installed. The range of camera angle of view and the need for continuous calibration of the adjusted camera are application constraints. It is necessary to simplify and repeat the definition of the relationship between the coordinates of the calibration measurement system and the object to be measured.

In order to solve the aforementioned problems, in [4], the author presented a two-axis parallel motion mechanism for

* Corresponding Author: Jin H. Huang, Department of Mechanical Engineering, Feng Chia University, Taichung, Taiwan, email: jhhuang@fcu.edu.tw

" This paper is an extension of work originally presented in 2022 IEEE International Conference on Consumer Electronics - Taiwan [1]"

vertical and horizontal motion. The device uses a servo motor as a positioning control, and is mounted on a camera device. That can precisely control the camera's imaging angle. Both X and Y axes operates are independently, and will not become the load of another servo motor for image capture and analysis. Using three cameras to build a semi-circular geometric multi-camera imaging platform system [5]. Those multi-camera are set with a semi-circular measuring rod and the optical axes of the multi-cameras co-intersect at a point in space. The mechanism can be used to adjust the spacing with the semi-circular measuring rod. With the superimposition feature of multiple cameras, the system can increase the range of overlapping areas of 3D reconstruction feature points and the correction parameters established inside. It's required for on-site measurement and reconstruction of 3D information. The field of view and simplify the procedure of repeatedly defining and determining the relationship between the coordinates of the calibration measurement system and the object to be measured.

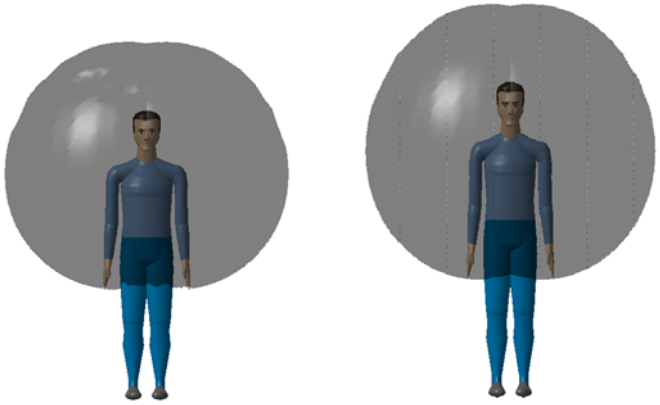
The image analysis techniques and photogrammetric image acquisition are precisely and quickly measure. The 3D-based machine vision in metrology is popular. The aircraft pilot's cockpit design needs to meet pilots' feasibility in 5~95% of body figuration. MIL-STD-1333 [6] is defined the pilot's reach zone and MIL-STD-1472 [7] is defined schematic diagram of the pilot's cockpit reach zone area. The analysis of actions related for the pilot's reach zone. That is necessary to detect the human head and movement of the upper and lower limbs.

The conventional technology can only measure the static human body of a single person, and it is necessary to manually adjust the human body measurement data multiple times, manually select the measurement mark points, and manually edit and superimpose the 3D human body model. This innovative technology provides dynamic human trajectory calculation reaction time, and can automatically mark human bone joints and superimpose standardized 3D human joint coordinates.

Most studies discuss pilots' eye gaze strategies, is lack of pilot's movement monitoring with optical tracking in the aircraft cockpit [8-12]. The aim of this research evaluate human-computer interaction and human factors engineering on multi-camera for pilot's cockpit measurement system. Using multiple depth-sensing cameras to perform dynamic capture and record the reaction time and trajectory of the subject's limbs, including real-time conversion of multi-view images corresponding to the 3D skeletal joint coordinate information of the human body.

2. Research methods

Fig. 1 is the aircraft pilot's cockpit design for pilots' feasibility in 5~95% of body figuration. When pilot wear a flight suit, comfort, easy operation and visual display, are the requirement to be concerned. And the maneuverability of control and display of information have to be effectively functioned under the constant high G (gravity) force condition. At present, the evaluation of pilots' human factors engineering, is used the research of physical model and interview as the basis for the adjustment of the aircraft cockpit design. The information presents from the research is difficultly analyzed and takes lots of time and manpower. The common improper design of human factors engineering are showed as: over-sized helmet, close instrument panel, the pilot body touches the control lever, the hand is too short to reach the control lever, and the pilot's long foot touches the instrument panel.



(a)5% (b)95%

Figure 1: The aircraft pilot's cockpit design for pilots' feasibility in 5~95% of body figuration

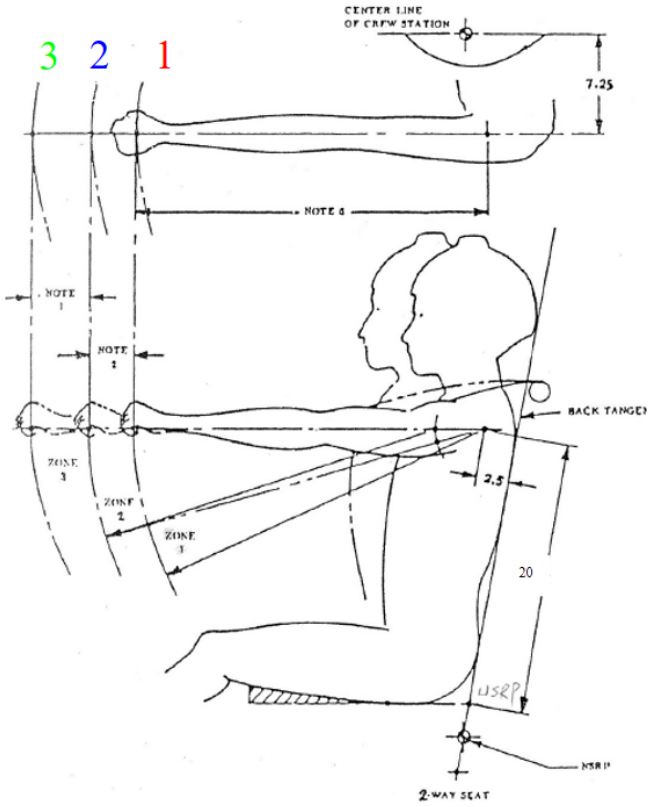


Figure 2: MIL-STD-1333 [6] is defined the pilot's reach zone.

Fig. 2 is MIL-STD-1333 [6], that defined pilot's reach zone. The pilot reach zones are divided into three areas. (1) ZONE 1 [Natural arm-reach mechanism zone (shoulder strap locked)]: Both hands can operate naturally to reach in ZONE 1, while shoulders are attached to the back of the seat, and arm and shoulder muscles do not need to be stretched forcefully. (2)ZONE 2[Maximum arm-reach mechanism zone (shoulder strap locked)]: The maximum arm-reach mechanism zone is defined while arm and shoulder muscles are stretched to the maximum. All the critical control devices in flight are included within ZONE 2.

(3)ZONE 3 [Maximum arm-reach mechanism zone (shoulder strap unlocked)]: The maximum mechanism arm-reach zone is defined while shoulders are as far forward as possible and arms fully reach. All the none-critical and none-essential control devices in flight are included within ZONE 3.

Fig 3 is MIL-STD-1472 [7] defined the touched area of aircraft cockpit. The motion of pilot's head and upper limbs are needed to detect in accord with the analysis of pilot's motion around touched area. To record and analyze the posture the pilot's upper body muscles touching the cockpit reach zone and motions around area, we proposed to use four cameras to build up a multi-view image. The pilot executes the instrument operation according to the instructions of different subjects. The system analyzes where each instrument should be located in tested pilot's touching area by image processing and motion detection of human upper limb.

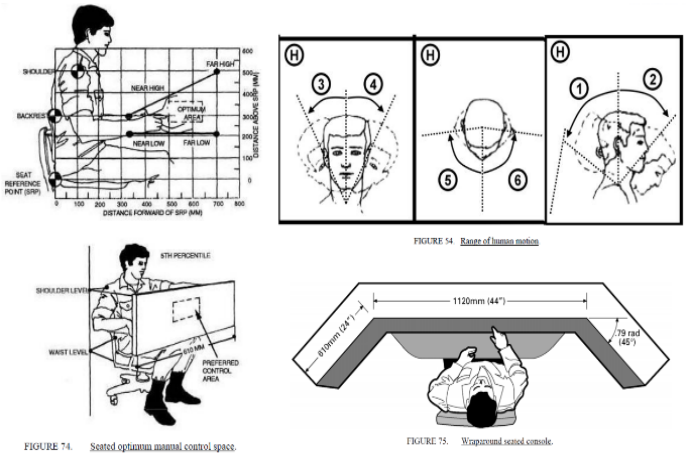
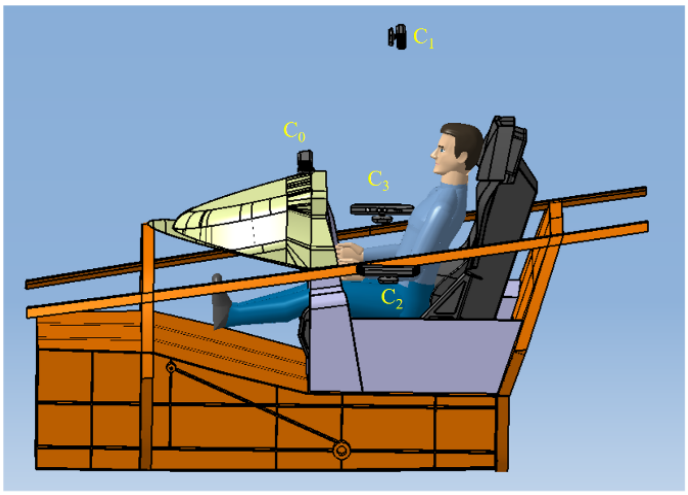


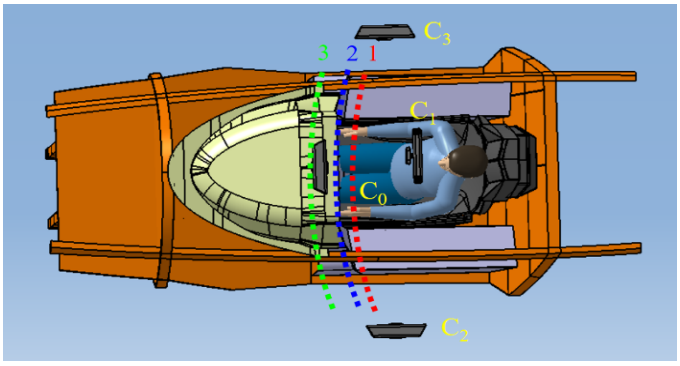
Figure 3: MIL-STD-1472 [7] defined the touched area of aircraft cockpit

Fig. 4 is the schematic diagram of multi-view image measurement system setup with human factors engineering. First, simulate the pilot operation in different positions and height of instrument panel. Then, utilize the concept of human skeleton detection to measure the position which the pilot reads or touches the instrument. Subsequently, develop an automatic measure evaluation system in pilot's reachable areas, with the utilization of multi-view imaging by four cameras and 3D machine vision method.

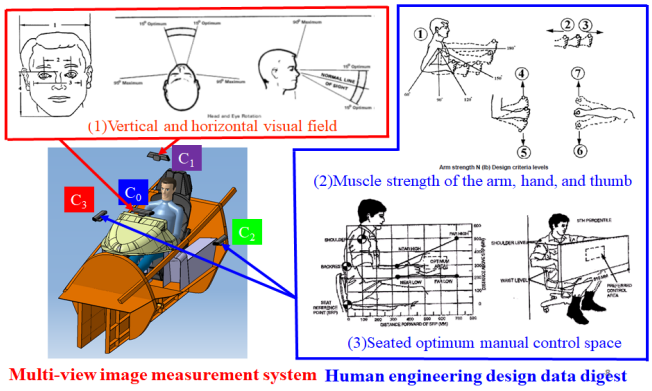
Fig. 5 is the multi-camera for pilot's cockpit measurement system. Using multiple depth image sensors to measure and track the operation of cockpit personnel with optical measurement technology, that measure pilot's reaction time and trajectory. Where, C_0 is placed in front of the pilot for pilot's head and eyes. C_1 is above the pilot's seat. The head is defined as the origin of the measurement. The reachable area of the upper limb in the X-Y plane is recorded in different conditions. Both C_2 and C_3 are arranged on the left and right sides of the pilot respectively. That is exclude the shielding of instruments and the human body and record the contact area of the X-Z plane. This technology proposes multi-view images to correspond to human body 3D skeletal joint coordinate information and explores human-computer interaction human factors engineering integration of limb reaction time and trajectory measurement system.



(a)

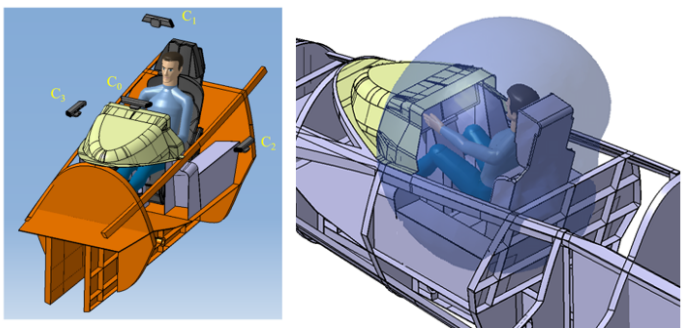


(b)

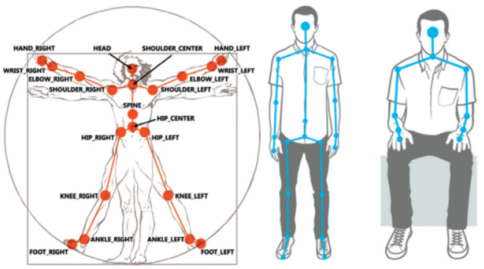


(c)

Figure 4: Schematic diagram of multi-view image measurement system setup with human factors engineering



(a) Multi-camera for measured pilot's reaction time and trajectory



(b) human body 3D skeletal joint coordinate information

Figure 5: The multi-camera for pilot's cockpit measurement system.

3. Experimental result

In order to verify the feasibility of a multi-view measurement system, the schematic diagram of simulated multi-view image setup is showed as Fig. 6. Where, fig 6(a) is a top view, Fig. 6(b) is a right view and Fig. 6(c) is a front view. Fig. 7 is the simulated multi-view image, in which Fig. 7(a) uses a hemispherical calibration target for 3D analysis, and the simulation result is shown in Fig. 7(b). The C_0 image is placed in front of the pilot to measure the head and eyes.

Fig. 7(c) is the C_1 image is on the top of the seat. The pilot's head is defined as the measurement reference point. The upper limbs in the X-Y plane can be touched under different constraints with MIL-STD-1333 definition. Figs. 7(d) and (e) are C_2 and C_3 image respectively, which are installed on the left and right sides of the pilot. Those exclude the cockpit instruments and human body shielding and record the touched area on the X-Z plane.

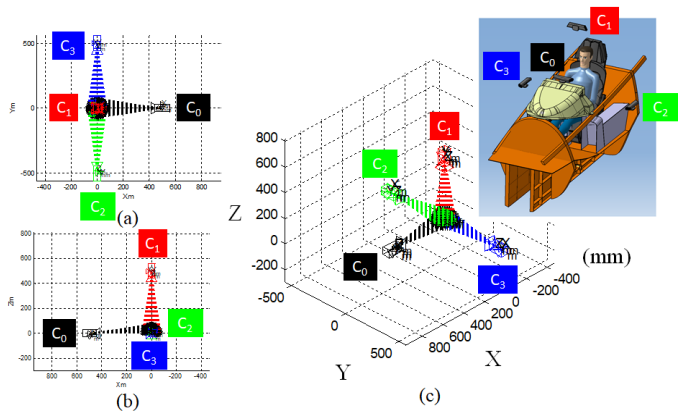


Figure 6: Schematic diagram of simulated multi-view image setup

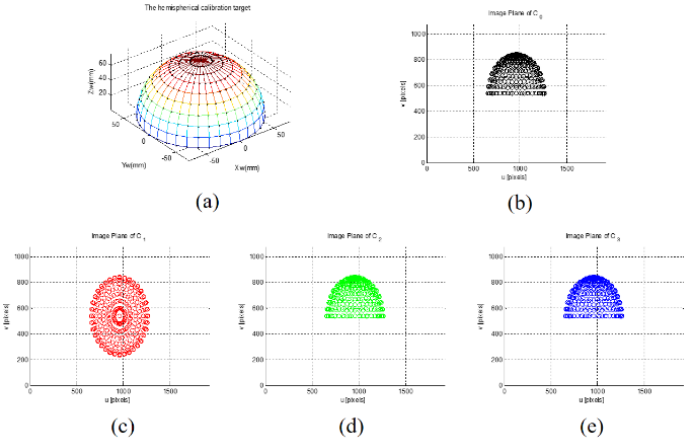
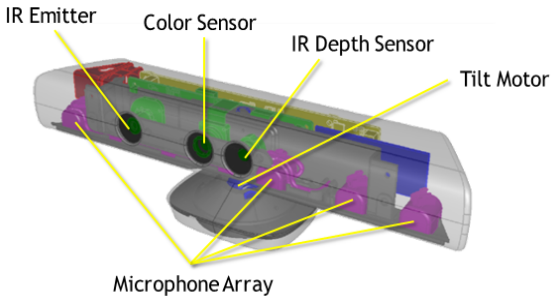


Figure 7: Simulated multi-view image

Fig. 8 shows that the Kinect camera includes a motor and color image (RGB camera in the middle). Those are include a 3D depth image (two lenses on the left and right), an infrared emitter, an infrared CMOS camera and sound (array microphone). 3D machine vision human skeleton detection results, as shown in Fig. 9. The Kinect camera defines the human skeleton joint coordinate map. This research selected Kinect camera to cost reduction. The Kinect v2 depth camera is cheap than high speed camera, and the measurement accuracy is mostly under 2 mm at 1.5 m [14]. Those condition would be meet our measurement system demonstration. The system utilizes four multi-view cameras to perform dynamic capture and record the reaction time and trajectory of the subject's limbs, including real-time conversion of multi-view images corresponding to the 3D skeletal joint coordinate information of the human body based on references. The proposed method have been practically tested in Figure 11. In Fig. 11 is captured 3D skeletal joint of the human body with Kinect.

(a) Physical diagram



(b) System architecture diagram

Figure 8: Kinect camera [13]

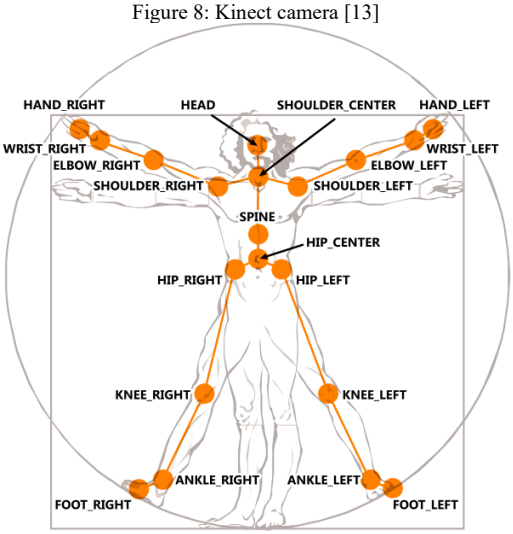
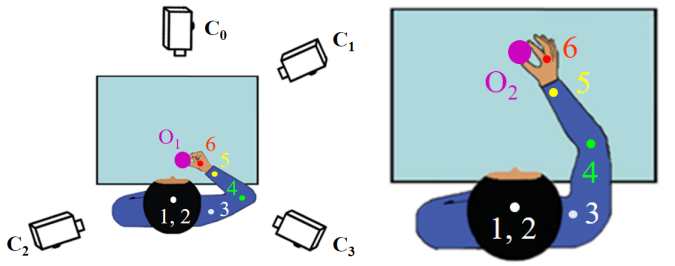


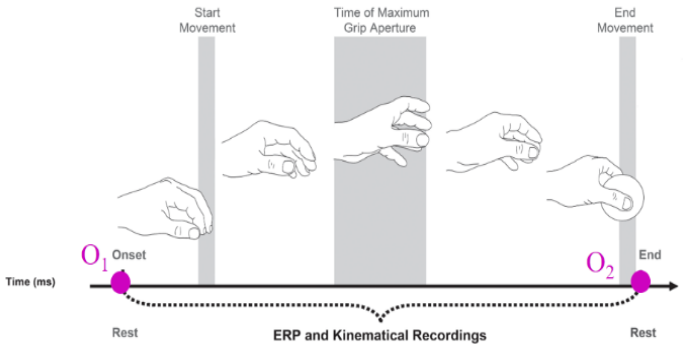
Figure 9: The Kinect camera defines the human skeleton joint coordinate map [13]

Pilot will isn't wear any equipment in our system, which is direct collection the joints information of the human skeleton. Use Kinect to obtain the coordinates of each joint point of the human body in space. The distance formula and vector inner product formula to calculate the distance and bending angle of the human joint points. Kinect is cheap, easy to carry, and used to evaluate the reaction time of the body of the test personnel. Due to the complex movements of the subject or the influence of environmental factors, occlusion may occur. Multi-angle shooting through multiple Kinects can successfully solve such problems. The integration of information collected by multiple Kinects can reduce occlusion. Real-time measurement and on-site operation testing analysis the cockpit instruments and human body.

Fig. 10 is the "Reach to Grasp" action sequence diagram. The six human skeleton joint coordinates verify the feasibility of the human reaction time and trajectory calculation of the cockpit personnel operating the measurement tracking system. Fig. 11 illustrates captured 3D skeletal joint coordinate information of the human body with Kinect

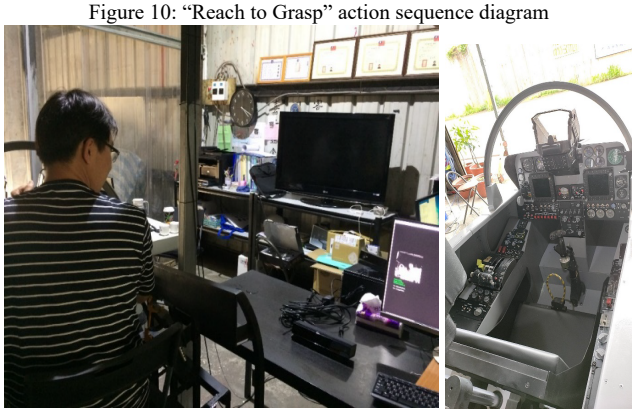


(a) Start (b) End

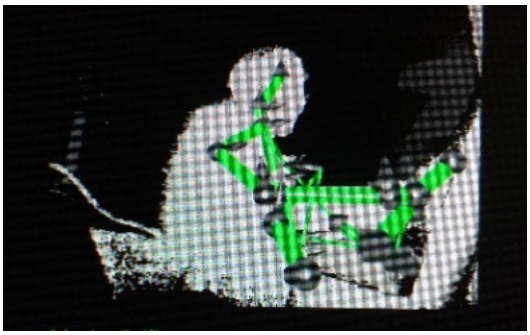


(c) "Reach to Grasp" action sequence

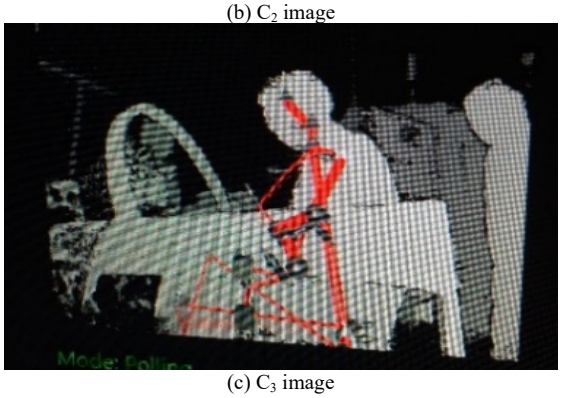
Figure 10: "Reach to Grasp" action sequence diagram



(a) Experimental setup



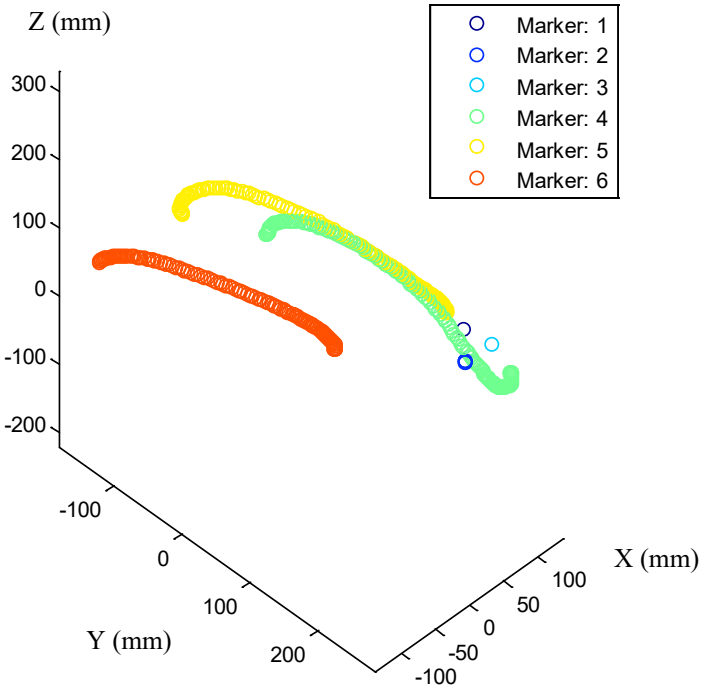
(b) C_2 image



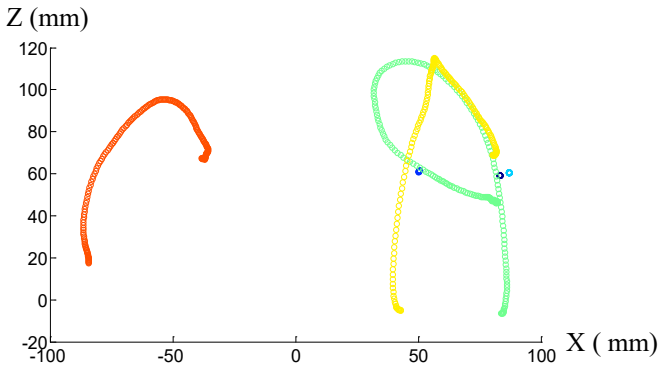
(c) C_3 image

Figure 11: Captured 3D skeletal joint coordinate information of the human body with Kinect

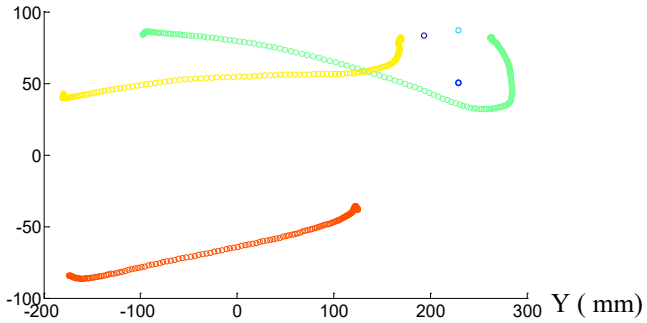
Fig. 12 is the 3D space trajectory of pilot's skeleton joints. Fig. 13 is the reaction time of the right arm of the pilot's body. Among them, the action of "Reach to Grasp" moves from point O_1 to point O_2 at 485 ms. The starting action is within 0.4-1.4 seconds, the maximum grasping action starts, the elapsed moving time MT1: 790 ms and the highest moving speed APV1: 1165 mm/s. Finally, the moving speed of the right arm feature points #4~#6 exceeds 800 mm/s, and the rest of the head and shoulder feature points #1~#3 are almost still.



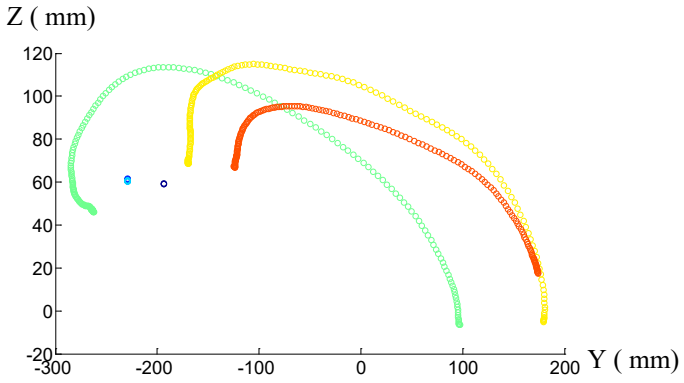
(a) 3D space trajectory



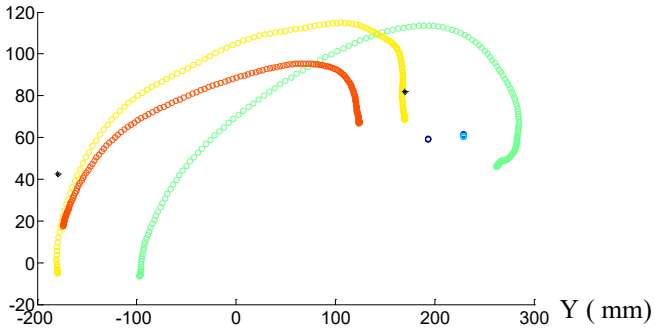
(b)C_0 image trajectory (front view)



(c)C_1 image trajectory (top view)



(d)C_2 image trajectory (left view)



(e)C_3 image trajectory (right view)

Figure 12: The 3D space trajectory of pilot's skeleton joints

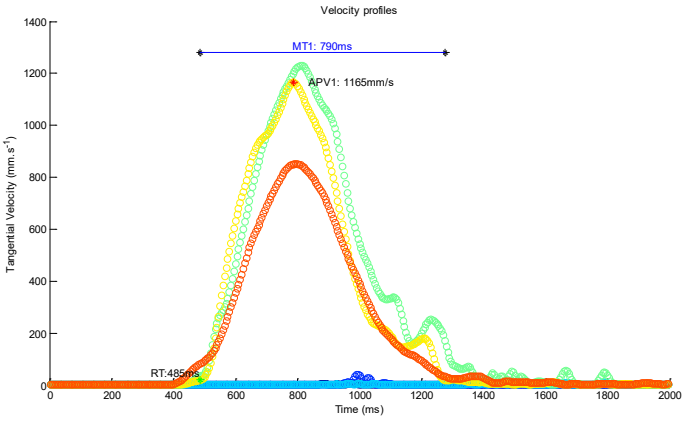


Figure 13: The reaction time of the right arm of the pilot's body.

4. Conclusion

The study successfully proposes a measurement and tracking system for movements of the pilot's body during flight operations. In order to overcome the limitation of image occlusion, the "Omni-directional Multi-view Image Measurement System" is used. The intersecting area can be increased by multiple cameras for 3D reconstruction. Measurement system flowchart is record skeleton detection of posture and touch area trajectory data, and then 3d data fitting with four multi-Kinect analysis. This system uses four multi-view images to eliminate the instrument and human body shielding and record the touched area. That could record the body reaction time (velocity and acceleration) and trajectory of the tested personnel. Real-time conversion of multi-view images corresponding to the 3D skeletal joint coordinate information of the human body, which measure the human-computer interaction human factors engineering integration of limb reaction time and trajectory measurement system. Finally, make prototypes, test and optimize, and achieve the research on the optimal cockpit touch area by conducting multi-view image simulation feasibility experiment framework and measurement process method. Using multiple depth-sensing cameras to perform low-cost, standardized automatic labeling of human skeleton joint dynamic capture.

Conflict of Interest

The authors declare no conflict of interest.

Reference

[1] Y.H. Chen, J. H. Huang, "Measurement and tracking system for movement of the pilot's body during flight operations," 2022 IEEE International Conference on Consumer Electronics - Taiwan, 539-540, 2022. DOI: 10.1109/ICCE-Taiwan55306.2022.9868976

[2] E. Kruk, M.M. Reijne, "Accuracy of human motion capture systems for sport applications; state-of-the-art review", European Journal of Sport Science, **18**(6), 806-819, 2018. DOI: 10.1080/17461391.2018.1463397

[3] P.F. Luo, Y.J. Chao, M.A. Sutton, W. H. Peters, "Accurate measurement of three-dimensional deformations in deformable and rigid bodies using computer vision", Experimental Mechanics, **33**, 123-132, 1993.

[4] Y.H. Chen, Y.S. Shiao, "Control simulation and experiment of two-axis parallel kinematic mechanism", 2016 Taiwan Power Electronics Symposium, 2006.

[5] C.H. Hwang, W.C. Wang, Y.H. Chen, "Camera calibration and 3D surface reconstruction for multi-camera semi-circular DIC system", Proc. SPIE 8769, International Conference on Optics in Precision Engineering and Nanotechnology (icOPEN2013), 2013.

[6] MIL-STD-1333B, "Aircrew station geometry for military aircraft", 09 JAN, 1987.

[7] MIL-STD-1472H, "Human Engineering", 15 SEP, 2020.

[8] O. Lefrançois, N. Matton, M. Causse, "Improving airline pilots' visual scanning and manual flight performance through training on skilled eye gaze strategies, Safety, **7**(4), 2021. DOI: 10.3390/safety7040070
[9] R. Li, B. Jumet, H. Ren, W. Song, ZTH Tse, "An inertial measurement unit tracking system for body movement in comparison with optical tracking", Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, **234**(7), 728-737, 2020. DOI: 10.1177/0954411920921695
[10] P. Biswas, "Using eye gaze controlled interfaces in automotive environments", Springer, 2016.
[11] M.G. Glaholt, "Eye tracking in the cockpit: a review of the relationships between eye movements and the aviators cognitive state", Defence Research and Development Canada Scientific Report DRDC-RDDC-2014-R153, 2014.
[12] E. Machida, M. Cao, T. Murao, H. Hashimoto, "Human motion tracking of mobile robot with Kinect 3D sensor", 2012 Proceedings of SICE Annual Conference (SICE), 2207-2211, 2012.
[13] www.developer/microsft.com/windows/kinect/.
[14] G. Kurillo, E. Hemingway, M.L. Cheng, L. Cheng, "Evaluating the accuracy of the azure kinect and kinect v2," Sensors (Basel). **22**(7), 2469, 2022. DOI: 10.3390/s22072469

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 2, 78-86 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Day-Ahead Power Loss Minimization Based on Solar Irradiation Forecasting of Extreme Learning Machine

Adelhard Beni Rehiara[*,1], Sabar Setiawidayat[2], Frederik Haryanto Sumbung[3]

[1]*Electrical Engineering Department, University of Papua, Manokwari, Indonesia*

[2]*Electrical Engineering Department, University of Widyagama, Malang, Indonesia*

[3]*Electrical Engineering Department, University of Musamus, Merauke, Indonesia*

ARTICLE INFO

Article history:
Received: 27 December, 2022
Accepted: 28 February, 2023
Online: 11 March, 2023

Keywords:
ELM
Forecasting
NASA
Solar irradiation
Power loss

ABSTRACT

Power losses exist naturally and have to be cared for in the operation of electrical power systems. Many researchers have worked on various methods and approaches to reduce losses by incorporating distributed generators (DG), particularly from renewable sources. These studies are based on the maximum unit penetration of the DGs, which is rarely achieved, resulting in inaccurate calculations. This paper proposes an advanced solution for calculating power losses by incorporating an Extreme Learning Machine (ELM) method for forecasting the solar irradiation. The ELM algorithm was used to create a model for forecasting solar radiation in the Manokwari region and its surroundings. Daily solar radiation in the region has been predicted using the model. NASA's 8016 data on temperature and solar irradiation were used to train the ELM model. With an MAE value of around 0.6392 and a training time of 4.4375 seconds, the test results demonstrate that the built model has good accuracy. The operation of a 1000 kWp solar power plant based on the ELM data forecasting can reduce the power loss of the existing distribution network around the location from 1.5095 kW/hour to 0.9068 kW/hour. Furthermore, the power plant operation can minimize the power loss by 39.9249 percent, from 36.2280 kW to 21.7640 kW.

1. Introduction

Solar energy is the bedrock of alternative energy sources since it may be used to develop other renewable energies [1]. In addition, photochemical power operations and other physical processes both heavily rely on solar energy. The earth's surface will receive solar energy via a radiation mechanism. The radiation will be filtered into the atmospheric layer, which contains gaseous substances as well as other solid forms including water vapour, dust, and aerosols, before it reaches the surface of the planet [2]. This filtration procedure has lowered the solar radiation's intensity to the point where it will not damage earthly life. The geographic position of a region on the earth's surface affects the amount of solar energy that reaches that region [3]. As a result, there are three ways to gather information about the intensity of solar radiation in a given area: directly using pyranometers, pyrheliometers, and Campbell Stokest measuring instruments; using satellite image data; and numerical simulation through computer modeling to determine the potential for future radiation. However, due to a lack of measuring tools, there is still extremely little and difficult-to-access sun irradiation data available in Indonesia. Since satellite image data is internationally available, it is widely used [4].

The escalation of world energy demand and the need for environmentally sustainable development have attracted human attention to renewable energy sources [5]. Among the renewable energy sources, solar power has become increasingly prevalent as a utility-scale renewable energy supply due to its simplicity of installation and maintenance. Furthermore, solar technology has matured, and mass manufacturing of PV panels as the core component of a solar power plant has reduced the cost. Thus, solar power integration into the grid is on the rise, and facilities might be present in the coming years [6].

Power losses in an electric power grid are caused by the current flowing in the conductor lines, where long conductor

[*]Corresponding Author: Adelhard B. Rehiara, Email: a.rehiara@unipa.ac.id

lines with large loads will result in significant power losses. However, this power loss cannot be avoided, so the only thing that can be done is to minimize the loss. Some of the methods that have been offered in previous research to minimize the losses include power injection through flexible alternating current transmission systems (FACTS) [7,8], distributed generators (DGs) [9–13], and optimal sizing and placement of capacitors [14–16], so that the voltage profile can be corrected, which results in minimal power losses.

In this paper, a model for forecasting solar irradiation was created for use in the Manokwari region and its surrounds, which are located at latitude -0.8457 and longitude 134.0504. The model was created using an extreme learning machine (ELM) algorithm that is modified from a neural network to improve learning speed. The model is used to forecast the local daily solar irradiation intensity. The data is then used to calculate the potency for minimizing power losses through the interconnection of a 1 MWp solar power plant to an existing power grid around the location. Then the objectives of this paper are given as: 1) designing a suitable model for the specific area based on the ELM algorithm, 2) forecasting the day-ahead solar irradiance for that location, and 3) investigating the possibility of power loss minimization in a power grid based on the forecasting data. This paper is an extension of work originally presented at the CyberneticsCom conference [1].

The frame of the paper is structured as follows: Section II provides problem statement. Section III describes the research approach and the methods used. The data source, results, and explanations of the findings are presented in Section IV. The paper's primary research findings are summarized in Section V.

2. Problem Statement

Previous research on power losses has been argued in [17], who have developed the Power Voltage Sensitivity Constant (PVSC), which has been developed to determine the location and size of multiple DG units so that active power losses in a distribution system can be reduced. The method was tested on an IEEE 69 reconfigured bus system under three different total loads and heavy conditions. It can be concluded from the results that the proposed methodology gives maximum loss reduction while considering DG size. Authors in [18] presented the use of the multiobjective cuckoo search (MOCS) algorithm to strategically place the unified power flower controller (UPFC) in order to reduce transmission losses. For the multiobjective issue under consideration, the pareto-optimal technique is used to extract the pareto-optimal solution. The best compromise option is extracted from the set of pareto-optimal solutions using the fuzzy logic method. A typical IEEE 30 bus test setup is used to evaluate the suggested method. The results show that the MOCS algorithm is relatively more effective in optimizing the multiobjective problem. Authors in [19] presented a multiobjective optimization methodology to optimally place a STATCOM in electric power distribution networks. Total cost and power loss are the objectives to be met so that the STATCOM placement can minimize these issues. The combination of multiobjective ant colony optimization (MACO) and the bacterial foraging optimization algorithm (BFOA) successfully solved the problems of minimization by testing the method's effectiveness in a 5-bus system. Authors in [20] proposed a method to control droop optimization strategically for inverter control of the islanded microgrid operation, which includes PV penetration and battery energy systems. Two-level controls are used to achieve maximum power loss minimization. A perturbation and observation (P&O) method is used to make the droop functions more adaptable. Load and inverter capacity are changed in three cases to check the effectiveness of the proposed method. Under various loading scenarios and system topologies, simulation studies have shown that it is capable of minimizing microgrid power loss while maintaining frequency and voltage stability. The fundamental disadvantage of this approach is that it only relies on the blind exploration of unknown functions, which degrades performance as the complexity of the grid increases and makes it possible that complex power systems will not achieve the lowest global power loss.

An extreme learning machine (ELM) is a single-layer feed-forward neural network (FFNN), which is a family of artificial neural networks (ANN). Implementing a single hidden layer minimizes the computation while minimizing the FFNN structure. This simplification can shorten calculation time by thousands of times while overcoming the FFNN's primary learning speed drawback. ELM is one of the most well-liked model-based systems because of the continued high accuracy of the system. The use of ELM algorithms is widespread in control [21,22], diagnosis [23,24] and forecasting [4,25]. When it comes to particular forecasting, ELM excels at predicting the solar irradiation of the Lamongan and Muara Karang regions [4], and the stock performance of PT. Telkom [25].

The existing methods have provided outstanding solutions for power loss minimization in the power grid through the penetration of solar power. However, the methods are based on maximizing the penetration of solar PV. This condition is not fully true, since solar irradiation is very dependent on weather conditions. Therefore, this disadvantage can be addressed by incorporating solar irradiation forecasting into the calculation of optimal power flow. The proposed method is capable of achieving detailed power loss minimization, which improves system reliability.

3. Methods

3.1. Extreme Learning Machine

A single hidden layer feed-forward neural network (SLFN), which has an effective training method, is the basis of an Extreme Learning Machine (ELM) [21]. This brilliance results from the absence of iteration in ELM. However, the hidden layer requires more neurons to provide effective prediction. The right number of neurons in the hidden layer of the ELM must also be determined using the trial-and-error method [4].

The ELM algorithm allows for the weightings to be selected at random without necessarily adjusting the input weights and hidden layer biases. The hidden layer output matrices can then be used to perform a generalized inverse operation to determine the SLFNs' output weight. The time for both training and performance has been shortened by this approach [21].

For a given n training set samples (x_j, t_j) where $x_j=[x_{j1}, x_{j2},..., x_{jn}]^T$ and $t_j=[t_{j1}, t_{j2},..., t_{jn}]^T$, an SLFN with N hidden neurons and an activation function $g(x)$ is expressed as [1,21,26]:

$$o_j = \sum_{j=1}^{n} \beta_i \, g\left(w_i x_j + b_i\right), \quad i = 1,2,\ldots,N \tag{1}$$

where $w_i=[w_{i1}, w_{i2},..., w_{in}]^T$, $\beta_i=[\beta_{i1}, \beta_{i2},..., \beta_{in}]^T$, b_i, and o_j are the connecting weight of the i-th hidden neuron to the input neuron, the connecting weights of the i-th hidden neuron to the output neurons, the bias of the i-th hidden node, and the actual network output with respect to input x_j respectively.

The standard SLFN can minimize the deviation between t_j and o_j, so that (1) can be rewritten as follows.

$$t_j = \sum_{i=1}^{N} \beta_i \, g\left(w_i x_j + b_i\right), \quad j = 1,2,\ldots,n \tag{2}$$

The simplification of (2) is $T=H\beta$ and the output weight matrix β can be solved using the least squares method as specified in (3).

$$\beta = H^{\dagger} T \tag{3}$$

The completed forms of the network output matrix T and hidden layer output matrix H are given as follows.

$$H = \begin{bmatrix} g\left(w_1 x_1 + b_1\right) & \cdots & g\left(w_N x_1 + b_N\right) \\ \vdots & \ddots & \vdots \\ g\left(w_1 x_n + b_1\right) & \cdots & g\left(w_N x_n + b_N\right) \end{bmatrix} \tag{4}$$

$$T = \left[t_1^T, \ldots, t_n^T\right]^T \tag{5}$$

3.2. Solar Power Plant

The sun emmits electromagnetics radiation with an effective temperature of 5777 K. Radiant energy, the ammount of energy within the radiation, is spread over the time and measured as radiant power also known as irradiance. The radiant power that reach the Earth surface is normaly measured in square meter through a pyranometer. Extraterrestrial irradiance (EXT) is very observable due to the absence of interference when the radiance traveling thru the space. The amount of irradiance reaching the earth's surface is approximately 1361.1 W/m^2, which is also known as global horizon irradiance (GHI) expressed in terms of clear-sky irradiance and is one of the most useful variables to calculate when working with solar data. This variable represents the best scenario for a photovoltaic system since it represents the maximum that could be received on the day, resulting in uninterrupted generation [2].

A solar power plant has at least one solar panel to convert energy directly from the sun. An inverter is also required to meet the output needs, and some accessories are involved to protect the process. The energy from a solar power plant depends on solar irradiation, while the constant parameters in calculating output power are the covered area and efficiency of the solar PV modules. Therefore, the output power of a solar power plant is formulated as follows [27].

$$P_{PV} = A_c N_{PV} S_{ir} \eta \tag{6}$$

where P_{PV}, A_c, Sir, and η are output power (W) dimension of solar PV modules (m^2), number of PV modules, solar irradiance (W/m^2) and efficiency of the solar PV modules.

3.3. Forward Backward Method

The voltage, power, and power loss in a radial distribution network are calculated in this research using the forward backward approach. This approach was created by reference [29] as it is described in their paper. By figuring out three fundamental computations, this method expands on the Distflow method to finish the analysis of power flow in distribution networks. The computations are used to determine the voltage magnitude, active power, and reactive power.

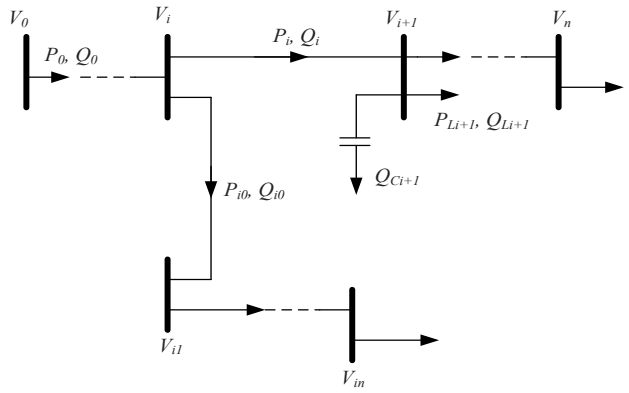


Figure 1: Single line diagram of a general distribution system

Consider the balanced three-phase, radial distribution feeder in Figure 1 with n branches/nodes and nc shunt capacitors. Without a branch from V_k to V_{kn} on the figure, the power and voltage at each node from V_0 to V_n can be calculated by the following equations [28–30].

$$P_{i+1} = P_i - \left(r_{i+1} \frac{P_i^2 + Q_i^2}{V_i^2} \right) - P_{Li+1} \tag{7}$$

$$Q_{i+1} = Q_i - \left(x_{i+1} \frac{P_i^2 + Q_i^2}{V_i^2} \right) - Q_{Li+1} + Q_{Ci+1} \tag{8}$$

$$V_{i+1}^2 = V_i^2 - 2\left(r_{i+1} P_i + x_{i+1} Q_i\right) + \left(r_{i+1}^2 + x_{i+1}^2\right) \frac{P_i^2 + Q_i^2}{V_i^2} \tag{9}$$

where P_{Li+1} and Q_{Li+1} are loads connected at node $i+1$; P_{i+1} and Q_{i+1} are the effective real and reactive power flows from node $i+1$; r_{i+1} and x_{i+1} are resistance and capacitance of the line leading to bus $i+1$; P_i and Q_i are real and reactive power flowing from bus i; V_i and V_{i+1} are the voltage magnitudes on bus i and bus $i+1$; and Q_{Ci+1} is an additional reactive power of the capacitor on the bus $i+1$.

The aforementioned equations are referred to as the forward equation, and each procedure is referred to as a forward update. Other calculating techniques and sequences in the Distflow are known as "backward updates" and "backward equations." The initial values of P_0 and Q_0 in this approximation method are determined by summing the active and reactive power of loads. V_0 is the initial voltage, which is also utilized as the system's base voltage in the approach method's calculation procedure [29]. The parameters in the bracket of equations (7) and (8) denote active and reactive power losses. At the end of each branch, the active and reactive power should be equal to zero [28–30].

3.4. Optimal Power Flow

An important method for improving a system's performance and lowering operational costs is optimization. A power system subject to dispersing loads amongst power plants is called optimal power flow (OPF). It is possible to decrease the overall fuel cost of all committed plants while still adhering to network limits. The OPF problem was generally expressed as follows [31–33]:

$$\text{Min } f(x, u) \tag{10}$$

subjected to

$$g(x, u) = 0 \tag{11}$$

$$h(x, u) \leq 0 \tag{12}$$

Where f, g and h are the objective functions, the equality constraints represent power flow equations and the system operating constraints, respectively.

The vector x is the dependent variable and consists of the voltage magnitude of load buses, the phase angle of all buses except that of the slack bus, the active power of the slack generator, and generators' reactive power. The vector of x is also called a state variable, and it is formulated in the following equation:

$$x^{\mathrm{T}} = [\mathrm{P_{G1}}, \mathrm{V_{L1}}, \ldots, \mathrm{V_{LNL}}, Q_{\mathrm{G1}}, \ldots, Q_{\mathrm{GNC}}, \mathrm{S_{l1}}, \ldots, \mathrm{S_{ITL}}] \tag{13}$$

The vector u is a control variable that includes the active output power of generators at generator buses P_G, the terminal voltage magnitude at generation bus bars V_G, the output of shunt VAR compensators Q_C, and the tap setting of the tap regulating transformers T. Therefore, the vector u can be modeled as follows:

$$u^{\mathrm{T}} = [\mathrm{P_{G1}}, \ldots, \mathrm{P_{GNG}}, \mathrm{V_{G1}}, \ldots, \mathrm{V_{GNG}}, \mathrm{Q_{C1}}, \ldots, \mathrm{Q_{CNC}}, \mathrm{T_1}, \ldots, \mathrm{T_{NT}}] \tag{14}$$

where S_l is transmission line loading and the subscripts NL, TL, NG, NC, and NT denote the number of load buses, transmission lines, generators, shunt VAR compensators, and regulating transformers, respectively.

The objective function in an OPF problem is to minimize the generators' costs in the operations. The cost function represents the relationship between operating costs and output power as expressed in equation (15), where a_i, b_i, and c_i are the coefficients of the fuel cost model.

$$F = \sum_{i=1}^{\mathrm{NG}} a_i P_{Gi}^2 + b_i P_{Gi} + c_i \tag{15}$$

There are certain equality and inequality restrictions in the problem of optimal power flow. The power flow equations are represented by the equality constraints g, as computed as follows:

$$\mathrm{P_{Gi}} - \mathrm{P_{Di}} = \sum_{j=1}^{NB} \left(\mathrm{V_i V_j Y_{ij}} \cos\left(\theta_{\mathrm{ij}} + \delta_{\mathrm{j}} - \delta_{\mathrm{i}}\right) \right) \tag{16}$$

$$\mathrm{Q_{Gi}} - \mathrm{Q_{Di}} = -\sum_{j=1}^{NB} \left(\mathrm{V_i V_j Y_{ij}} \sin\left(\theta_{\mathrm{ij}} + \delta_{\mathrm{j}} - \delta_{\mathrm{i}}\right) \right) \tag{17}$$

Where P_{Gi} and Q_{Gi} are the i-th generator's active and reactive power; P_{Di} and Q_{Di} are the i-th bus's active and reactive demand. V_i and V_j are the voltage magnitudes at buses i and j; Y_{ij} models the element of the admittance bus matrix at row i and column j; NB is the number of buses; δ_i, δ_j, and θ_{ij} are the angels of V_i, V_j, and Y_{ij}, respectively.

The inequality constraint h limits the physical devices as well as system security. The inequality constraints consist of generator constraints in equations (18) to (20), transformer constraints in equation (21), and shunt capacitor constraints in equation (22). System security constraints, on the other hand, are given in equations (23) and (24).

$$P_{Gi}^{\min} \leq P_{Gi} \leq P_{Gi}^{\max} \quad i = 1, \ldots, NG \tag{18}$$

$$Q_{Gi}^{\min} \leq Q_{Gi} \leq Q_{Gi}^{\max} \quad i = 1, \ldots, NG \tag{19}$$

$$V_{Gi}^{\min} \leq V_{Gi} \leq V_{Gi}^{\max} \quad i = 1, \ldots, NG \tag{20}$$

$$T_i^{\min} \leq T_i \leq T_i^{\max} \quad i = 1, \ldots, NT \tag{21}$$

$$Q_{Ci}^{\min} \leq Q_{Ci} \leq Q_{Ci}^{\max} \quad i = 1, \ldots, NG \tag{22}$$

$$V_i^{\min} \leq V_i \leq V_i^{\max} \quad i = 1, \ldots, NL \tag{23}$$

$$S_{ij} \leq S_{ij}^{\max} \quad i = 1, \ldots, TL \tag{24}$$

where P_{Gi}^{min} and P_{Gi}^{max} are the minimum and maximum active power of i-th generator and Q_{Gi}^{min} and Q_{Gi}^{max} denote the minimum and maximum reactive power of i-th generator; S_{ij} and S_{ij}^{max} are line flow and maximum line flow between bus i and j.

3.5. Model Validation

The developed model needs to be evaluated in order to detect and prevent any potential problems that might appear after running the ELM model. The mistakes will be eliminated through the validation process, making the model sufficiently precise for the simulation to match reality as predicted. The model is verified

using the mean absolute error (MAE). It can be expressed as follows [4]:

$$MAE = \frac{\sum_{i=1}^{Y} |\hat{y}_i - y_i|}{Y} \quad (25)$$

where $\hat{y}_i$, y_i and Y are forecasting data, real data and number of samples, respectivelly.

4. Result and Discussion

4.1. Datasets

NASA [34] provided the datasets used to train the ELM model, which are located at latitude -0.8457 and longitude 134.0504. The data is based on Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) that starts to provide data from 1980. The elevation from MERRA-2 is average for a 0.5 x 0.625 degree latitude/longitude region about 944.25 meters. For the entire year 2021, there are 8016 data items in the dataset for both solar irradiance and ambient temperatures. The solar irradiance data ranges from 0 to 1033.38 Wh/m^2, which is clear-sky surface shortwave downward irradiance combined with the all-sky insolation clearness index. In same way the ambient temperature ranges from 18.47 to 26.5 $^{\circ}$C, which is the MERRA-2 temperature at 2 meters. The model can be used in the research area based on location-specific data. In fact, the model can be used in any other field that has the necessary data for the defined domains and is available to be trained.

4.2. Training and Validation

The training is carried out using the Windows X-compatible MatLab 2021b program. The PC's hardware includes a Core i7-2600 CPU running at 3.40 GHz with 4 GB of RAM.

The datasets of 8016 from NASA are used to train the ELM model. The training time of 4.4375 seconds demonstrates the ELM's superiority, and its precision is sufficient, as indicated by a little modest MAE value. The test data is made up of data for 24 hours, which is the average amount of time for the entire year 2021. Training and testing with 5000 neurons take approximately 4.4375 seconds and 0.0625 seconds, respectively. The validation of the model was done using mean absolute error (MAE), with an error of about 0.6392 for the developed model.

4.3. Simulations

The simulations are performed to forecast the Manokwari region's daily solar irradiation and ambient temperatures. Then the results are presented in Figures 2 to 5.

The distribution of solar irradiation data is depicted in Figure 2 as fluctuating between 0 and a maximum of 1022.9 Wh/m^2. On the other side, hourly fluctuations occur at 14:00, when the difference between the minimum and maximum is rather large, measuring about 142.45 Wh/m^2. The graph also demonstrates that the forecasted results were located within these gaps [1].

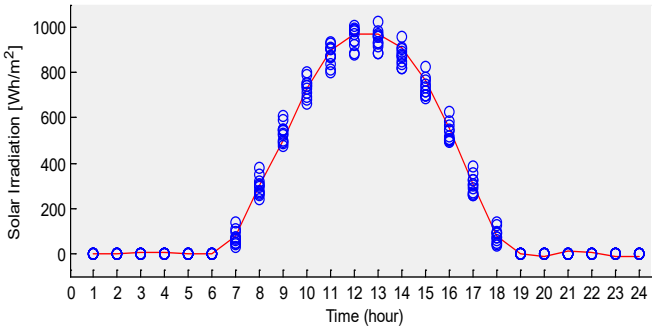


Figure 2: Solar irradiation distribution

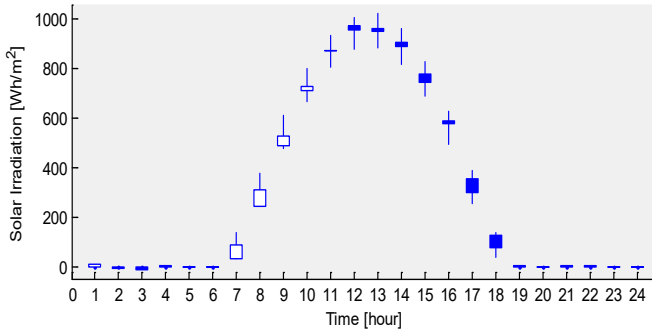


Figure 3: Forecasting of solar irradiation

As shown in Figure 3, a candlestick diagram is used to present the data between the minimum and maximum as well as factual and forecasted data. A thin line in this graph joins the minimum and maximum data points. On the other side, a thick line connects the error rate, which is the discrepancy between the actual and forecasted data. According to the simulation results in Figure 3, the highest error rate occurs at 8:00, around 55.75 Wh/m^2. The thick line is currently white, implying that the difference is negative in size as a result of the projected result being greater than the actual data [1].

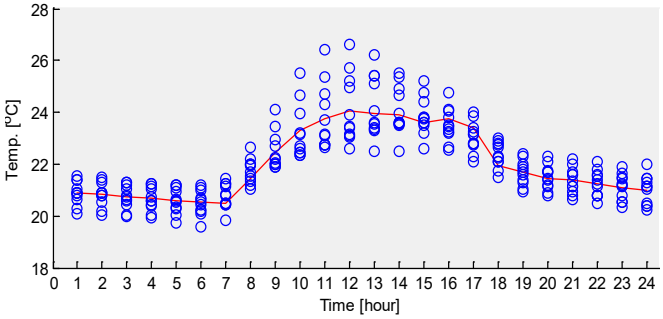


Figure 4: Ambient temperature distribution

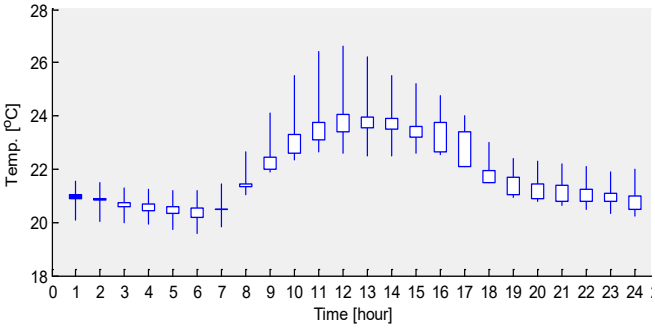


Figure 5: Forecasting of ambient temperature

The ELM model also outputs ambient temperature in accordance with solar irradiation forecasts. The forecasted environmental temperatures are given in Figures 4 and 5 [1]. As seen in Figure 4, the data distribution simultaneously displays a relatively significant variation in the Manokwari region, in contrast to the distribution of data from solar irradiation. This region is close to the equator and directly adjacent to the Pacific Ocean, which indicates a considerable change in environmental circumstances.

The candlestick diagram in Figure 5 depicts data changes between maximum and minimum data for the remainder of the year, with the highest data difference occurring at noon and being 4.02 ˚C. The figure also shows the discrepancy between the actual and anticipated data, with the largest variance of approximately 1.10 ˚C occurring around 16:00 [1].

4.4. Model Performances

Using the same ELM data for both training and testing, a straightforward feedforward neural network (FFNN) is used to assess the model's performance. The results of the performance test indicate that more than 1000 neurons cannot be operated on the same machine, while the FFNN requires 203.3281 seconds for training and 0.1250 seconds for testing. The average deviations for both approaches are 0.03 ˚C and 0.28 Wh/m^2, respectively [1]. The comparisons revealed that the ELM model has the same accuracy as the FFNN in extremely fast computing.

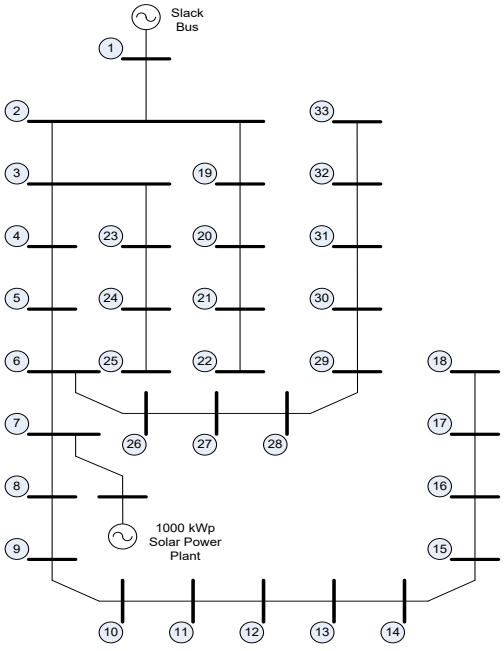


Figure 6: Single line diagram

4.5. Power Losses Minimization

Surrounding the research location, a distribution line, namely the Rajawali feeder, is operated. The feeder consists of 33 load buses connected to 32 lines, as figured in Figure 6, with data attached in Tables 1 and 2 [29].

Based on Figure 6, it can be seen that the distribution lines are connected in a radial connection. The longest line is from bus 1 to bus 18, where some tabs are taken from bus 6 to connect buses 26 and 33 and a tab is added for the solar power plant penetration. The other branches of the feeder are taken from bus 2 to connect buses 19 and 22 and from bus 3 to connect buses 23 and 25.

Table 1: Bus data

No	Bus Type*	Voltage (pu)	Angle (deg)	Load		Generation	
				P (kW)	Q (kVar)	P (kW)	Q (kVar)
1	1	1.00	0.00	0.00	0.00	0.00	0.00
2	2	1.00	0.00	100.00	60.00	0.00	0.00
3	2	1.00	0.00	90.00	40.00	0.00	0.00
4	2	1.00	0.00	120.00	80.00	0.00	0.00
5	2	1.00	0.00	60.00	30.00	0.00	0.00
6	2	1.00	0.00	60.00	20.00	0.00	0.00
7	3	1.00	0.00	200.00	100.00	1000.00	0.00
8	2	1.00	0.00	200.00	100.00	0.00	0.00
9	2	1.00	0.00	60.00	20.00	0.00	0.00
10	2	1.00	0.00	60.00	20.00	0.00	0.00
11	2	1.00	0.00	45.00	30.00	0.00	0.00
12	2	1.00	0.00	60.00	35.00	0.00	0.00
13	2	1.00	0.00	60.00	35.00	0.00	0.00
14	2	1.00	0.00	120.00	80.00	0.00	0.00
15	2	1.00	0.00	60.00	10.00	0.00	0.00
16	2	1.00	0.00	60.00	20.00	0.00	0.00
17	2	1.00	0.00	60.00	20.00	0.00	0.00
18	2	1.00	0.00	90.00	40.00	0.00	0.00
19	2	1.00	0.00	90.00	40.00	0.00	0.00
20	2	1.00	0.00	90.00	40.00	0.00	0.00
21	2	1.00	0.00	90.00	40.00	0.00	0.00
22	2	1.00	0.00	90.00	40.00	0.00	0.00
23	2	1.00	0.00	90.00	50.00	0.00	0.00
24	2	1.00	0.00	420.00	200.00	0.00	0.00
25	2	1.00	0.00	420.00	200.00	0.00	0.00
26	2	1.00	0.00	60.00	25.00	0.00	0.00
27	2	1.00	0.00	60.00	25.00	0.00	0.00
28	2	1.00	0.00	60.00	20.00	0.00	0.00
29	2	1.00	0.00	120.00	70.00	0.00	0.00
30	2	1.00	0.00	200.00	600.00	0.00	0.00
31	2	1.00	0.00	150.00	70.00	0.00	0.00
32	2	1.00	0.00	210.00	100.00	0.00	0.00
33	2	1.00	0.00	60.00	40.00	0.00	0.00

*Type of bus: 1:slack; 2:load; 3:generator

Table 2: Line data

No	Line (From - To)	Impedance (Ohm)		No	Line (From - To)	Impedance (Ohm)	
		R	jX			R	jX
1	1 - 2	0.0922	0.0470	17	14 - 15	0.5910	0.5260
2	2 - 3	0.4930	0.2511	18	15 - 16	0.7463	0.5450
3	2 - 19	0.1640	0.1565	19	16 - 17	1.2890	1.7210
4	3 - 4	0.3660	0.1864	20	17 - 18	0.7320	0.5740
5	3 - 23	0.4512	0.3083	21	19 - 20	1.5042	1.3554
6	4 - 5	0.3811	0.1941	22	20 - 21	0.4095	0.4784
7	5 - 6	0.8190	0.7070	23	21 - 22	0.7089	0.9373
8	6 - 7	0.1872	0.6188	24	23 - 24	0.8980	0.7091
9	6 - 26	0.2030	0.1034	25	24 - 25	0.8960	0.7011
10	7 - 8	0.7114	0.2351	26	26 - 27	0.2842	0.1447
11	8 - 9	1.0300	0.7400	27	27 - 28	1.0590	0.9337
12	9 - 10	1.0440	0.7400	28	28 - 29	0.8042	0.7006
13	10 - 11	0.1966	0.0650	29	29 - 30	0.5075	0.2585
14	11 - 12	0.3744	0.1238	30	30 - 31	0.9744	0.9630
15	12 - 13	1.4680	1.1550	31	31 - 32	0.3105	0.3619
16	13 - 14	0.5416	0.7129	32	32 - 33	0.3410	0.5302

A solar power plant has been prepared to be installed on the Rajawali feeder, which is planned to be integrated into the grid through the feeder. The power plant has an output power of 1000 kW, and the plant consists of 3498 PV panels with an output power of 260 Wp. The PV panel data is provided as follows.

Optimum operating voltage (Vmp)	: 30.60 V
Optimum operating current (Imp)	: 8.50 A
Open–circuit voltage (Voc)	:37.70 V
Short–circuit current (Isc)	: 9.15 A
Maximum power (Pmax)	: 260 W
Efficiency	:16 %
Operating module temperature	: -40 to 85 ˚C
Maximum system voltage	:1000 VDC
Maximum series fuse rating	:15 A
Power tolerance	: 0–3%
Solar cell type	: Monocrystalline
Number of cells	: 60
Dimensions (mm)	: 1636x992x45

The solar power plant will be used as a model to investigate the use of ELM forecasting results to investigate the penetration effect of the solar power plant in increasing the voltage profile and thus improving power losses. The results will be compared with the existing operating conditions, which probably have high power losses in operation.

The simulation is carried out using forecasting data from both NN and ELM methods. The forecasting results will then be converted into electrical energy, which will be integrated into the grid according to the specifications of the solar panels used and calculated using equation (6). The results of hourly load forecasting will also be influenced by the pattern of changes in load on each bus, which varies according to peak load times. The simulation results show an improvement in the voltage profile and power loss, as shown in Tables 3 and 4, and also illustrated in Figure 7.

Table 3: Average voltage magnitude (p.u.) and load profile (kW)

Bus	Existing	NN	ELM	Load	Bus	Existing	NN	ELM	Load
1	1.000	1.000	1.000	135.06	18	0.928	0.941	0.941	59.64
2	0.997	0.998	0.998	85.47	19	0.995	0.996	0.996	74.16
3	0.986	0.987	0.987	120.79	20	0.992	0.991	0.991	84.24
4	0.983	0.985	0.985	66.00	21	0.990	0.990	0.990	93.02
5	0.976	0.981	0.981	96.91	22	0.989	0.989	0.989	80.38
6	0.968	0.978	0.978	80.29	23	0.985	0.984	0.984	73.44
7	0.963	0.984	0.984	99.46	24	0.983	0.978	0.978	133.77
8	0.955	0.976	0.976	113.71	25	0.983	0.975	0.975	63.01
9	0.954	0.970	0.970	73.32	26	0.962	0.974	0.974	88.08
10	0.953	0.965	0.965	94.15	27	0.961	0.971	0.971	67.16
11	0.948	0.962	0.962	73.55	28	0.956	0.960	0.960	65.13
12	0.941	0.958	0.958	69.21	29	0.953	0.952	0.952	62.47
13	0.936	0.951	0.951	85.46	30	0.951	0.948	0.948	65.37
14	0.935	0.949	0.949	69.00	31	0.949	0.944	0.944	77.83
15	0.933	0.947	0.947	82.24	32	0.948	0.943	0.943	78.92
16	0.929	0.944	0.944	140.82	33	0.948	0.942	0.942	99.92
17	0.928	0.942	0.942	98.08	Avg	0.962	0.968	0.968	86.37

The outer buses on the feeder are those of 18, 22, 25, and 33, which have the lowest average voltage on the lines. After the penetration of the solar power plant, the average voltages are increasing on the buses, except for bus number 22, which is highly influenced by the slack bus voltage, as shown in Table 3. Overall, the average voltage of the feeder is increased by the solar power plant from 0.962 to 0.968 for both NN and ELM forecasting. The voltage profile improvement also occurs in many cases of power plant injection into a grid, which acts as a distributed generator, as reported in references [9–11,14,17].

Table 4: Power Loss Minimization (kW)

Time	Existing	NN	ELM	Time	Existing	NN	ELM
1	1.4050	1.4050	1.4050	14	1.6290	1.0760	0.4550
2	1.4270	1.4270	1.4270	15	1.6020	1.0760	0.4660
3	1.4110	1.4110	1.4110	16	1.5960	1.0760	0.4560
4	1.4810	1.4810	1.4810	17	1.5510	1.0760	0.4560
5	1.4840	1.4840	1.4840	18	1.5460	1.0760	0.4480
6	1.4960	1.0760	0.4270	19	1.5640	1.5640	1.5640
7	1.5020	1.0760	0.4320	20	1.5060	1.5060	1.5060
8	1.5210	1.0760	0.4320	21	1.4780	1.4780	1.4780
9	1.5060	1.0760	0.4380	22	1.4020	1.4020	1.4020
10	1.5270	1.0760	0.4350	23	1.4050	1.4050	1.4050
11	1.5960	1.0760	0.4420	24	1.4020	1.4020	1.4020
12	1.5980	1.0750	0.4580	Σ	36.2280	29.9520	21.7640
13	1.5930	1.0760	0.4540	%	100.0000	17.3236	39.9249

The data in Table 4 shows that at night, the power loss forecast for the line will be the same as the existing condition. Therefore, the power loss simulation is carried out using the forward backward sweep (FBS) method. Furthermore, when the sun starts to shine, the power loss is analyzed by the Optimal Power Flow (OPF) method. This is a fact of the system because it is a pure radial distribution system without solar power penetration. Then the power loss problem needs to be solved with the FBS method because it will cause simulation calculation errors with the OPF method that is used for a complex power system. In a different way, during the day, there is energy penetration on the system's bus 7, transforming the system into a multi-machine system connected to a radial distribution system, where power loss problems should be counted using the OPF method.

When the solar power starts to penetrate at 6 a.m., the ELM simulation shows a decrease in power loss until it reaches a maximum at around 12 p.m., and after noon, the power loss slowly increases as the solar radiation reaching the PV panel is reduced. The ELM method is smaller than the NN method because the ELM prediction gives more accurate results compared to the NN forecast, which yields a high percentage of power loss minimization. Data on the table shows that average power loss without solar power penetration is about 1.5095 kW/hour, which is reduced in average by the penetration of about 1.2480 kW/hour and 0.9068 kW/hour of NN and ELM forecasting, respectively.

On the other hand, the simulation results show that the power loss simulated through NN forecasting tends to be constant throughout the solar power plant's operation. This resulted in less solar power penetration, and its contribution to the power loss is also less, as indicated by the small percentage of power loss, which is only 17.3236 % compared to the ELM forecasting percentage that reached 39.9249 %.

The standard deviations of power losses have decreased from 6.9441 in the existing condition to 5.7440 and 4.2015 for both NN and ELM, respectively. This indicates that the performance of the system has improved with the penetration of solar power plants. However, the smallest standard deviation values indicate that power loss minimization represents a more accurate result of the ELM algorithm.

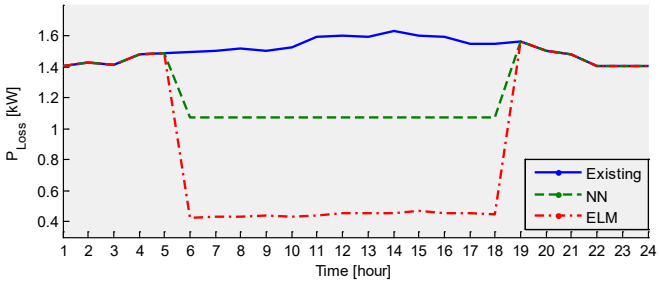


Figure 7: Power losses improvement

Figure 7 depicts the difference in power loss for the following day of the normal calculation versus the improved results due to the penetration of the 1000 kWp solar power plant on bus 7. The figure shows that there is a high power loss improvement predicted with ELM compared to the NN prediction, which tends to be stable throughout the operating time of the solar power plant. The prediction results show the accumulated power loss through NN predictions of 29.9520 kW and ELM predictions of 21.7640 kW, which is slightly lower than the normal calculation of about 36.2280 kW. This prediction result is not high for predictions in the day ahead, but it will change significantly for long-term forecasting.

Overall, the proposed method is done in the sequence of forecasting solar irradiation a day ahead in [1], injecting the solar irradiation into a power grid, calculating power flow in the grid, and investigating the power losses by comparing the results with another method. It is found that by applying the proposed method in a microgrid system, the average voltage is raised to 0.968 p.u. and further power losses in day-ahead operation can be minimized to 0.79% of the total load of 2763.84 kW. The proposed method is also valid in terms of accuracy as measured by standard deviation.

5. Conclusions

In this study, an ELM model has been developed to forecast solar data for the following day as well as data on the ambient temperature. The ELM model includes 5000 neurons in the hidden layer and was trained using annual datasets of about 8016 items. The forecasting process takes 0.0625 seconds, while training takes 4.4375 seconds of CPU time. The model has been verified using the MAE approach and has a relatively low error rate of 0.6392, making it considered accurate enough to be used to forecast solar irradiation and the surrounding air temperature at the sampling sites. The ELM model's performances have also been compared with those of a straightforward feed-forward neural network (FFNN), which has the same accuracy but requires less time to train and evaluate.

Power loss minimization for the day-ahead operation of a 1000 kWp solar panel based on the ELM data forecasting can reach 21.7640 kW, or about 39.9249 percent, compared to the existing operation without solar penetration of about 36.2280 kW. The power loss is reduced by the penetration of the solar power plant from 1.5095 kW/hour to 0.9068 kW/hour.

Acknowledgment

The first author would like to thank his colleagues from Electrical Engineering Department, Engineering Faculty of Papua University, Manokwari, who provided insight and expertise that greatly assisted this research.

References

[1] A.B. Rehiara, S. Setiawidayat, "Day Ahead Solar Irradiation Forecasting Based on Extreme Learning Machine," in 2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), 63–66, 2022, doi:10.1109/CyberneticsCom55287.2022.9865532.

[2] J.M. Bright, Introduction To Synthetic Solar Irradiance, 1-1-1–32, doi:10.1063/9780735421820_001.

[3] B. Brahma, R. Wadhvani, "Solar Irradiance Forecasting Based on Deep Learning Methodologies and Multi-Site Data," Symmetry, **12**(11), 2020, doi:10.3390/sym12111830.

[4] M. Abdillah, W.A. Pramudito, T.A. Nugroho, D.N. Fitria, "Solar irradiance forecasting using kernel extreme learning machine: case study at Lamongan and Muara Karang regions, Indonesia," 17, 2022.

[5] V.V.V.S.N. Murty, A. Kumar, "Optimal Energy Management and Techno-economic Analysis in Microgrid with Hybrid Renewable Energy Sources," Journal of Modern Power Systems and Clean Energy, **8**(5), 929–940, 2020, doi:10.35833/MPCE.2020.000273.

[6] D. Sampath Kumar, O. Gandhi, C.D. Rodríguez-Gallegos, D. Srinivasan, "Review of power system impacts at high PV penetration Part II: Potential solutions and the way forward," Special Issue on Grid Integration, 210, 202–221, 2020, doi:10.1016/j.solener.2020.08.047.

[7] S. Monshizadeh, "Comparison of Intelligent Algorithms with FACTS Devices for Minimization of Total Power Losses," Advances in Intelligent Systems and Computing, 926(Query date: 2022-12-02 17:21:04), 120–131, 2020, doi:10.1007/978-3-030-15032-7_10.

[8] S.P. Dash, "Optimal location and parametric settings of FACTS devices based on JAYA blended moth flame optimization for transmission loss minimization in power systems," Microsystem Technologies, **26**(5), 1543–1552, 2020, doi:10.1007/s00542-019-04692-w.

[9] A. Alam, "Optimal placement of DG in distribution system for power loss minimization and voltage profile improvement," 2018 International Conference on Computing, Power and Communication Technologies, GUCON 2018, (Query date: 2022-12-02 17:21:04), 837–842, 2019, doi:10.1109/GUCON.2018.8674930.

[10] S. Essallah, "Optimal Multi-Type DG Integration and Distribution System Reconfiguration for Active Power Loss Minimization using CPSO Algorithm," 2019 International Conference on Control, Automation and Diagnosis, ICCAD 2019 - Proceedings, (Query date: 2022-12-02 17:21:04), 2019, doi:10.1109/ICCAD46983.2019.9037947.

[11] G.A.S. Gandhi, "Optimal allocation of DG for minimization of power loss and total investment cost using an analytical approach," 2020 21st National Power Systems Conference, NPSC 2020, (Query date: 2022-12-02 17:21:04), 2020, doi:10.1109/NPSC49263.2020.9331891.

[12] S. Ansari, J. Zhang, R.E. Singh, "A review of stabilization methods for DCMG with CPL, the role of bandwidth limits and droop control," Protection and Control of Modern Power Systems, **7**(1), 2, 2022, doi:10.1186/s41601-021-00222-x.

[13] N. Pragallapati, S.J. Ranade, O. Lavrova, "Cyber Physical Implementation of Improved Distributed Secondary Control of DC Microgrid," 2021 1st International Conference on Power Electronics and Energy (ICPEE), doi:10.1109/icpee50452.2021.9358705.

[14] M. Kang, "Optimal placement and sizing of DG and shunt capacitor for power loss minimization in an islanded distribution system," Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, 245(Query date: 2022-12-02 17:21:04), 43–52, 2018, doi:10.1007/978-3-319-94965-9_5.

[15] M.B. Essa, "Distribution power loss minimization via optimal sizing and placement of shunt capacitor and distributed generator with network

reconfiguration," Telkomnika (Telecommunication Computing Electronics and Control), **19**(3), 1039–1049, 2021, doi:10.12928/TELKOMNIKA.v19i3.15223.
[16] D. Bhowmik, "Optimal placement of capacitor banks for power loss minimization in transmission systems using fuzzy logic," Journal of Engineering Science and Technology, **13**(10), 3190–3203, 2018.
[17] S. Nawaz, "A new technique to solve DG allocation problem for distribution power loss minimization," ICIC Express Letters, Part B: Applications, **9**(7), 701–706, 2018, doi:10.24507/icicelb.09.07.701.
[18] N.T. Rao, "Comparative study of Pareto optimal multi objective cuckoo search algorithm and multi objective particle swarm optimization for power loss minimization incorporating UPFC," Journal of Ambient Intelligence and Humanized Computing, **12**(1), 1069–1080, 2021, doi:10.1007/s12652-020-02142-4.
[19] M. Sankaramoorthy, "A hybrid MACO and BFOA algorithm for power loss minimization and total cost reduction in distribution systems," Turkish Journal of Electrical Engineering and Computer Sciences, **25**(1), 337–351, 2017, doi:10.3906/elk-1410-191.
[20] N. Vazquez, "A Fully Decentralized Adaptive Droop Optimization Strategy for Power Loss Minimization in Microgrids with PV-BESS," IEEE Transactions on Energy Conversion, **34**(1), 385–395, 2019, doi:10.1109/TEC.2018.2878246.
[21] R. Adelhard Beni, C. He, S. Yutaka, Y. Naoto, Z. Yoshifumi, "An Adaptive Internal Model for Load Frequency Control Using Extreme Learning Machine," TELKOMNIKA, **16**(6), 2879–2887, 2018, doi:http://dx.doi.org/10.12928/telkomnika.v16i6.11553.
[22] Y. Lu, W. Yu, J. Wang, D. Jiang, R. Li, "Design of PID Controller Based on ELM and Its Implementation for Buck Converters," International Journal of Control, Automation and Systems, **19**(7), 2479–2490, 2021, doi:10.1007/s12555-019-0989-1.
[23] P. Winangun, I.M. Widyantara, R. Hartati, "Extreme Learning Machine Based Diagnostic Approach with Linear Kernel for Classifying Lung Disorders," Majalah Ilmiah Teknologi Elektro, **19**(1), 83–88, 2020, doi:10.24843/MITE.2020.v19i01.P12.
[24] C. Jia, H. Zhang, "ELM Neural Network-based Fault Diagnosis Method for Mechanical Equipment," in 2019 Chinese Automation Congress (CAC), 5257–5261, 2019, doi:10.1109/CAC48633.2019.8996333.
[25] M.N. dan I.C. dan I. Indriati, "Stock Price Earning Ratio Prediction Using the Kernel Extreme Learning Machine Algorithm (Case Study: PT TELKOM)," Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, **4**(10), 3455–3462, 2020.
[26] G. Huang, S. Song, J.N.D. Gupta, C. Wu, "Semi-Supervised and Unsupervised Extreme Learning Machines," IEEE Transactions on Cybernetics, 44(12), 2405–2417, 2014, doi:10.1109/TCYB.2014.2307349.
[27] F. Odoi-Yorke, A. Woenagnon, "Techno-economic assessment of solar PV/fuel cell hybrid power system for telecom base stations in Ghana," Cogent Engineering, **8**(1), 1911285, 2021, doi:10.1080/23311916.2021.1911285.
[28] S.F. Mekhamer, S.A. Soliman, M.A. Mostafa, M.E. El-Hawary, "Load flow solution of radial distribution feeders: a new approach," in 2001 IEEE Porto Power Tech Proceedings (Cat. No.01EX502), **3**, 5, 2001, doi:10.1109/PTC.2001.964918.
[29] B.R. Wihyawari, "Radial Power Flow Analysis of Rajawali Feeder in Manokwari Power Grid," in ICTVT 2021, State University of Yogyakarta, 2021.
[30] J.A.M. Rupa, S. Ganesh, "Power Flow Analysis for Radial Distribution System Using Backward/Forward Sweep Method," International Journal of Electrical and Computer Engineering, **8**(10), 1628–1632, 2014.
[31] A.B. Rehiara, "Optimal Power Flow of the Manokwari Power Grid Regarding Penetration of 20 MW Combined Cycle Power Plant," in 2019 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA), 53–57, 2019, doi:10.1109/ICAMIMIA47173.2019.9223393.
[32] L.T. Al-Bahran, A.Q. Abdulrasool, "Multi objective functions of constraint optimal power flow based on modified ant colony system optimization technique," IOP Conference Series: Materials Science and Engineering, **1105**(1), 012015, 2021, doi:10.1088/1757-899X/1105/1/012015.
[33] H. Bouchekara, "Solution of the optimal power flow problem considering security constraints using an improved chaotic electromagnetic field optimization algorithm," Neural Computing and Applications, **32**(7), 2683–2703, 2020, doi:10.1007/s00521-019-04298-3.
[34] NASA, NASA/POWER CERES/MERRA2 Native Resolution Hourly Data, 2021.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 1, 78-83 (2023)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

On the Polytopic Modelling & Robust H_∞ Control of Nonlinear Systems Subject to Cyber-attack: Application to Attitude Stabilization of Quadrotor

Bezzaoucha-RebaÏ Souad *

EIGSI- La Rochelle, 17041, France

ARTICLE INFO

Article history:
Received: 05 November, 2022
Accepted: 08 January, 2023
Online: 24 January, 2023

Keywords:
Quadrotor
Stabilization
Polytopic representation
Stealthy attacks

ABSTRACT

In the present contribution, a robust output H_∞ control ensuring the stability, reliability and security for nonlinear systems when actuator attacks (data deception attacks) occur. A new design method based on the polytopic rewriting of the attacked system as an uncertain one subject to external disturbances will be detailed. Robust polytopic state feedback observer stabilizing controller based on the PDC (Parallel Distributed Compensation) polytopic framework with disturbance attenuation for the obtained uncertain system will also be considered. The obtained methodology is used to ensure the stability and security of a quadrotor/UAV subject to stealthy actuator attacks. State and attacks estimations signals are given in order to highlight the efficiency of the developed approach.

1 Introduction

Based on previous contribution [1], this paper is an extension of the original work which aims to ensure a robust attitude stabilization of a quadrotor subject to stealthy actuator attacks. The modelling and control aspect were considered in this first contribution, where in the following the observer design for both state and stealthy attacks is added. Robust polytopic state feedback stabilizing controller based on PDC (Parallel Distributed Compensation) polytopic framework observer with disturbance attenuation (guaranted by the H_∞ norm) for the obtained uncertain system is also considered.

Design and implementing feedback control strategies that are robust against cyber-attacks is of critical importance nowadays. Assuming that the behavior of the system is driven via actuator commands, the actuator data deception attack corresponds to a manipulation of an attacker on the communication channels between the plant and the controller. The actuator commands are then corrupted and it becomes necessary to integrate this data in the control system design and make it as robust as possible to these stealthy attacks.

The objective is then the design of a resilient control for a system where an attacker corrupts control packets; and of course, to detect and reduce/attenuate the effect of these corrupted signals on the well-behaviour of the considered system.

In the following contribution, the novelty comparing the work originally presented in [1] is about the estimation and robust control part where both state and stealthy attacks are now estimated with an H_∞ disturbance attenuation property.

One solution in order to represent and implement heuristic knowledge to control nonlinear systems when remaining the study relatively simple consists in the use of the polytopic Takagi-Sugeno (T-S) structure. Indeed, this representation was initially proposed by [2] and [3], and proven its efficiency in various applications in the past decades [4].

Based on the polytopic T-S approach, a number of most important issues in control systems have been addressed in the past few years. These includes stability analysis [5], state and output feedback control [6], [7], performances and robustness [8], [4], as well as recently cyber-security [9]–[12].

Solving the considered problem (H_∞ control of stabilization), the nonlinear behaviour of the quadrotor, including stealthy attacks, is represented in terms of an uncertain polytopic system subject to bounded external disturbances. The stabilization and robust H_∞ control, based on state estimation feedback is then deduced based on classical Lyapunov theory leading to a set of matrices inequalities (constraints) to solve. These constraints, solvable through convex optimization techniques allows to obtain polytopic controllers and observers that guarantee both stability and robustness of the closed-loop system.

Indeed, this paper focuses on the problem of observer-based

*Corresponding Author: BEZZAOUCHA-REBAÏ, Souad. EIGSI-La Rochelle; 26 Rue François de Vaux de Foletier, 17000, France. Email: souad.bezzaoucha@eigsi.fr

H_∞ control for polytopic T-S systems under actuator data deception attacks. Sufficient conditions for the simultaneous controller and observer design with a desired H_∞ disturbance attenuation level are derived in terms of linear matrix inequalities wich can be easily solved by using available software package (Matlab for example).

The present paper objective is to contribute to the cyber-security and resilience design of Unmanned Aerial Vehicles (UAVs). It is known that the wireless control used to monitor drones makes them defenceless to a large variety of cyber-attacks, which may have severe consequences on the system behaviour/security/integrity/performances.

In this contribution, stealthy attacks disturbing and destabilizing the control and navigation system of the UAV are studied. We aim to propose a robust control ensuring the safety and security of the UAv despite these assaults.

This contribution strategy was previously developed for cyber-attacks estimation [1], [10] and is now adapted to the quadrotor robust control. The estimation of the system states and stealthy signals will be given.

Considering the estimable premise variables, the attacked system will be presented as an uncertain T-S model. Based the resulting system, a PDC observer based control will be designed.

The paper organized as follows: a brief state of the art is presented in section 1; the system modelling with the actuator data deception attack and the uncertain system representation are detailed respectively in section 2 and 3. Section 4 is about the robust output H_∞ observe-based T-S controller. Section 5 is about the approach illustration through an application to quadrotor attitude stabilization with simulation results. The final section, 6 is about conclusion and perspectives.

The applied methodology solving the considered problem is the following:

1. The modelling aspect: such that the nonlinear behaviour and threats attacks are both represented in a polytopic T-S form based on the sector nonlinearity transformation (SNT); it is important to note that in this representation, there is no approximation or any loose of information. The main advantage of this method consists into an exact rewriting of the original nonlinear equations.

2. In order to be able to implement the state feedback observer based PDC H_∞ control law, the main constraint in the obtained model (5) is about the immeasurable state and time-varying parameters present in the weighting functions h_i & μ_j, and the control law. In order to overcome this difficulty, it is imperative to standardize the system equations in order to have only measurable and/or estimable premise variables. For that, an uncertain representation of the system equations (5) is proposed in section 3 in order to obtain a more convenient model for the study; i.e. (19).

3. Based on the chosen structure for the observer and control law, a robust H_∞ T-S control of the nonlinear system is considered in section 4. The objective, in addition to the system stabilization, is to ensure an attenuation of the external perturbation, guaranteed thanks to the H_∞ norm.

4. The final step of our study would be the application of the developed approach to our case study; i.e. the attitude stabilization of the quadrotor.

.

2 System Modelling: a Polytopic representation

In the following section, based on the nonlinear state space model of a system, a polytopic representation will be deduced applying the sector nonlinearity approach (SNT).
Assuming that our system is subject to actuator data deception attacks (modeled as unknown, but bounded, multiplicative time-varying parameters); our nonlinear model under these attacks may be represented by the following state space system equations:

$$\begin{cases} \dot{x}(t) &= \sum_{i=1}^{r} h_i(t)(A_i x(t) + B_i(t)u(t)) \\ y(t) &= Cx(t) \end{cases} \tag{1}$$

s.t. matrices $B_i(t)$ are defined as:

$$B_i(t) := B_i + \Gamma^u a^u(t) \tag{2}$$

In this model, B_i is called the nominal input matrix (i.e. when none attack occurring); Γ^u is known as the binary incidence matrix, which indicates the data channels that can be accessed by the attacker; finally $a^u(t)$ represents the actuator data corruption signal.
The stealthy signals $a^u(t)$ are unknown (in terms of value), but bounded (limits assumed to be known) $a^u(t) \in [{a_2}^u, {a_1}^u]$. Applying the SNT transformation, the following representation is proposed:

$$a^u(t) = \mu_1(a^u(t)){a_1}^u + \mu_2(a^u(t)){a_2}^u \tag{3}$$

with

$$\begin{cases} \mu_1(a^u(t)) &= \dfrac{a^u(t) - {a_2}^u}{{a_1}^u - a_2^u} \\ \mu_2(a^u(t)) &= \dfrac{{a_1}^u - a^u(t)}{{a_1}^u - a_2^u} \end{cases} \tag{4}$$

$$\mu_1(a^u(t)) + \mu_2(a^u(t)) = 1, \ \forall t$$

The equations (1) of the attacked system is then rewritten as:

$$\begin{cases} \dot{x}(t) &= \sum_{i=1}^{r}\sum_{j=1}^{2} h_i(t)\mu_j(t)(A_i x(t) + \mathcal{B}_{ij}u(t)) \\ y(t) &= Cx(t) \end{cases} \tag{5}$$

s.t. $\mathcal{B}_i^j(t)$ are defined as:

$$\mathcal{B}_{i1} = B_i + {a_1}^u\Gamma^u, \quad \mathcal{B}_{i2} = B_i + {a_2}^u\Gamma^u; \quad i = 1, 2 \tag{6}$$

3 Uncertain System Representation

Based on contributions [11] and [12], where the data deception representation of cyber-attacks via a time-varying model and thanks to a polytopic form of an uncertain system representation, a state

and actuator data deception attacks observer is proposed and given by the following equations:

$$\begin{cases} \dot{\hat{x}}(t) = \sum_{i=1}^{r}\sum_{j=1}^{2} h_i(\hat{x}(t))\mu_j(\hat{a^u}(t)) \\ \qquad \left(A_i\hat{x}(t) + \mathcal{B}_{ij}u(t) + L_{ij}(y(t) - \hat{y}(t))\right) \\ \dot{\hat{a}}^u(t) = \sum_{i=1}^{r}\sum_{j=1}^{2} h_i(\hat{x}(t))\mu_j(\hat{a^u}(t)) \\ \qquad (K_{ij}(y(t) - \hat{y}(t)) - \alpha_{ij}^u\hat{a}^u(t)) \\ \hat{y}(t) = C\hat{x}(t) \end{cases} \tag{7}$$

s.t. $L_{ij} \in \mathbb{R}^{n_x \times m}$, $K_{ij}^u \in \mathbb{R}^{n \times m}$ and $\alpha_{ij}^u \in \mathbb{R}^{n \times n}$ are solution of a $LMI\text{-}H\infty_2$ attenuation conditions ensuring both estimation errors for the states and malicious input parameters to converge to zero.
Let us now define the estimation errors $e_x(t)$ and $e_{a^u}(t)$ (for the state and cyber-attacks) as:

$$\begin{array}{rcl} e_x(t) & = & x(t) - \hat{x}(t) \\ e_{a^u}(t) & = & a^u(t) - \hat{a^u}(t) \end{array} \tag{8}$$

The system equations (5) can be rewritten as follows [13], [11]:

$$\begin{cases} \dot{x}(t) = \sum_{i=1}^{r}\sum_{j=1}^{2}[h_i(\hat{x}(t))\mu_j(\hat{a^u}(t))(A_ix(t) + \mathcal{B}_{ij}u(t)) + \\ \qquad \delta_{ij}(t)(A_ix(t) + \mathcal{B}_{ij}u(t))] \\ y(t) = Cx(t) \end{cases} \tag{9}$$

with $\delta_{ij}(t)$ are defined by the following equations:

$$\delta_{ij}(t) = h_i(x(t))\mu_j(a^u(t)) - \mu_i(\hat{x}(t))\mu_j(\hat{a^u}(t)) \tag{10}$$

satisfying:

$$-1 \leq \delta_{ij}(t) \leq 1 \tag{11}$$

Let us introduce now:

$$\Delta A(t) = \sum_{i=1}^{r}\sum_{j=1}^{2} \delta ij(t)A_i = \mathcal{A}\Sigma(t)E_A \tag{12}$$

$$\Delta B(t) = \sum_{i=1}^{r}\sum_{j=1}^{2} \delta_{ij}(t)\mathcal{B}_{ij} = \mathcal{B}\Sigma(t)E_B \tag{13}$$

with

$$\mathcal{A} = \left[\underbrace{A_1 \quad A_1}_{2\,times} \quad \dots \quad \underbrace{A_r \quad \dots \quad A_r}_{2\,times} \right] \tag{14}$$

$$\mathcal{B} = \left[\mathcal{B}_1^1 \quad \dots \quad \mathcal{B}_r^2 \right] \tag{15}$$

$$\Sigma(t) = \text{diag}(\delta_{11}(t), \dots, \delta_{r2}(t)), \tag{16}$$

$$E_A = \left[I_{n_x} \quad \dots \quad I_{n_x} \right]^T, \; E_B = \left[I_{n_u} \quad \dots \quad I_{n_u} \right]^T \tag{17}$$

Thanks to (11) and definitions (16), we have:

$$\Sigma^T(t)\Sigma(t) \leq I \tag{18}$$

Using the above definitions (12)-(17), system (11) is then written as an uncertain system given by:

$$\begin{cases} \dot{x}(t) = \sum_{i=1}^{r}\sum_{j=1}^{2} h_i(\hat{x}(t))\mu_j(\hat{a^u}(t)) \\ \qquad ((A_i + \Delta A(t))x(t) + (\mathcal{B}_{ij} + \Delta B(t))u(t)) \\ y(t) = Cx(t) \end{cases} \tag{19}$$

4 Robust Output H_∞ T-S Control

The objective in the following section is to determine polytopic T-S controller and observer gains ensuring that:

- The system given by (19) is asymptotically stable in the presence of data deception attacks.
- The attenuation of the external perturbations like (i.e. actuator attacks) is guaranteed by the H_∞ norm. i.e. find for a given scalar $\gamma > 0$, an observer (7) and a PDC controller (20) such that the attenuation condition (26) is satisfied. The resulting conditions to be solved will be given in Lemma 2.

Considering the nonlinear system subject to data deception attacks given by the system equations (1), and the polytopic T-S observer to estimate the unmeasurable state variables and unknown time-varying parameter (actuator attack signal $a^u(t)$) given by system equations (7) with the following PDC (Parallel Distributed Compensation) controller defined by:

$$u(t) = -\sum_{k=1}^{r} h_k(\hat{x}(t))\Omega_k\hat{x}(t) \tag{20}$$

By combining the uncertain system equations (19), the observer equations (7), the polytopic PDC controller and the estimation errors definitions (8), the following uncertain system with bounded external disturbances is obtained:

$$\dot{x}_a(t) = \sum_{i=1}^{r}\sum_{j=1}^{2}\sum_{k=1}^{r} h_i(\hat{x})\mu_j(\hat{a^u})h_k(\hat{x})(\Phi_{ijk}x_a(t) + \Psi_{ij}\omega(t)) \tag{21}$$

s.t. $x_a(t) = \left(\begin{array}{ccc} x(t) & e_x(t) & e_{a^u}(t) \end{array} \right)^T$ represents the augmented (extended) state vector; $\omega(t) = \left(\begin{array}{cc} a^u(t) & \dot{a^u}(t) \end{array} \right)^T$ represents the exogenous input (signal attack $a^u(t)$ and its derivative), supposed unknown but bounded.
Matrices Φ_{ijk} and Ψ_{ij} are defined as follows:

$$\Phi_{ijk} = \begin{pmatrix} \Phi_{ijk}^1 & (\mathcal{B}_{ij} + \Delta B(t))\Omega_k & 0 \\ \Delta A(t) - \Delta B(t)\Omega_k & A_i - L_{ij}C & 0 \\ 0 & -K_{ij}C & -\alpha_{ij}^u \end{pmatrix} \tag{22}$$

with $\Phi_{ijk}^1 = A_i - \mathcal{B}_{ij}\Omega_k + \Delta A(t) - \Delta B(t)\Omega_k$ and

$$\Psi_{ij} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ \alpha_{ij}^u & I \end{pmatrix} \tag{23}$$

From (21) and the system output equation (1), the resulting closed-loop system becomes:

$$\begin{pmatrix} \dot{x}_a(t) \\ y(t) \end{pmatrix} = \sum_{i=1}^{r}\sum_{j=1}^{2}\sum_{k=1}^{r} h_i(\hat{x})\mu_j(\hat{a^u})h_k(\hat{x}) \begin{pmatrix} \Phi_{ijk} & \Psi_{ij} \\ \overline{C} & 0 \end{pmatrix} \begin{pmatrix} x_a(t) \\ \omega(t) \end{pmatrix} \tag{24}$$

s.t.

$$y(t) = Cx(t) = \begin{pmatrix} C & 0 & 0 \end{pmatrix} x_a(t) = \overline{C}x_a(t) \tag{25}$$

Before presenting the stabilization and control conditions, the following definition and lemma are remembered:

Definition 1 *Given a positive scalar γ, the system equation (24) is said to be stable with H_∞ attenuation level γ if it is exponentially stable with:*

$$\int_\infty^0 \{(y^T(t))_\infty (y(t))_\infty - \gamma^2 \omega^T(t)\omega(t)\}dt < 0 \tag{26}$$

where γ is the desired level of disturbance attenuation.

Lemma 1 *Based on Lyapunov theory, the continuous-time system (24) is stable with an H_∞ disturbance attenuation γ if there exists a positive symmetric matrix $P = P^T > 0$ s.t.*

$$\begin{bmatrix} \Phi_{ijk}^T P + P\Phi_{ijk} & P\Psi_{ij} & \overline{C}^T \\ * & -\gamma^2 I & 0 \\ * & * & -I \end{bmatrix}_{i,k=1,\dots,r,j=1,2} < 0 \tag{27}$$

In order to relax the stability conditions given in Lemma 1, the following formulation is proposed [14], [4]:

Lemma 2 *[15] For a given positive scalar γ, if there exist matrices P, Z_{ijk}, where $P = P^T > 0$ and Z_{iji} are symmetrical, $Z_{kji} = Z_{ijk}^T$, $i \neq k$, $i,k = 1,\dots,r$, $j = 1,2$ satisfying the following matrix inequalities, then for the uncertain polytopic T-S system (19), the controller (20) makes the H_∞ norm of fuzzy system (24) less that γ*

$$\begin{bmatrix} \Phi_{iji}^T P + P\Phi_{iji} & P\Psi_{ij} \\ * & -\gamma^2 I \end{bmatrix}_{i=1,\dots,r,j=1,2} < Z_{iji} \tag{28}$$

$$\begin{bmatrix} (*)^T P + P(\Phi_{ijk} + \Phi_{kji}) & 2P\Psi_{ij} \\ * & 2\Psi_{ij}^T P \end{bmatrix}_{i\neq k,j=1,2} < Z_{ijk} + Z_{kji} \tag{29}$$

$$\begin{bmatrix} Z_{1j1} & \dots & Z_{1jr} & \overline{C}^T \\ \vdots & \ddots & \vdots & \vdots \\ Z_{rj1} & \dots & Z_{rjr} & \overline{C}^T \\ \overline{C} & \dots & \overline{C} & -I \end{bmatrix}_{j=1,2} < 0 \tag{30}$$

By replacing Φ_{ijk} and Ψ_{ij} by their expressions, with some change of variables and classical linearization procedure (Schur's complement and bounded real lemma), the obtained constraints can be easily solved using convex optimization tools and/or the use of a dedicated resolution tool for bilinear constraints like the PenBMI Matlab toolbox (see [16], [17] and [18] for some examples). The proposed solution presents the advantage of a simultaneous design of both the controller and the observer gains using a single-step procedure rather that a classical two-steps procedure of resolution like the one presented in [19].

5 Numerical Example

In the following, let us consider the study case of a dynamic modeling and control for quadrotor.
The objective of this work is to ensure the quadrotor safe behaviour and stabilization via an observer based control design. Indeed, stealthy actuator attacks aiming to disturb and destabilize the control and navigation system of the UAV are here considered. These attacks are modeled as unknown but bounded time-varying signals affecting the system matrix $B(t)$.
The first step to this aim will be the quadrotor Polytopic modelling; then, considering the actuator stealthy attacks, the resulting system (attacked one) will be written as an uncertain one, as detailed in previous sections . The proposed control and observer design approach will be then applied to illustrate the its efficiency thanks to simulation results.

5.1 Polytopic Model of a UAV

In this section, we address the polytopic T-S modelling of a UAV. The considered representation is used in order to rewrite the nonlinear behaviour of the quadrotor into a polytopic-Multiple Model way, without any linearisarion, loss of information or approximation. The nonlinear dynamic of the quadrotor is given by the following model:

$$\begin{cases} \ddot{\phi}(t) &= \frac{1}{I_x}\left[(I_y - I_z)\dot{\theta}\dot{\psi} - K_{fax}\dot{\phi}^2 - J_r\dot{\theta}\overline{\Omega} + l\,U_2\right] \\ \ddot{\theta}(t) &= \frac{1}{I_y}\left[(I_z - I_x)\dot{\psi}\dot{\phi} - K_{fay}\dot{\theta}^2 - J_r\dot{\phi}\overline{\Omega} + lU_3\right] \\ \ddot{\psi}(t) &= \frac{1}{I_y}\left[(I_x - I_y)\dot{\theta}\dot{\phi} - K_{fay}\dot{\psi}^2 + lU_4\right] \end{cases} \tag{31}$$

s.t. $\overline{\Omega}$ is given by $\overline{\Omega} = \omega_1 - \omega_2 + \omega_3 - \omega_4$. The motors control inputs, denoted $U_i, i = 1,2,3,4$, are given as a function of the rotors angular velocities as follows:

$$\begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{pmatrix} = \begin{pmatrix} K_t & K_t & K_t & K_t \\ -K_t & 0 & K_t & 0 \\ 0 & -K_t & 0 & K_t \\ K_d & K_d & K_d & K_d \end{pmatrix} \begin{pmatrix} \omega_1^2 \\ \omega_2^2 \\ \omega_3^2 \\ \omega_4^2 \end{pmatrix} \tag{32}$$

The angles (given in [rad]), ϕ, θ and ψ represent the Roll, Pitch, and Yaw angles respectively.
We denote the moment of inertia among axes x, y and z as I_x, I_y and I_z respectively.
J_r, K_t and K_d are the rotor inertia, propeller thrust and drag coefficients and K_{fax}, K_{fay}, K_{faz} the frictions' aerodynamic coefficients. Interested readers can see [1] and [20] for more calculation details.
From the SNT tranformation, the nonlinear system model (31) can be in a straightforward way written as a quasi-LPV model given by:

$$\begin{cases} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= Cx(t) \end{cases} \tag{33}$$

with suitable state, output and input vectors:

$$x(t) = \begin{pmatrix} \phi & \theta & \psi & \dot{\phi} & \dot{\theta} & \dot{\psi} \end{pmatrix}^T$$

$$y(t) = \begin{pmatrix} \phi & \theta & \psi & \dot{\phi} & \dot{\theta} & \dot{\psi} \end{pmatrix}^T$$

$$u(t) = \begin{pmatrix} \omega_1 & \omega_2 & \omega_3 & \omega_4 & \omega_1^2 & \omega_2^2 & \omega_3^2 & \omega_4^2 \end{pmatrix}^T$$

the state matrices are given by:

$$A(t) = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\frac{K_{fax}}{I_x}\dot{\phi} & 0 & \frac{I_y - I_z}{I_x}\dot{\theta} \\ 0 & 0 & 0 & 0 & -\frac{K_{fay}}{I_y}\dot{\theta} & \frac{I_z - I_x}{I_y}\dot{\phi} \\ 0 & 0 & 0 & \frac{I_x - I_y}{I_z}\dot{\theta} & 0 & -\frac{K_{faz}}{I_z}\dot{\psi} \end{pmatrix}$$

$$C = I_6$$

and

$$B(t) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{J_r}{I_y}\dot{\theta} & -\frac{J_r}{I_x}\dot{\theta} & \frac{J_r}{I_y}\dot{\theta} & -\frac{J_r}{I_x}\dot{\theta} & -l\frac{K_t}{I_x} & 0 & l\frac{K_t}{I_x} & 0 \\ -\frac{J_r}{I_y}\dot{\phi} & \frac{J_r}{I_y}\dot{\phi} & -\frac{J_r}{I_y}\dot{\phi} & \frac{J_r}{I_y}\dot{\phi} & 0 & -l\frac{K_t}{I_y} & 0 & l\frac{K_t}{I_y} \\ 0 & 0 & 0 & 0 & l\frac{K_d}{I_z} & -l\frac{K_d}{I_z} & l\frac{K_d}{I_z} & -l\frac{K_d}{I_z} \end{pmatrix}$$

Presuming that the variation of angular velocities occurs between known minimum and maximum values, and applying the SNT approach [2], [5], [7], when choosing the following premise variables:

$$z_1(t) = \dot{\phi}, \quad z_2(t) = \dot{\theta}, \quad z_3(t) = \dot{\psi}$$

the resulting polytopic model is then deduced:

$$\begin{cases} \dot{x}(t) &= \sum_{i=1}^{r} h_i(t)(A_i x(t) + B_i u(t)) \\ y(t) &= Cx(t) \end{cases} \tag{34}$$

5.2 *Uncertain System Modelling under Attacks*

In the previous subsection, and for the nominal case (no attacks), the polytopic model of the quadrotor-UAV (34) was deduced from its nonlinear dynamics. In the following, the data deception signal will be included and, the global system under attacks will be represented as an uncertain system.
The quasi-LPV system (34) under actuator attacks may be represented as the following:

$$\begin{cases} \dot{x}(t) &= \sum_{i=1}^{r} h_i(t)(A_i x(t) + B_i(t) u(t)) \\ y(t) &= Cx(t) \end{cases} \tag{35}$$

where $B_i(t)$ is given by:

$$B_i(t) := B_i + \Gamma^u a^u(t) \tag{36}$$

Based on the results presented in section 3, the system and attacks observer is given by the system equations (7), and the nonlinear system subject to actuator attacks is modeled thanks to system (19). The objective now is to apply the proposed approach in order to design the robust control law (20) and the observer gains.

5.3 *Simulation results*

The designed observers and controller are implemented and tested through a numerical simulation of a quadrotor robust attitude stabilization despite stealthy actuator data deception attacks.
The design goals and the controller structure (20) based on the state feedback control law is applied to the the nonlinear system equations (33) and (35) subject to the actuator stealthy attacks (36). The control gains are obtained by applying the developed polytopic approach given in Lemma 2 and solving the LMI constraints (28), (29) and (30).

The resulting figures illustrate the stability, robustness and convergence of the system states regarding the attacks.
The state, their estimates and estimations error are illustrated in the following figures:

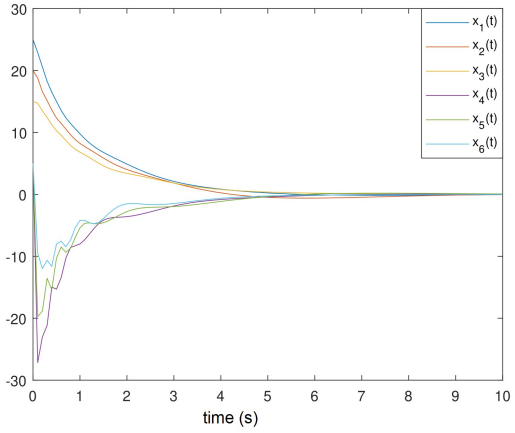


Figure 1: System states estimation errors

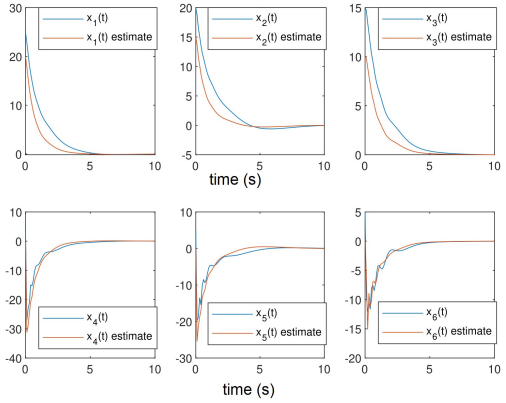


Figure 2: System states estimation errors

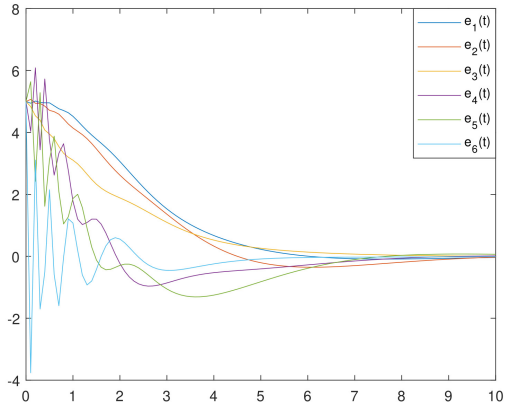


Figure 3: System states estimation errors

The stealthy attack signal $a^u(t)$ and its estimate is represented in figure 4.

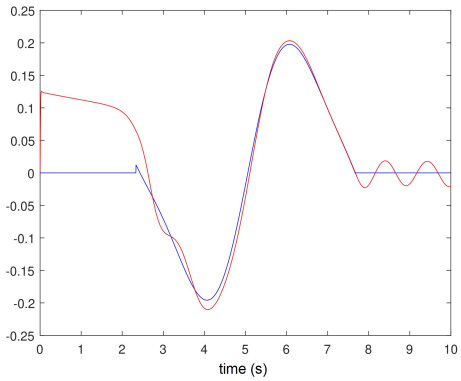


Figure 4: Stealthy attack signal

From the obtained results, one can conclude to the effectiveness of the proposed approach.
Indeed, the system states (angles and velocities) are converging asymptotically despite the stealthy attacks (unknown behaviour), where the estimation errors for both states and data deception attacks tend to zero for the state and with an attenuation level γ for the stealthy signal.

6 Conclusion

This paper contribution aimed to propose a polytopic Takagi-Sugeno strategy for the modelling and H_∞ robust control of a quadrotor subject to stealthy actuator attacks. The attacked UAV system under attacks was modeled as an uncertain polytopic T-S fuzzy one; which allowed us to generalize existing results for the state feedback observer based control design.
The nonlinear system was represented under an uncertain shape (with observable premise variables) allowing the implementation of the observer and control design. The attenuation of the external like perturbation (attack) was guaranteed thanks to the H_∞ norm. Numerical simulations were given in order to illustrate the effectiveness of the proposed approach. As an extension of this work, a real application example is also under investigation.

Conflict of Interest The author declares no conflict of interest.

References

[1] S. Bezzaoucha Rebai, "Robust Attitude Stabilization of Quadrotor Subject to Stealthy Actuator Attacks", in the 2022 International Conference on Control, Robotics and Informatics (ICCRI), Danang, Vietnam, April 2-4, 2022, doi: 10.1109/ICCRI55461.2022.00018.

[2] T. Takagi, M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control", IEEE Transactions on Systems, Man and Cybernetics **15**(1), 116-132, 1985, doi 10.1109/TSMC.1985.6313399.

[3] M. Sugeno, G. Kang, "Structure identification of fuzzy model", Fuzzy Sets and Systems, **28**(1), 15-33, 1988, https://doi.org/10.1016/0165-0114(88)90113-3.

[4] A. Benzaouia, A. El Hajjaji, "Advanced Takagi-Sugeno Fuzzy systems, Delay and saturations", Studies in Systems, Decision and Control 8, 2014, Springer books, https://doi.org/10.1007/978-3-319-05639-5.

[5] K. Tanaka, H. Wang, "Fuzzy Control Systems Design and Analysis: A Linear Matrix Inequality Approach", Ed Hardcover, John Wiley and Sons, Inc., New York, 2001, doi:10.1002/0471224596.

[6] K. Tanaka, T. Ikeda, H. Wang, "Fuzzy regulators and fuzzy observers: relaxed stability conditions and LMI based designs", IEEE Transactions on Fuzzy Systems **6**(2), 250-265, 1998, doi: 10.1109/91.669023.

[7] K. Tanaka, T. Hori, H. Wang, "A Multiple Lyapunov Function Approach to Stabilization of Fuzzy Control Systems", IEEE Transactions on Fuzzy Systems **11**(4), 582-589, 2003, doi: 10.1109/TFUZZ.2003.814861.

[8] T. Guerra, A. Kruszewski, L. Vermeiren, H. Tirmant, "Conditions of output stabilization for nonlinear models in the Takagi-Sugeno's form", Fuzzy Sets and Systems **157**(9), 1248-1259, 2006, ttps://doi.org/10.1016/j.fss.2005.12.006.

[9] S. Bezzaoucha, H. Voos, M. Darouach, "A contribution to Cyber-Security of Networked Control Systems: an Event-based Control Approach", in 3^{rd} International Conference on Event-Based Control, Communication and Signal Processing, Funchal, Madeira, Portugal, 2017, doi: 10.1109/EBCCSP.2017.8022805.

[10] S. Bezzaoucha Rebaï, H. Voos, "Stability Analysis of Power Networks under Cyber-Physical Attacks: an LPV Descriptor Approach", in the 6^{th} International Conference on Control, Decision and Information Technologies, Paris, France, 2019, doi: 10.1109/CoDIT.2019.8820425.

[11] S. Bezzaoucha Rebaï, H. Voos, "Simultaneous State and False-Data Injection Attacks Reconstruction for NonLinear Systems: an LPV Approach", in the 3^{rd} International Conference on Automation, Control and Robots, ICACR2019, Prague, Czech Republic, 2019, doi: 10.1145/3365265.3365280.

[12] S. Bezzaoucha Rebai, "A Cyber-Security Contribution to Estimation and Event-Based Control Scheduling Co-Design for Polytopic and T-S Fuzzy Models Using A Lyapunov Approach", Springer Nature - International Journal of Fuzzy Systems, 2022, https://doi.org/10.1007/s40815-022-01282-3

[13] S.Bezzaoucha, B. Marx, D. Maquin, J. Ragot, "Nonlinear joint state and parameter estimation: Application to a wastewater treatment plant", Control Engineering Practice, **21**(10), 1377-1385, 2013, https://doi.org/10.1016/j.conengprac.2013.06.009.

[14] L. Xiaodong, Z. Gingling, "New approaches to H_∞ controller design based on fuzzy observers for T-S fuzzy systems via LMI", Automatica **39**(9), 1571-1582, 2003, doi: 10.1016/S0005-1098(03)00172-9.

[15] M. Oudghiri, M. Chadli, A. El Hajjaji, "One step procedure for robust outpur fuzzy control", in the 15^{th} meditarranean conference of control automation, Athens, 1-6, 2007, doi: 10.1109/MED.2007.4433964.

[16] M. Kocvara, M. Stingl, "PENNON – a code for convex nonlinear and semidefinite programming", Opt. Methods and Software, **18**(3), 3170-333, 2003, https://doi.org/10.1080/1055678031000098773.

[17] M. Kocvara, M. Stingl, "PENBMI, Version 2.0", See www.penopt.com for a free developer version, 2004.

[18] D. Henrion, J. Lofberg, M. Kocvara, M. Stingl, "Solving polynomial static output feedback problems with PENBMI", Proceedings of the 44th IEEE Conference on Decision and Control, 7581-7586, 2005, doi: 10.1109/CDC.2005.1583385.

[19] J. Lo, M. Lin, "Observer-based robust H_∞ control for fuzzy systems using two-steps procedure", IEEE Trans Fuzzy Systems, **12**(3), 350-359, 2004, doi: 10.1109/TFUZZ.2004.825992.

[20] F. Torres, A. Rahbi, D. Lara, G. Romero, C. Pégard, "Fuzzy State Feedback for Attitude Stabilization of Quadrotor", International Journal of Advanced Robotic Systems, InTech, 2016, https://doi.org/10.5772/61934.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 80-88 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Human-Centered Design, Development, and Evaluation of an Interface for a Microgrid Controller

Mohammed Mahfuz Hossain[1], Thomas Ortmeyer[*,2], Everett Hall[2]

[1] *Engineering Science Dept., Clarkson University, Potsdam, NY 13699, USA*

[2]*ECE Dept. Clarkson University, Potsdam, NY 13699, USA*

ARTICLE INFO

Article history:
Received: 10 February, 2023
Accepted: 15 April, 2023
Online: 15 May, 2023

Keywords:
Microgrid
Human factors testing
Situational awareness
Interface design

ABSTRACT

Many millions of people have adverse effects on their lives, both socially and economically, when a power outage occurs. Along with other electrical events, the lack of Situational Awareness (SA) is one of the root causes of power system outages. In order to promote adequate situational awareness, both power system and microgrid interfaces should communicate the necessary information in a helpful format at the right time. It is particularly difficult to present this information to microgrid operators in an accessible and timely manner. A human-centered design approach is used to develop two human-machine interfaces for the Potsdam, NY microgrid project. A detailed description of the process is provided in this extended paper.

1. Introduction

A reliable, quality power supply is vital to our standard of living on a day-to-day basis. Electric power supply interruptions have great negative impact on our social and economic life. Significant research has focused on identifying the root causes of power outages and have reported Situational Awareness (SA) as one of the key causes. In this regard a study has been done to design and develop interface for Potsdam Microgrid. An overview of the main study results is presented in [1]. In this paper, the step-by-step process of conducting the experiment is described in detail.

2. SA and its challenges

One of the most widely used SA models can be found in [2]. Where the author defined SA as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their "comprehension" of the current situation, and finally, level three is the "projection" of the future state. Further details of the three-level SA model are found in [3].

Power system operations require real-time assessments, monitoring, and activity control. Most importantly, power system operations require the coordination of electricity production at thousands of generators, long transmission line networks, and tens of thousands of electrical buses. All of this is required to ultimately deliver electricity to millions of users by means of the distribution network [4]. The complexity of the power system infrastructure is continuously rising, and utility companies have been increasingly facing challenges to make decisions in a timely and accurate manner. [5] and [6] identified SA challenges for power transmission and distribution which seriously downgrade operators' SA.

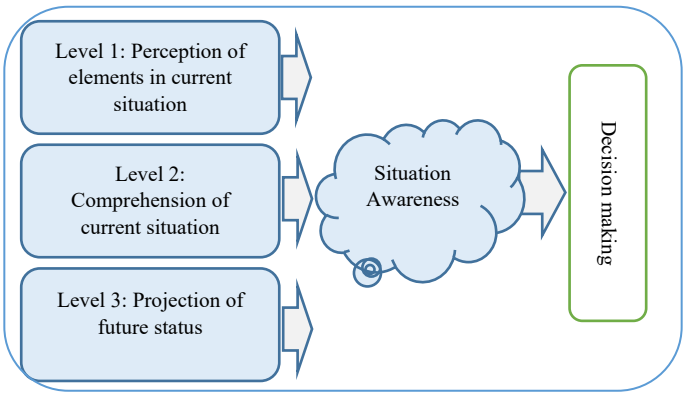


Figure 1: Model of Situation awareness adapted from [3]

Attentional tunneling involves situations when even all the needed information is presented, it is not fully attended by the person monitoring the system. The attention narrows as the scanning behavior is dropped.

Data overload is related to the volume and rapid change of data that creates an information intake pace that is hard to assimilate. Typically, the operator has to scan through thousands of pages of

*Corresponding Author: Thomas Ortmeyer, tortmeye@clarkson.edu

SCADA data tables. When coupled with weather reports, alarms, contingency analysis data, and state estimator calculations, operators tend to become flooded with data, creating severe losses in SA [7].

Requisite memory trap. It is related to the capacity limitations to retain information in the working memory. Control room operators have to monitor more than twenty different pieces of information continuously [8].

Workload, anxiety, fatigue, and other stressors. System operators need to seek data, sort through what is available, and integrate information for decision-making in time-pressured environments. This requires high mental workloads, fatigue, and other stressors, leading to an increased number of opportunities for errors.

Misplaced salience occurs when software systems fail to highlight the most critical information. Operators have to visually fight the allure of several flashing lights and a wide variety of colors to identify the relevant information needed.

Errant mental models. Mental models tell how to combine information taken from different places. Poor comprehension and projection of the situation result when using incomplete or wrong mental models to interpret information.

Complexity creep. The more system features and governing rules, the more difficult it is to understand how the system works, slowing down the ability to extract and interpret information.

Out-of-the-loop syndrome. Highly automated systems can leave operators with low awareness of the state of the system. A germane example is the August 2003 blackout [9], where operators did not realize that the diagnostic tools were off-line and not updating in real-time.

There is a need for a good SA design solution to provide support for human limitations and avoid known problems with human information processing. This study developed and tested interface to promote/support operators SA.

3. Interface and Simulator Design

Situation Awareness (SA) is the key to user-centered design [10]. Endsley defined SA as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" [3] pg. 13. The user interface design process is adapted from [11], shown in Figure 2.

3.1. Requirements analysis

In the first phase of interface design, it is necessary to transform the goals of the project/users into specific system requirements. Typically, the requirements are collected using cognitive task analysis (CTA) [11] , [12] presented a form of CTA called a goal-directed task analysis (GDTA) to identify the goals, decisions, and SA requirements of operators. GDTA is used/necessary to understand the interface design specifications with a detail description of not only the specific data operator needs but also indicates the way the operator integrates the data to develop an understanding of the current situation and make a decision [13]. To collect the microgrid specific requirements a set of questions are prepared. In this study, a number of practicing power system operators were interviewed in preparation for developing these questions. The developed questions are:

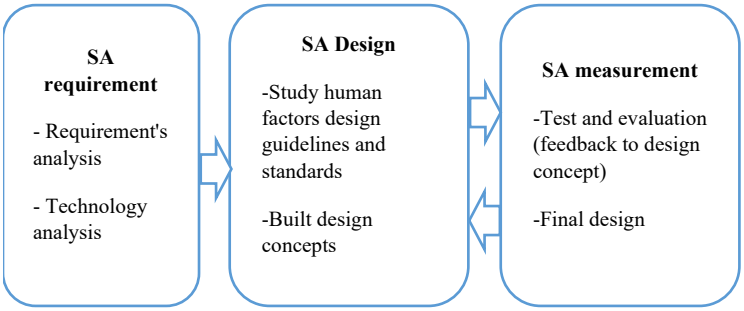


Figure 2: Interface design process [11]

1. What are your main goals during normal operations?
2. What are the functions of the human operator under normal operations? (Daily activity) or what tasks do you need to do to achieve your goals?
3. What do you need to know to achieve your goals? Or what would you ideally like to know? How do you get that information? How do you use that information?
4. What information elements are necessary for properly monitoring the system? How many displays (or any other information sources) do you need to constantly monitor?
5. Do you think these displays (or information sources) are enough for maintaining a safe operation? Or is there anything else that needs to be monitored to improve reliability? What do you do if you do not have the information needed to completely assess the situation (e.g., how do you seek that information and how often does it happen?)
6. Do you think people can make mistakes because the system does not provide all the information needed or the information displays are too complex? If so, what types of mistakes or consequences are likely to occur? Are there any improvements you would like to make?
7. How confident are you that the information you are receiving is reliable and valid? How does that affect your decisions?
8. How do you define normal condition operations? (e.g., what are the threshold values/limits?
9. How do you identify (or become aware of) a problem/contingency? Does the system alert you in some way? Is there an alarm system (audio, visual, or both)?
10. Do you think these problem identification techniques/methods are good enough? Or do you need better system notifications? Are there any improvement recommendations?
11. If there are multiple types of alarms, is there a prioritization mechanism for addressing problems?
12. How do you determine what actions need to be done to solve the problem? (e.g., do you have to do manual computations? Are there any particular skillset needed or specialized knowledge that helps the decision-making?)

13. What types of problems typically occur? What actions do you perform to solve each type of problem?
14. How much time do you have to (or is expected for you too) solve the problem?
15. Do you think that solving the problem through this method or series of actions are good enough? Or is there anything you need for facilitating your job (e.g., do you have all the information/tools you need to make decisions quicker)? Could you please elaborate on that?
16. What is the workforce needed for smooth operations under normal conditions? (area of operation)
17. Have you ever felt the need for getting information outside of your working area?
18. How would be the ideal way of getting that piece of external information? Via phone? Or Do you think all information should be available for you through electronic communication (e.g., computer software displays)?
19. What is the skill level necessary for the operator (BS, MS or high school grad)?
20. What is the required training for the operator?
21. Do you use any manual operations or documented operation guidelines?
22. Which decisions are automated and, which ones cannot be automated or require final operator input?
23. Do you have anything to support awareness of future projections? (i.e.: that indicates near future worst-case contingencies, trending information)

An interview is not sufficient to create a GDTA table properly. Therefore, this study also relied on literature to identify the goals of power system operators. In [14], the authors interviewed Specialist Reliability Analysis and Operation (SRAO), and the Reliability Coordinator/System Operator (RCSO) positions from two U.S. power companies to develop GDTA. Figure 3 shows the overall goal of the operator: *to keep all elements and voltages within limits in real-time and for first contingency* [14].

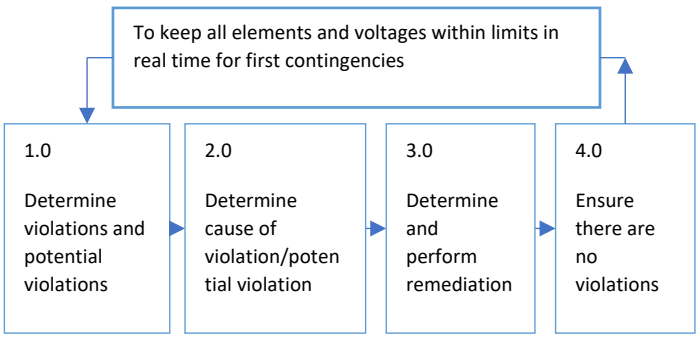


Figure 3: The overall goal of power system operators [14].

Contingencies are an unexpected failure of any system component, for example, transmission line, generator, circuit breaker or other electrical elements. By planning for first contingency means, the operator can attempt to prevent the uncontrolled cascading loss of system elements that results in widespread load interruption. Four primary goals were identified as well under the overall goal. A primary goal is determining if any violations have occurred. A second primary goal includes the cause of a violation or potential violation. The next two primary goals are to remediation of any violation or possible violation.

3.2. Design alternatives

SA is necessary for effective decision making that will lead system operators to take appropriate corrective actions. Initially, four preliminary interface display conceptss were developed that were based on the principles and best practices identified from the literature review.

Alternative #1 is shown in Figure 4. Some of the key features are:

- Switches and breakers are shown in a conventional way
- Bus names and pu values are shown to the side of each bus icon
- Loads are shown in MW and MVAR units
- Dynamically sized pie charts show line loading percentages in per cent of the thermal limit.

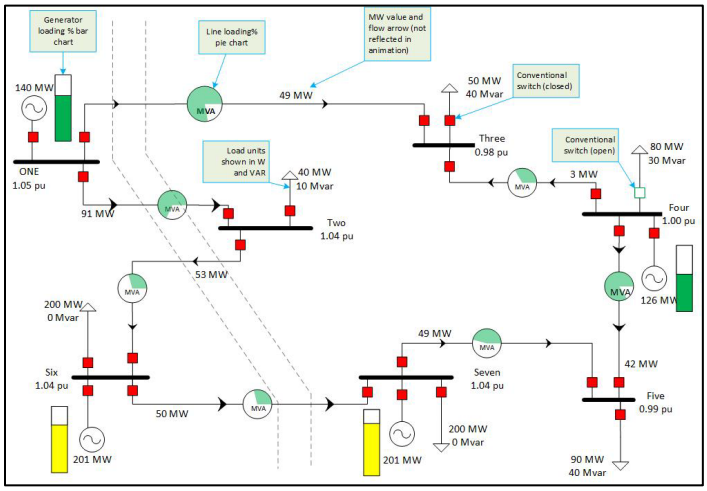


Figure 4: Alternative#1 display overview

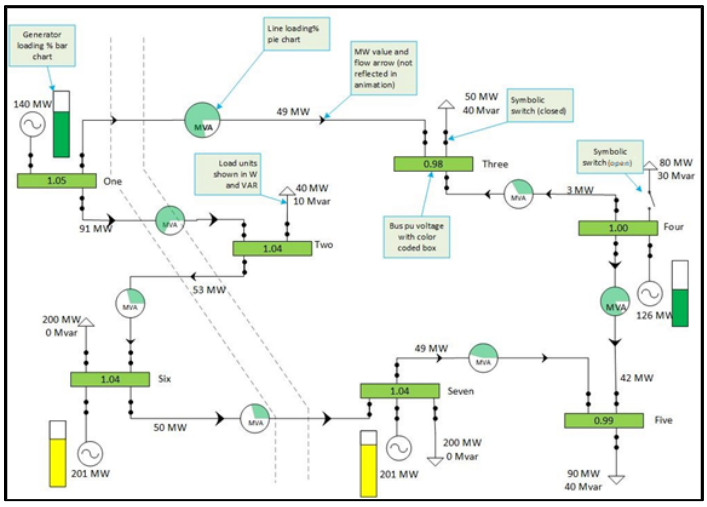


Figure 5: Alternative#2 display overview

Alternative #2 is shown in Figure 5. Key features are:

- Switches and breakers are shown with SPST switch symbols to discern open or close status easily
- Buses are presented in colored boxes (This color could be reversed based on the operator's choice)
- Bus per unit values are written inside the box
- Loads are shown in MW and MVAR units

Alternative#3 is shown in Figure 6. Some of the key features are:

- An overview box displays key information
- Pie charts show generator loading in percent of rating. The pie chart switches color depending on the situation (normal, alarm, etc.)
- Per Unit bus voltage are shown with a trend line
- Line names are displayed, and the operator has the option to display loading in percent, MVA, or amps.
- The set of overloaded lines are displayed in a single box, as are the set of heavily loaded lines.
- Loads are shown with MW or MVAR trend line
- Last update time is also logged

Alternative #4 is shown in Figure 7. Some of the key features are:

- The overview box has key information, such as overloads, trip and recent alarms
- Selecting an alarm will highlight the fault location/s on the coordinated views
- Generator that are out of service for maintenance are grayed out
- Spinning reserve will be shown in pink color
- Loads are shown with MW or MVAR trend line

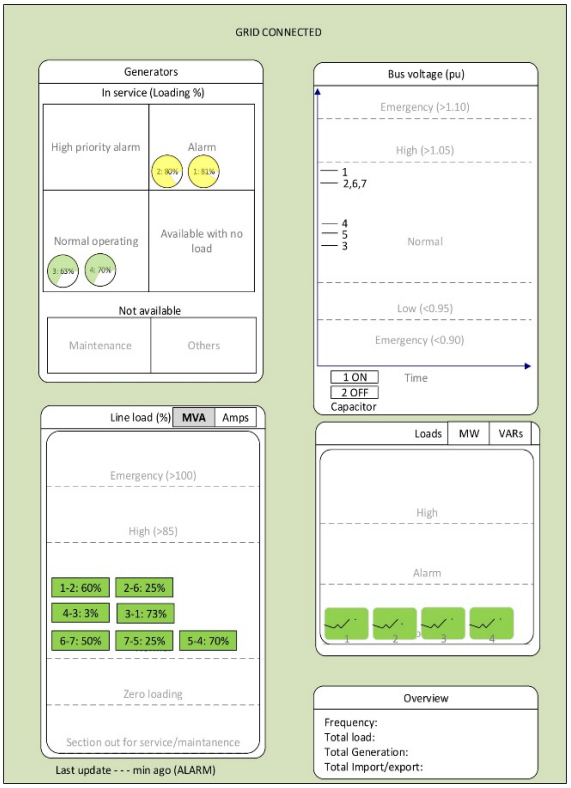


Figure 6: Alternative#3 display overview

3.3. Feedback from the experts

These four concepts were presented to power grid operators and supervisors from our industry partners, using simulated screens/animations. A brief feedback survey was done with power system operators. We received four survey responses from these participants. The purpose of this survey was to select one concept

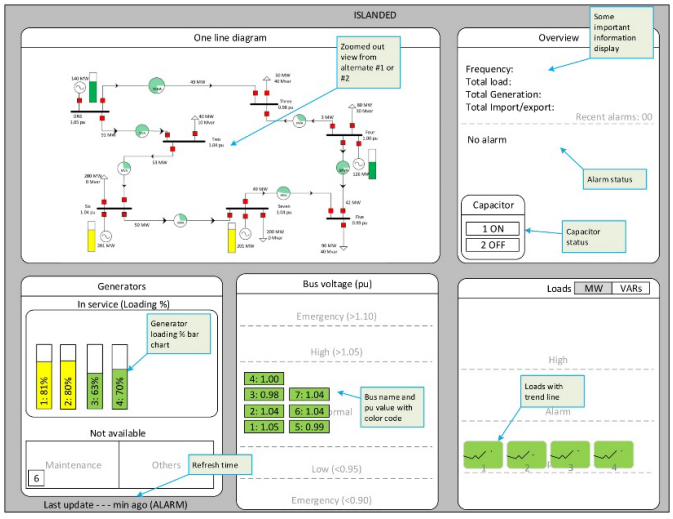


Figure 7: Alternative#4 display overview

and upgrade that interface to integrate it with load flow simulation software to make a fully real time simulation of the microgrid. Survey questions for these four alternatives are as follows:

- How satisfied are you with the interface design?
- Do you think the overview display screen would be appropriate for small screens, e.g. tablet or mobile devices?
- What other important information do you think is necessary to display but is missing in the overview display?
- What other information would be nice to have in the overview display?
- Any suggestions for improvement
- Preference for the design alternatives
- Overall suggestions
- How do you want to view bus voltage values?

Responses: Overall average score of satisfaction for the alternative#1, 2, 3, and 4 are 4, 3.5, 2.25, and 4.25 respectively. As a result of this process, parts of Alternatives #1 and #4 were selected for further study, and revised based on operator feedback. For Alternative #2, it was found that professionals do not want to change from conventional displays for circuit breakers and buses i.e., red for closed, and green for open breakers.

These revised concepts were then developed using LabVIEW software for the HMI and Matlab code for the system state and power flow analysis to develop a real time simulator that was then evaluated using SA measures.

3.4. Real Time Simulator

Among the alternatives presented in an earlier section, Alternatives #1 and #4 were selected to move forward for testing. Two HMI's were created for the proposed seven-bus Potsdam, NY.

In both cases, line loading pie charts were replaced with bar charts. The reason behind this is, pie charts are not easy to interpret by the human cognition process. "Pie charts force us to compare either 2-D areas or the angles formed by each pie. Our visual perception handles neither of these comparisons easily or accurately." [15]. Also, operators control each circuit breaker from the one-line diagram. Selection of a given circuit breaker opens a pop-up window to confirm the action. As the status of the microgrid is critical, the one-line also has a block at its center that confirms the status (grid connected or isolated).

Interface A (Figure 8) consists of five blocks of information. These are arranged to provide an intuitive understanding of the status of the microgrid. The first block presents a conventional one-line diagram.

The second block on the right side of the interface presents overview information of the total microgrid. This section provides total microgrid frequency, generation, import/export, and load. In addition, this section presents alarms. One important intended feature of this alarm list is a coordinated view, meaning the selection of one alarm should show/highlight the affected elements in the other blocks of the interface. In the next block (bottom middle) of the interface, the microgrid load is presented in a load curve. The load curve is provided to help in users SA level 3 (projection of the system condition). The load block shows the load curve trend, and includes operator-initiated load shedding capability on the individual busses. Finally, the generator status block presents both generator status and loading as well a capacitor status.

Interface B figure 8, has the same color conventions and diagram symbols as Interface A. Interface B has a single block design, with the one-line diagram enlarged in this block. Load and voltage data are displayed on the one line. Alarms are shown directly on the one line. In Figure 9, the Bus 1 generator is overloaded. As a result, the generator bar graph is red. When it becomes overloaded, the bar graph will pulse, and an audible alarm will sound until it is acknowledged. Operator actions for generator, capacitor and load shedding are initiated directly from the one-line in this interface, rather than from a subpanel. In addition, detailed views of the generator (Figure 10) and bus status (Figure 11) are available for selection by the operator to show additional detail that is in the subblocks of Interface A.

In both HMI's, an overload of violation causes an alarm to sound. For generator and line overloads, the bar graph turns red and blinks. Also, the red "Alarm On" turns on. For bus voltage violations, the bus name turns red and blinks to indicate the alarm state. When the alarm comes on, the operator mutes the alarm by pressing the "Mute Alarm" button. When the alarm is muted, this button turns orange and displays "Unmute Alarm". When the system is in an alarm state, the operator mutes the alarm, and then resolves the contingency by adjusting generator outputs, switching lines or capacitors in or out, or implementing load shedding from the HMI interface. Contingencies studied include separating from the bulk power grid, line trips and generator trips.

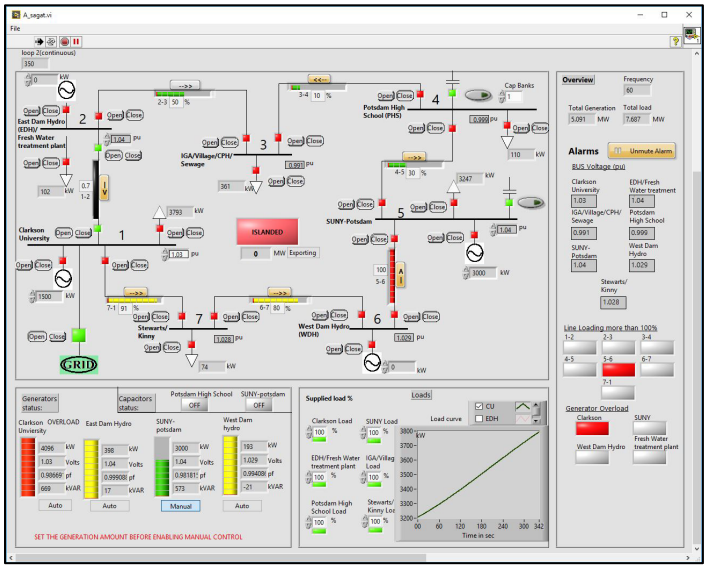


Figure 8: Potsdam microgrid interface-A using LabVIEW software.

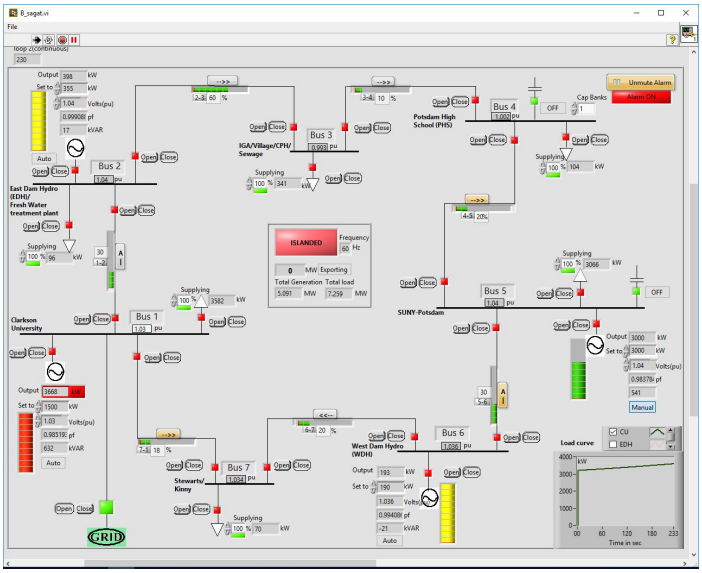


Figure 9: Potsdam microgrid interface-B using LabVIEW software.

These two HMI options were implemented in the LabView software. The Potsdam Microgrid was represented in Matlab with using a Newton-Raphson load flow. TCP/IP communications were used to connect the load flow and HMI interfaces. The load flow had an update rate of 1 second, and it was able to operate in real time. Each second, the load flow algorithm received updated breaker status and generation and load inputs from the HMI. It then ran the load flow, and sent bus voltages and line and generator power flows back to the HMI for display and alarm generation. The user was able to open and close circuit breakers and adjust generator and load settings directly in the HMI.

The next section presents the situational awareness assessment methods evaluation techniques used in this study, and the SA results that provide comparative analysis of the two interfaces.

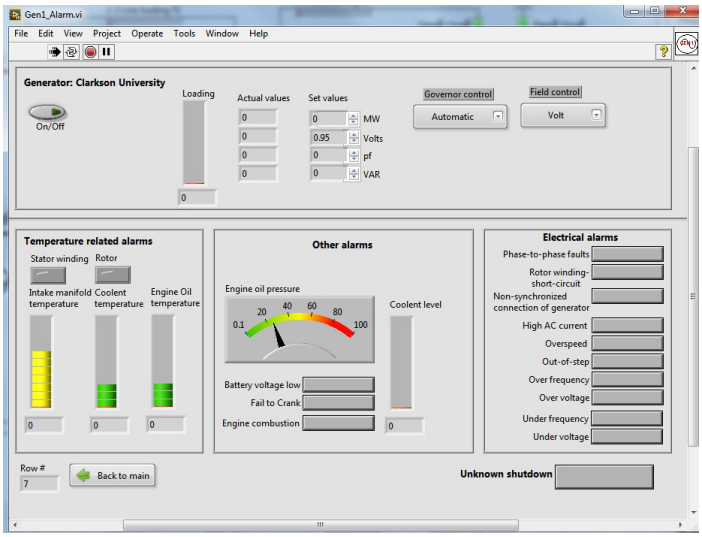


Figure 10: Detail view of the generator window.

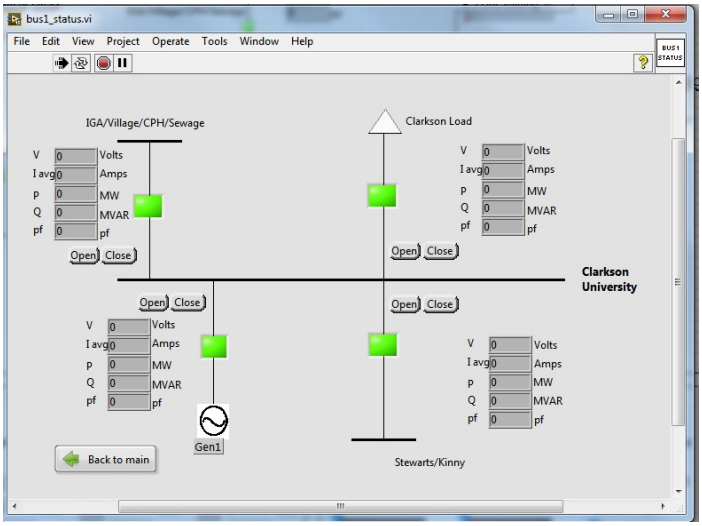


Figure 11: Detail Bus view window.

4. SA assessment/Human factors testing

Assessment or evaluation is an important part of any SA design process. Evaluation is important to avoid any unforeseen issues that can negatively impact an operator's SA. SA is an internalized mental concept, and adequately assessing SA can be difficult. In [16], the authors suggested two evaluation methods: direct and indirect measures of SA (presented in Figure 12). They proposed some techniques for those measures as well. In this study, SAGAT (Situation Awareness Global Assessment Technique) is used as a direct measure and the performance measure method is used as an indirect measure for the microgrid interfaces.

In [17], the authors have used 28 students to participate in their study to assess their simulated interfaces. For the microgrid interface assessment, 28 undergrad and/or graduate students participated in the study. Students were selected from those who have completed or were enrolled in the courses Power Transmission and Distribution, Power System Engineering, High-Voltage Techniques and Measurements, or Energy Conversion. Student subjects are compensated $20/hr for their participation in the study.

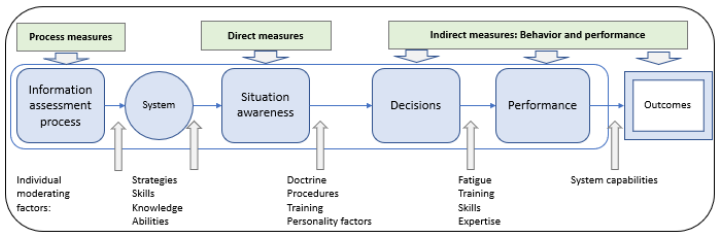


Figure 12: Approaches to SA measurement. Adapted from [16]

4.1. The indirect performance measure of SA

Indirect performance measures consist of techniques that assess SA based on the operators' overall performance. This approach assumes a direct relationship between SA and performance. However, good SA is necessary but not sufficient for good performance. A person might have good SA, but lack the skill, training and/or knowledge that is required for good performance. This section describes the human factors testing methodology used to compare the operators/students' performance between the two interfaces developed using LabVIEW.

In this study, a training manual and video were developed. Participants were given hands-on training with a sequence of 5 trial cases. Half of the participants were trained with interface-A, another half trained with interface-B. Finally, they were tested with 5 trials in each interface (5 x 2 = 10 trials for each participant). Action times were recorded for evaluation as a measure of their performance with the interfaces. Each student received the same contingency (line trip) sequence. In each trial, thirty seconds into the simulation a line trip/outage occurs. This trip causes an overload on one or more of the transmission lines or generators and/or voltage violations. Any of this cause an audible alarm to sound. This event requires three tasks of the user:

1- acknowledge the violation
2- solve each violation through operator action
3- confirm that the system does not have any violation

For the Potsdam microgrid, this study analyzes three measures:

- time taken to acknowledge violation,
- time taken to solve the violation and
- time taken to confirm that there is no violation now/system is a normal state.

Operators/subjects can solve the violations by increasing/decreasing generation kW output, switching capacitors, and/or load shedding. Faster response times for each measure indicate better operator performance.

With each test participant, the time taken to mute the alarm and the time to solve the contingency were recorded. While evaluating responses, the study excludes all the measurements that took 60 seconds or more to solve the contingency. In addition, the test recorded the time to acknowledge no violation scenarios (time is taken to click the unmute button after the system came back to normal state). The study excludes those test points that do not have unmute time (could be for the technical issue) or where participants took more than 20 sec to unmute. Considering all the facts 121

observations for Interface A and 127 observations for Interface B were analyzed.

Table 1: Average times in seconds taken for each interface.

	Interface-A (seconds)	Interface-B (seconds)
Mute time (time between any alarm ON to click MUTE button)	1.73	1.54
Solution time (time between any alarm ON to solve the contingency (all alarm OFF))	19.51	21.51
Acknowledge time (time between all alarm OFF to click the UNMUTE button)	3.63	3.28

From the Table 1 results it is evident that Interface B took less time in both mute time and acknowledgement time. However, Interface A took slightly less time in solving the contingency. Interface A has a dedicated subpanel for alarms to the right side of the screen in Figure 8, with the display having individual indicators for bus voltage and line and generator overload violations. An enlarged version is shown in Figure 13. The Alarm light and the Mute-unmute pushbutton are in this subpanel. Interface B has the Alarm indicator and mute/unmute button in the upper right corner of the one-line diagram that is show in Figure 9. The overloaded device or bus voltage violation are noted by flashing bar graphs or lights on the one line at the location of the violation. An enlarged view of this part of the one-line is also shown in Figure 13. This difference between the alarm display and mute/unmute button location are considered to be the most likely cause of the differences in test results between the HMI's.

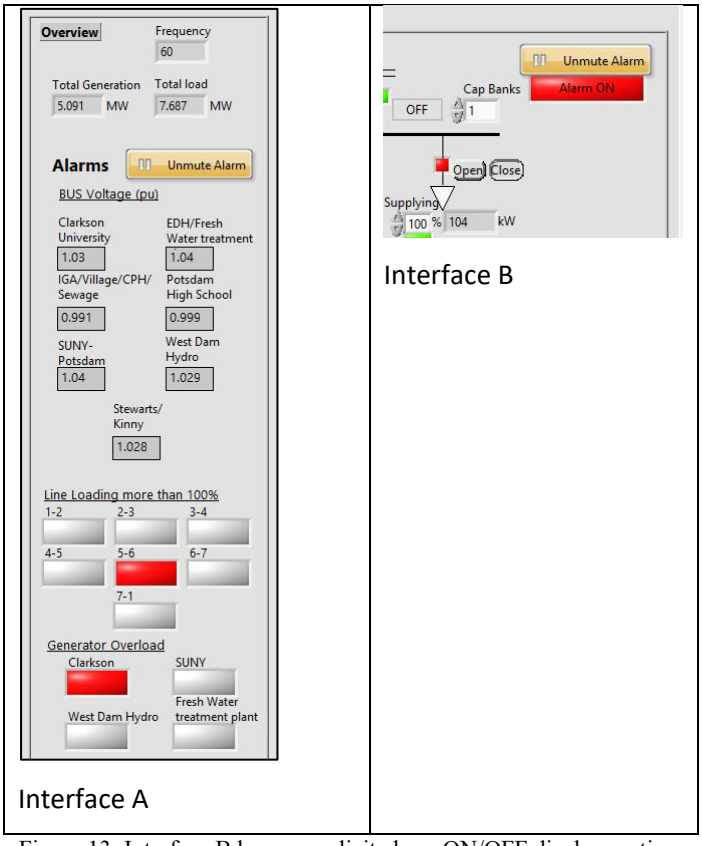


Figure 13: Interface B has an explicit alarm ON/OFF display section.

4.2. The direct measure of SA

SAGAT is a direct measure of SA. This is the human in the loop testing system. The simulation was frozen at several points in time. During this period, a series of questions are asked to the operators to determine his or her knowledge of the situation at the specific moment. Some advantages of the SAGAT technique are:

- it assesses global SA,
- avoids a retrospective recall,
- minimizes biasing,
- performs in a realistic environment.

On the contrary, some disadvantages of the SAGAT technique are that it requires stoppages in the scenarios. This may have a negative impact on the real-time scenarios. As this study is using simulated scenarios, the stoppages are not a factor.

Data were collected using a modified version of the SAGAT technique adopted from [18]. Participants got about five minutes of training to get used to the system/interface. Then simulated scenarios of line trip/outage followed by some voltage/line/generator overload occurred. In these cases, participants were not required to do any activity other than monitoring the system condition/s. A maximum of five halts was used to collect data. Each halt consists of ten queries, with a maximum time limit of two minutes permitted per halt. Each participant was surveyed using an online survey instrument ('Google Forms'). Per standard SAGAT procedure, participants are barred from viewing operational information during the halt. Responses from the first halt were considered a training response. Each query was set based on the goals, decisions, and information requirements of the operators. Selected SAGAT questions and their level of SA's are in (parenthesis):

Q1. What is the approximate total load in MW? (L1)

Q2. Is the Microgrid currently importing OR exporting energy? (L1)

Q3. Are there any lines currently loaded between 80% - 100% (in Yellow condition)? (L2)

Q4. At this moment which Buses are in voltage violation (pu lower than 0.95 OR higher than 1.05)? (L2)

Q5. Which buses have capacitor banks? (L1)

Q6. Currently how many lines are in an outage state? (L2)

Q7. How many lines are overloaded at this time? (L2)

Q8. Are there any generators currently loaded over 100% (in Red condition)? (L2)

Q9. Currently how many generators are set to manual control? (L1)

Q10. Within the next 10 min, what do you think is going to happen about the load? (L3)

The actual system conditions are recorded at the time of the simulated halt. The accuracy of the responses are compared with the actual scenarios at that time. The time delay from the beginning of the simulation to each halt in different scenarios are presented

in Table 2, 3 and 4 present SAGAT results from the experiment. Total response count is less for halt#1 and halt#5 because each participant with each display faces four halts. The very first response was considered as a training response, regardless of the interface type. Thus, if anyone is presented with 'interface A' first then the participants were given 4 more halt (up to 350 - 355 seconds). However, then for the 'interface B' the participants were given four halts (up to 268-277 seconds). The following questions were developed and used in this study to conduct the SAGAT measure:

Question 1 (What is the approximate total load in MW) results indicate better performance of 'interface A' than 'interface B'. Total load is displayed in 'interface B' right in the middle of the screen whereas in 'Interface A' in one corner. It is possible that the positioning of this information on the screen can have an influence on this result.

Question 2 (Is the Microgrid currently importing OR exporting energy) results shows almost the same performance in both interfaces. It is observed that both interfaces have a similar look and positioning of this information. Question 3 (Are there any lines currently loaded between 80% - 100% (in yellow condition)?) 'Interface A' has significant preference over 'interface B'. 'Interface A' has a separate section for alarm display. In contrast, participants had to scan all over the 'interface B' to get the pieces of information (check Figure 8 and 9. Question 4 At this moment which Buses are in voltage violation (pu lower than 0.95 OR higher than 1.05)?). The results demonstrate that 'interface B' better performed than 'interface A'. 'Interface A' has a separate alarm section to display voltage violation information. However, 'interface B' displays flashing red light in voltage violation scenarios. Question 5 (Which buses have capacitor banks?) 'Interface A' has a clear preference over 'interface B'. Observations are made that a separate section of capacitors in 'interface A' made the privilege over 'interface B'. Question 6 (Currently how many lines are in an outage state?) same explanation as described under Question 3. Question 7 (How many lines are overloaded at this time?) same explanation as described under Question 3. Question 8 (Are there any generators currently loaded over 100% (in red condition)?) same explanation as described under Question 3. Question 9 (Currently how many generators are set to manual control?) 'Interface A' has little better performance over 'interface B'. Notes are made that 'interface A' has the information's displayed together at a place. In contrast, in 'interface B' participants had to scan through the full display to get the pieces of information. Question 10 (Within the next 10 min, what do you think is going to happen about the load?). 'Interface B' has little better performance over 'interface A'. Note to be made is that, in terms of responses count this is negligible.

Table 2: Timing of the halts during the experiment

Halt Number	Time delay of a halt in the simulated scenario, (seconds)
1	70-75
2	150-152
3	230-240
4	268-277
5	350-355

Table 4 depicts that both on level 1 and level 2, 'interface A' performed better than 'interface B'. However, level 3 results show almost similar performance on both the interfaces. However, note that there is only one level 3 question (Q10) in this study, and further SA testing at level 3 is indicated.

Table 3: SAGAT Results by Question

	Interface A			Interface B		
Questions	Wrong response	Total response count	% Error	Wrong response	Total response count	% Error
1	3	105	2.9	11	105	10.5
2	15	105	14.3	16	104	15.5
3	21	105	20.0	35	104	33.5
4	22	105	20.9	17	105	16.5
5	9	105	8.6	15	104	14.4
6	12	105	11.4	20	104	19.2
7	14	105	13.3	24	104	23.1
8	3	105	2.9	9	104	8.7
9	18	105	17.1	21	104	20.2
10	17	105	16.1	16	104	15.4

Table 4: SAGAT results showing SA level performance.

		Interface A			Interface B		
SA level	#of Questions	wrong response	Total events	% Error	wrong response	Total events	% Error
L1	4	45	420	10.71	63	417	15.11
L2	5	72	525	13.71	105	521	20.15
L3	1	17	105	16.19	16	104	15.38

5. Conclusion

Both power system and microgrid interfaces should communicate the necessary information in a helpful format at the appropriate time in order to promote adequate situational awareness. For the Potsdam, NY microgrid, four HMI concepts were developed, and two human-machine interfaces were simulated and tested using a human-centered design approach. This paper provides detailed information about the design, development, and evaluation process. Both direct and indirect measures are used to evaluate the designed interface. Study results underscore the importance of both direct and indirect measures while doing human factors testing. The indirect/performance measures showed better performance of Interface A in solving the contingency, while Interface B scored better in the muting and acknowledging time tests. The direct/SAGAT method provided further evidence participant performance was more accurate with Interface A than Interface B.

Acknowledgement

This work was supported by the National Science Foundation, Project 1534035 "PFI:BIC Developing Advanced Resilient Technology to Improve Disaster Response Capability."

References

[1] M. H. Mahfuz, T. Ortmeyer, and E. Hall, "Development of a Microgrid Controller Interface Using Human-Centered Design Approach," in IEEE Power and Energy Society General Meeting, 2022. doi: 10.1109/PESGM48719.2022.9917191.

[2] P. M. Salmon et al., "What really is going on? Review of situation awareness models for individuals and teams," Theor Issues Ergon Sci, vol. 9, pp. 297–323, 2008.

[3] M. R. Endsley and D. G. Jones, "What Is Situation Awareness?," in Designing for Situation Awareness, CRC Press, 2011, 13–30. doi: doi:10.1201/b11371-4 10.1201/b11371-4.

[4] J. D. Weber and T. J. Overbye, "Voltage contours for power system visualization," IEEE Transactions on Power Systems, **15**, 404–409, 2000, doi: 10.1109/59.852151.

[5] M. R. Endsley and E. S. Connors, "Situation awareness: State of the art," in 2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century, IEEE, Jul. 2008, 1–4. doi: 10.1109/PES.2008.4596937.

[6] E. S. Connors, "Situation awareness for the power transmission and distribution industry," in EPRI's XIII annual power switching safety and reliability conference and seminar, 2009.

[7] C. Tu, X. He, Z. Shuai, and F. Jiang, "Big data issues in smart grid – A review," Renewable and Sustainable Energy Reviews, **79**, 1099–1107, 2017, doi: https://doi.org/10.1016/j.rser.2017.05.134.

[8] C. Schneiders, J. Vanzetta, and J. F. Verstege, "Enhancement of situation awareness in wide area transmission systems for electricity and visualization of the global system state," in 2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), 2012, 1–9. doi: 10.1109/ISGTEurope.2012.6465665.

[9] G. Andersson et al., "Causes of the 2003 major grid blackouts in North America and Europe, and recommended means to improve system dynamic performance," IEEE Transactions on Power Systems, **20**(4), 1922–1928, 2005, doi: 10.1109/TPWRS.2005.857942.

[10] M. R. Endsley and D. G. Jones, "User-Centered Design," in Designing for Situation Awareness, CRC Press, 2011, 3–12. doi: doi:10.1201/b11371-3 10.1201/b11371-3.

[11] M. Endsley and D. Jones, "Design Process," in Disigning for situation awareness, CRC Press, Taylor and Francis Group, 2011, pp. 43–60.

[12] S. Chipman, J. M. Schraagen, and V. Shalin, Introduction to cognitive task analysis. 2000.

[13] M. R. Endsley, "A Survey of Situation Awareness Requirements in Air-to-Air Combat Fighters," Int J Aviat Psychol, 1993, doi: 10.1207/s15327108ijap0302_5.

[14] E. S. Connors, M. R. Endsley, and L. Jones, "Situation awareness in the power transmission and distribution industry," in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications, 2007, 215–219.

[15] S. Few, Now You See It: Simple Visualization Techniques for Quantitative Analysis, 1st ed. USA: Analytics Press, 2009.

[16] M. R. Endsley and D. G. Jones, "Evaluating Design Concepts for SA," in Designing for Situation Awareness, CRC Press, 2011, 259–284. doi: 10.1201/b11371-18.

[17] A. M. Rich, D. A. Wiegmann, and T. J. Overbye, "Visualization of power systems data: A human factors analysis," PSERC, 2001.

[18] M. R. Endsley, "Measurement of Situation Awareness in Dynamic Systems," Human Factors: The Journal of the Human Factors and Ergonomics Society, **37**(1), 65–84, 1995, doi: 10.1518/001872095779049499.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 2, 87-96 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Development and Analysis of Models for Detection of Olive Trees

Ivana Marin[1], Sven Gotovac[2], Vladan Papić[*,2]

[1]*Faculty of Science, University of Split, Split, 21000, Croatia*

[2]*Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Split, 21000, Croatia*

ARTICLE INFO

Article history:
Received: 26 January, 2023
Accepted: 04 March, 2023
Online: 11 March, 2023

Keywords:
Tree detection
Olive tree
Remote sensing
Deep learning

ABSTRACT

In this paper, an automatic method for detection of olive trees in RGB images acquired by an unmanned aerial vehicle (UAV) is developed. Presented approach is based on the implementation of RetinaNet model and DeepForest Phyton package. Due to fact that original (pretrained) model used in DeepForest package has been built on images of various types of trees but without images of olive trees, original model detection was unsatisfactory. Therefore, a new image dataset of olive trees was created using sets of images chosen from five olive groves. For neural network training, individual olive trees were manually labeled, and new models were generated. Each model has been trained on different set of images from selected olive groves. Pretrained model and new models were compared and evaluated for various test scenarios. Obtained results showed high precision and recall values of proposed approach and great improvement in performance compared to the pretrained model.

1. Introduction

It is predicted that close to 10 billion people will live on Earth by 2050 [1]. At the moment, about 37% of the total land surface is used for food production [2], and it is estimated that the necessary increase in food production between 2010 and 2050 will be between 35% to 56% [3]. Needed increase in production can be achieved by increasing the share of agricultural land and/or increasing productivity on existing agricultural land by applying the so-called precision or smart agriculture [4]. Olive (*Olea Europea*) is one of the most widespread plants and plantations in the world. Olive oil is a basic ingredient in Mediterranean cuisine, and it is popular all over the world. Worldwide, consumption of olive oil has been constantly increasing [5]. According to the latest reports of the International Olive Council (https://www.internationaloliveoil.org), worldwide olive oil production for 2020/2021 crop year was just above 3.000.000 tons. Spain is the largest producer of olives and olive oil in the world (close to 50% of world production) and EU countries in total produce around 70% of world production. In 2019, the global olive oil market size was above 13 billion US dollars, and it is projected to reach 16.64 billion US dollars by 2027, with annual growth of 3.2% during the forecast period (2020-2027) [6].

Therefore, olive trees and olive oil are economically very important for the producing countries. On the demand side, world consumption of olive oil has also witnessed a substantial growth in the course of the three past decades [7]. This makes olive growing and oil production a good choice for research and implementation of new approaches aiming to respond to the challenges in food production. Complex systems such as those in agriculture should be continuously monitored, measured, and analyzed. The above implies the use of new information and communication technologies [8]. Remote sensing is the process of detecting and monitoring physical characteristics of larger areas [9] using satellites, aircraft, and drones. Therefore, farmers don't need to physically visit all parts of the land to gather data that can be used to analyze different aspects of the crop and yield. The application of artificial intelligence and machine learning in agriculture is increasingly intensive due to its ability to understand, learn and react to different situations (based on learning) in order to increase the efficiency and quality of production.

Images collected for agricultural applications can be obtained from satellites such as ESA Sentinel-2A. However, these types of images depend on weather conditions (cloudiness) and have a low spatial resolution (Sentinel up to 10 m), which is not satisfactory for certain treatments [10]. It is to be expected that the temporal and spatial resolution will improve over time, but problems with clouds will certainly remain. The use of drones for data collection

[*] Corresponding Author: Vladan Papić, FESB, Ruđera Boškovića 32, Split, Croatia, vpapic@fesb.hr

enables higher spatial resolution, the time of recording images is determined by the user, and data can be collected even in cloudy weather. A greater number of camera types are available (RGB, multispectral, hyperspectral, thermal). The collected data are of significantly higher quality than those collected by satellite [11]. Also, UAV-based imaging implies lower operational costs compared to imaging systems on manned aircraft or satellites, so it can be considered a preferred solution for monitoring smaller regions. The collected images can be used after applying different computer vision algorithms for different types of applications, such as counting and estimating the size of trees [12,13], assessing fruit maturity [14], assessing crops [15], plant diseases [16], etc.

Importance of counting and identification of olive trees in aerial images can be explained by multiple reasons. Perhaps the most obvious reason is that the number of trees is a fundamental criterion for the access to public grants by olive tree farmers. Another reason is the fact that crop yield estimation is based on the number of trees in the orchard (along with other parameters such as number and volume of fruits). Furthermore, irrigation plans and water management are based on inventory and arrangement of the trees in the orchard [17][18]. Also, detection and localization of individual tree is prerequisite for more advanced analysis of plant health and fruit status using remote sensing technology.

Counting of trees by humans is prone to errors but, first of all, it is tiresome and time-consuming. Therefore, automatization of this process is lately in focus of research community [19]. Availability of various sources of aerial images such as high-resolution satellite images, images acquired by unmanned aerial vehicles (UAVs) combined with advanced image processing algorithms, makes this task solvable.

The availability of different sensors has enabled different approaches in the detection of individual trees during the last decade. For example, some authors use hyperspectral and airborne laser scanning (ALS) for tree detection and classification [20]. In contrast to hyperspectral sensors that can use several hundreds of narrow frequency bands (10-20 nm) for detection, multispectral sensors usually use 3 to 15 frequency bands. The width of these bands is usually slightly larger. For example, the multispectral camera used by the popular UAV DJI Phantom 4 has, in addition to the RGB sensor (visible spectrum), 5 more monochrome sensors with a width of 32 or 52 nm. Captured wavelengths are: blue (450 nm ± 16 nm), green (560 nm ± 16 nm), red (650 nm ± 16 nm), red edge (730 nm ± 16 nm), near-infrared (840 nm ± 26 nm). Images obtained from hyperspectral sensors contain much more data than images from multispectral sensors and have a greater potential to detect differences among land and water features. However, multispectral sensors are very popular for precision agriculture because they are much cheaper than hyperspectral sensors. Also, from available multispectral information, various vegetation indexes can be calculated [21]. Vegetation indexes calculated for each image pixel can be used to enhance the presence of green, vegetation features and thus may distinguish plants from the other objects present in the image [22-24]. One of the most frequently used and implemented vegetation indexes calculated from multispectral information as normalized ratio between the red and near infrared bands is the Normalized Difference Vegetation Index (NDVI). NDVI correlates with chlorophyll, which in turn correlates with plant health (Figure 1). Based on calculated NDVI and utilization of red band thresholding, the algorithm for detection of olive trees, resulted in an overall estimation error of 1.3% [22]. Jan Peters et al. proposed an object-based classification method for detection of olive trees from multi-spectral images [23]. This approach was comprised of a four-step model: image segmentation, feature extraction, classification, and result mapping. Obtained overall accuracy was 84.3%.

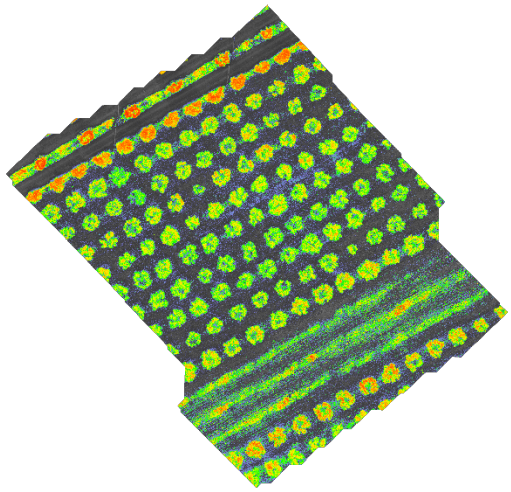


Figure 1: Example of a NDVI olive orchard image obtained with multispectral camera.

Sensor data can be used for more advanced image processing and analysis. The most popular classical methods of image analysis include machine learning (K-means, support vector machines - SVM), wavelet-based filtering, vegetation indices and regression analysis [25, 26]. In the image processing procedure, a preprocessing step is common (image segmentation, contrast enhancement and edge detection, color model selection, noise removal by filtering, feature extraction by various transformations, dimensionality reduction), after which object-based image analysis (OBIA) is performed [27].

An approach that uses classical image methods for the automatic detection and recognition of a single tree and labelling is presented in [28]. Authors pre-processed the images with the unsharp masking followed by improved multi-level thresholding-based segmentation. The circular Hough transform was applied for the identification of the circular blobs that presented single trees. Another study presented an algorithm that used RGB satellite images for a classification system. The system consists of several steps: it includes image pre-processing, image segmentation, feature extraction and classification [29]. All images were preprocessed to suppress the additive noise. Next, the region of interest was segmented from the pre-processed images using K-Means segmentation, through which statistical features were extracted and classified. The best classification results reported in that paper were achieved with Random Forest that outperformed other tested algorithms by an overall accuracy of 97.5%.

As in many other areas, deep learning has played an increasingly important role in the field of image processing in agriculture in recent years [30]. Changes in lighting, camera position and camera distance (height) to the ground significantly affect the performance of classical methods compared to methods

that use deep learning. Compared to classical methods, the approach using deep learning requires larger computing resources and larger databases of labeled images for learning. The aforementioned limitations have been overcome or largely removed in recent years due to the availability of advanced graphics processors and tools for easy labeling of learning images. Also, publicly available image databases such as PASCAL Visual Object Classes (PASCAL VOC), Microsoft Common Objects in COntext (COCO) and ImageNet, which contain thousands of object classes and millions of images and are available to researchers for model training, are also useful in this area. Deep learning models can be tuned and trained to detect fruits on these bases using transfer learning. However, it can be noted that the mentioned bases do not contain images of orchards [31].

One recent example of implementation of deep learning for identification and mapping of trees can be found in [32]. In the presented approach, the UAV RGB photograph of the forest was automatically segmented into several tree crown objects using color and 3D information and the slope model. After that, an object-based CNN classification was applied for each crown image. Classification results of the presented system showed good results in classifying seven tree classes, including several tree species with more than 90% accuracy. Another recent paper presents deep learning-based approach for estimating the biovolume of individual trees [33]. In this paper, authors used Mask R-CNN and UAV images for olive tree crown and shadow segmentation.

DeepForest is an open-source (MIT license) Python package that uses deep learning object detection networks to predict bounding boxes corresponding to individual trees in RGB imagery [34]. In order to make training models for tree detection simpler, DeepForest use the RetinaNet model [35, 36] from the TorchVision package [37]. More precisely, the model was trained on images from 40m x 40m windows obtained from 1km x 1km maps downloaded from National networks of ecological observatories (NEON) using a semi-supervised LiDAR-based algorithm to generate millions of moderate-quality annotations for model pretraining. In the next step, the pretrained model was retrained on over 10,000 hand annotations of RGB imagery from six NEON sites which further improved generalization abilities. Obtained model can be used directly to make predictions for new data or used as a foundation for retraining the model using labelled data from a new application.

Individual tree detection may not seem particularly difficult computer vision task at first, but it can be a demanding task for various reasons. Perhaps the biggest problem are closely planted trees forming joint crowns. In olive growing, this type of problems is related to extensive types of orchards (orchards with lower productivity per hectare, low mechanization level, small amount of labor relative to the area under cultivation) which is not usual for larger plantations with larger production of olive fruits and oil. Other challenges are related to varying sizes of trees in an orchard, misaligned plantation of trees, different types of soil and vegetation under trees, etc. As a result, there is quite vivid research activity in this field.

Although the use of other types of sensors, such as multispectral ones, could make the detection and labeling of individual trees simpler (as could be assumed by analyzing Figure 1), in this work we are focused on the use of RGB sensors as the most widespread and cheapest. In order to simplify and speed-up the process, only 2D information from the obtained terrain maps was used. Our approach is based on the implementation of deep neural networks for detection, more precisely on adaptation of the DeepForest package. Due to fact that original (prebuilt) model used in DeepForest package has been built on images from various types of trees but without images of olive trees, it was expected that results obtained on that model would not be good enough for implementation on olive groves. Therefore, a new image dataset of olive trees was created, individual olive trees were labeled, and new models were created. New models were built using different sets of images chosen from five olive groves. Since those olive groves had different characteristics, choice of olive groves used for model creation was important for detection results. This paper is an extension of work originally presented in conference 2022 International Conference on Software, Telecommunications and Computer Networks [38]. In this work, comparing to conference paper, a more detailed explanation of multiple models creation will be given. Also, in addition to detailed comparison of various models, analysis of the detection results for one olive orchard monitored in different seasons of year will be done.

Contributions of this paper are following: we propose a methodology for automatic olive trees detection based on adaptation of publicly available open source DeepForest package. Image dataset of five olive orchards were annotated and used for further research. Also, we present analysis on the variability of the olive trees detection results with the same neural network model in the case of olive grove surveillance at different times of year.

The remainder of the paper is organized as follows: in Section 2 the proposed methodology is described, along with the test sites description and used software tools. In Section 3, a detailed description of used procedure is given. Section 4 presents the results of the tree detection based on implementation of our models. The discussion and conclusion are then presented in Section 5.

2. Materials and Methods

2.1. Study Sites

For this study, five olive orchards were surveyed (Figure 2). Two orchards were in Mravinci, north of Split, Croatia (Mravinci01: 43°31’40.7” N, 16°30’57.8” E and Mravinci02: 43°31’37.4” N, 16°30’57.8” E). Both olive orchards at this location can be classified as extensive. They are characterized with irregular pruning and non-uniform shaping of trees. Most of the trees are free vise shaped while smaller number of trees have monoconical and globe shaped plants. Other three orchards were located in Tinj, south-west of city of Benkovac in Zadar County, Croatia (Tinj01: 44°00’49” N, 15°28’13.2” E, Tinj02: 44°01’10.6” N, 15°28’19.5” E and Tinj03: 44°00’26” N, 15°29’31.0” E). All olive trees at this location were vise shaped and rather heavy pruned. Since olive orchard at Tinj03 location is quite large (12.000 olive trees), surveillance with UAV did not cover all plants. In four separate flights (Flight01, Flight02, Flight03, Flight04), around 5% of total area was covered (around 600 olive plants detected and annotated). Each flight for this olive grove was

used for generating separate map and, in this paper, each is treated individually for the analysis.

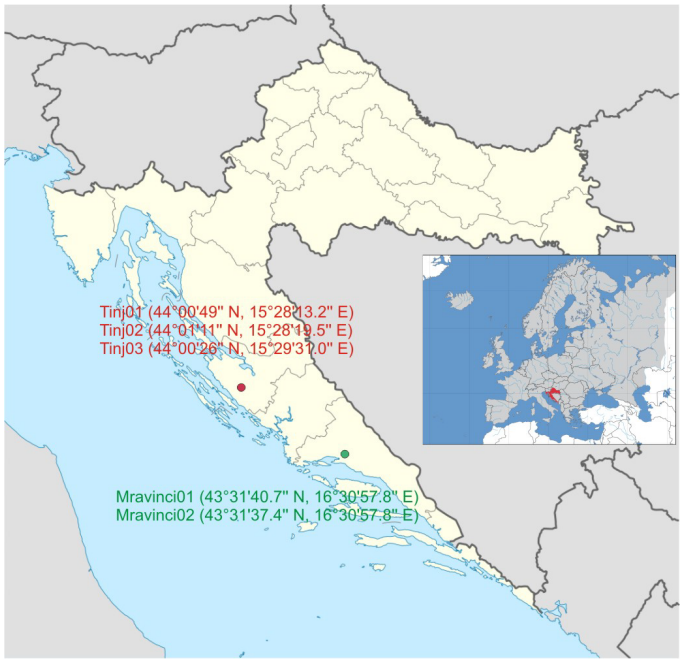


Figure 2: Locations of surveyed olive groves (Dalmatia region, Croatia). Three olive orchards in Tinj and two in Mravinci.

Also, one of the orchards (Tinj02) has been mapped in two different seasons of year (May and December) in order to analyze differences in detection performance not only for different orchards but also for the same orchard surveilled at different seasons and times of day.

Both observed regions have a Mediterranean climate characterized by dry summers and mild, wet winters. The UAV flights were performed on five dates: 10 May 2021 (Mravinci01), 11 May 2021 (Tinj01 and Tinj02), 20 December 2021 (Tinj02), 18 January 2022 (Mravinci02) and 5 April 2022 (Tinj03).

2.2. UAV for Images Acquisition

Nine image datasets were acquired using high resolution sensors onboard UAV platform to monitor the olive groves. RGB and multispectral images were collected using the camera on DJI Phantom 4 Multispectral drone. DJI Phantom 4 Multispectral drone is equipped with camera with six 1/2.9" CMOS (Complementary metal–oxide–semiconductor) image sensors. One CMOS sensor is RGB sensor for visible light imaging while other five sensors are used for multispectral imaging. Each sensor has 2.08 megapixels (MP).

For this research, we used only information from RGB sensor. In order to collect images needed for making the map of an olive orchard, UAV was programmed to fly at 35 m above ground altitude (AGL) with airspeed of 5 m/s. The forward and sideway image overlaps were 75%. Ground sampling distance (GSD) was 2 cm.

2.3. Software Tools

DJI Terra (https://www.dji.com/hr/dji-terra) was used as a flight planner software. Also, this software was used for stitching of the collected multispectral and RGB images and production of 2D terrain maps of monitored olive orchards (Figure 3). Since maps generation is a compute-intensive process, minimum hardware configuration for map reconstruction using DJI Terra is 16GB RAM and a NVIDIA graphics card with at least 4GB VRAM. For this purpose, one NVIDIA GForce RTX 3060 GPU with 12 Gb VRAM was used.

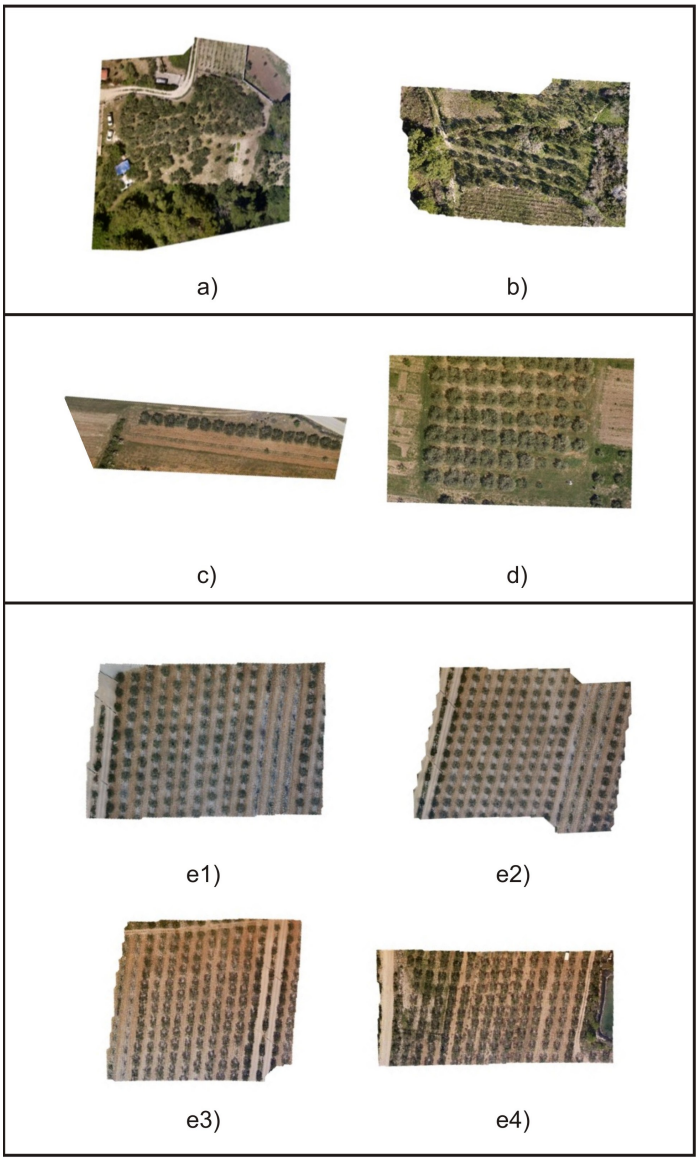


Figure 3: Maps of olive groves. a) Mravinci01, b) Mravinci02, c) Tinj01, d) Tinj02, e1) Tinj03 – Flight01, e2) Tinj03 – Flight02, e3) Tinj03 – Flight03, e4) Tinj03 – Flight04.

Labeling of individual olive trees on generated olive orchard maps was done using Computer Vision Annotation Tool (https://www.cvat.ai/). It is a free (for individual data scientists and small teams) web-based image and video annotation tool used for labeling data for computer vision algorithms. Labeled annotations for object detections was done in Pascal VOC format [39]. Each label is defined with four values (x_{min}, y_{min}, x_{max}, y_{max}): where x_{min} and y_{min} are coordinates of the upper left corner of the rectangle label and x_{max} and y_{max} are coordinates of the lower right corner of the rectangle label.

Implementation and evaluation of object (olive tree) detectors and image processing was done using Python 3.10.4. programming language with the DeepForest package that comes with the prebuilt RetinaNet model from the torchvision package. Proposed implementation has been done on Windows operating system although the package has been tested also on MacOS, and Linux.

2.4. RetinaNet Detector

Popular object detection models can be broadly classified into two categories: two-stage and single-stage detectors. Two-stage detectors are using one model to extract regions of objects (first stage), and a second model is used to classify and further refine the localization of the object (second stage). Single-stage detectors have only one model which skip the region proposal stage of two-stage models and run detection directly over a dense sampling of locations. Comparing to two-stage detectors, these types of models usually have faster inference (possibly at the cost of performance) [40]. RetinaNet is a single-stage detector which is fast and has accuracy comparable to two-stage detectors. RetinaNet uses a feature pyramid network (FPN) [41] which enables the detection of objects at multiple scales and introduces a new loss, the Focal loss function [35], to alleviate the problem of the extreme foreground-background class imbalance. Focal Loss function approach addresses this problem that occurs in single-stage detectors by assigning less weight to easily classified examples and focusing on correcting misclassified ones. RetinaNet's network architecture FPN backbone is on top of a feedforward ResNet architecture [42] with the goal of generating rich, multi-scale convolutional feature pyramid. RetinaNet attaches two subnetworks to this backbone, one for classifying anchor boxes and one for regressing from anchor boxes to ground-truth object boxes (Figure 4).

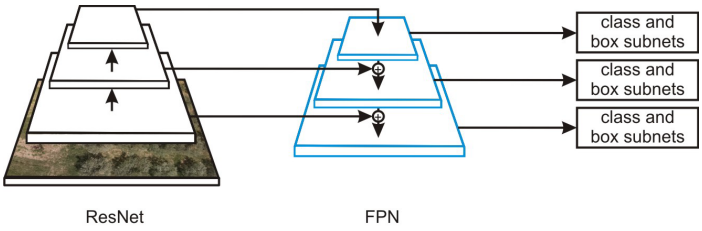


Figure 4: ResNet network architecture.

3. Procedure

After collecting sets of images for five olive orchards (parameters described in Section 2.2.), maps of olive groves for each flight were obtained with DJI Terra software. Each map was annotated using CVAT i.e. individual olive trees were labeled as a ground truth. However, generated maps have higher resolution than the images used for training the prebuilt RetinaNet model from the DeepForest package. Furthermore, their resolution may, generally, vary depending on the flight parameters and used sensor.

In order to get better predictions, it is necessary to divide each map into smaller windows that are more similar to the data on which the DeepForest model was trained. When forecasting, the input map is divided into smaller overlapping windows and then the model in each window tries to detect trees. Detections from all windows are then collapsed into detections (predictions) on the entire map, while redundant filtering is carried out frame by the non-max suppression method. This method keeps only the highest reliability frame from all detections whose predicted limit frames match more than the default intersection over union (IoU) threshold (Figure 5).

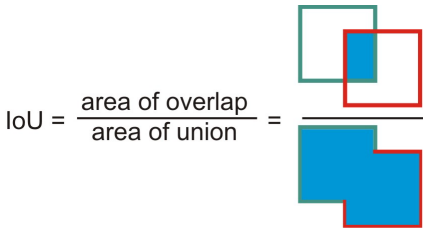


Figure 5: IoU illustration.

First detection results were obtained using the pretrained DeepForest model. Different window sizes (ranging from 600 x 600 to 1000 x 1000 pixels with a step of 50) with different "overlaps" (10-40% with a step of 5%) were tested and evaluated on maps Tinj1 and Tinj2 (Figure 3). During inference, the model tries to detect olive trees on each window, and afterwards, detections from all windows are compressed into detections on the whole map.

Finally, the windows size of 750 x 750 pixels was chosen with an overlap of 20%. For the pretrained model, the best predictions were obtained using confidence limit (τ) of 0.3.

Table 1: Models trained, N – total number of trees in maps (ground truth objects)

Model	Trained on maps	N
M1	Tinj02	69
M2	Tinj03 – Flight01	133
M3	Tinj02, Tinj03 – Flight01	202
M4	Tinj02, Tinj03 – Flight01, Tinj03 – Flight04	356
M5	Tinj03 – Flight01, Tinj03 – Flight04	287
M6	Tinj02, Tinj03 – Flight01, Mravinci02	277
M7	Tinj02, Tinj03 – Flight01, Tinj03 – Flight04, Mravinci02	431
M8	Tinj01, Tinj03 – Flight01, Tinj03 – Flight04, Mravinci02	381

Since the prebuilt model had been trained on various types of trees and not olives, further steps were needed in order to improve predictions and reduce the number of other trees being detected as olive trees. Therefore, an adaptation of the pretrained RetinaNet model to the local data using transfer learning was done. During this step, eight new models were trained. For each model, different labeled maps were used for training (Table 1).

Again, for new models, various windows sizes with different "overlaps" were tested. After tests, an image size of 1000 × 1000 pixels with 40% overlap was chosen.

Each network was trained for five epochs with stochastic gradient descent with a momentum of 0.9, a learning rate equal to 0.001 and a confidence threshold of 0.7. All trained models used a confidence threshold of 0.5 at inference time. As already written, a confidence threshold of 0.3 was chosen for the pretrained model because, in this case, all predicted bounding boxes had low confidence scores.

4. Results

4.1. Performance Evaluation

In order to evaluate the proposed methodology, exact number of olive trees in the evaluation areas was determined by a human observer (ground truth). The performance assessment of the methodology was approached by comparing the actual number of plants, and their distribution with the results of detection of a deep neural network for eight created models (and pretrained model).

A number of metrics defined below are proposed for quantitative assessment.

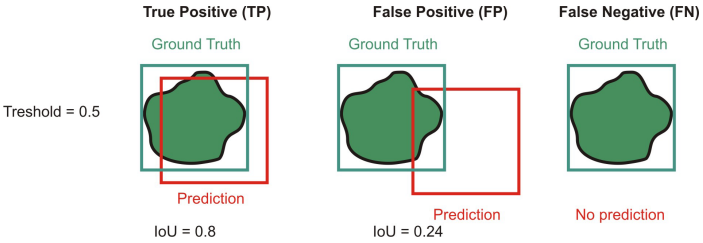


Figure 6: Examples of TP, FP and FN bounding boxes when default IoU = 0.5.

Using the calculated IoU value each predicted bounding box is classified into one of the following categories (illustrated in Figure 6):

- True Positive (TP): detection is correct (predicted frame matches with correct) if valid IoU >= threshold,
- False Positive (FP): the detection is wrong (a frame is provided for the object which is not in the picture, or the intended frame does not match the correct one) if IoU < threshold,
- False Negative (FN): the object in the image is not detected.

Precision: presents the hit ratio for the trees found by the algorithm.

$$Precision = \frac{TP}{TP+FP} = \frac{TP}{all\ detections} \quad (1)$$

where TP (true positives) is the number of olive trees correctly identified by the algorithm, and FP (false positives) is the number of instances wrongly proposed by the algorithm as potential olive trees.

Recall: presents the proportion of the trees correctly found by the algorithm.

$$Recall = \frac{TP}{TP+FN} = \frac{TP}{all\ ground\ truth\ boxes} \quad (2)$$

where FN (false negatives) is the number of olive trees that were not identified.

F1 score: the harmonic mean of precision and recall,

$$F1\ score = \frac{1}{\frac{1}{Precision}+\frac{1}{Recall}} = 2x\frac{Precision\ x\ Recall}{Precision+Recall} \quad (3)$$

For the calculation of previous metrics, IoU value of 0.5 was used (the most commonly used values for IoU are 0.5 and 0.75 [43]).

Average precision (AP): summarizes the precision recall curve into one number, it can be interpreted as the area under the precision-recall (PR) curve.

$$AP = \int_0^1 p(r)dr \quad (4)$$

where p(r) is Precision for particular Recall value.

Precision – Recall (PR) curve is obtained by plotting points (r(τ), p(τ)) where r(τ) and p(τ) denote precision and recall at confidence threshold τ. In practice, the area is calculated under interpolated monotone curve instead of the actual "zig-zag" PR curve. Average precision is calculated according to MS COCO [44]. For AP calculation, 101 recall points on the PR curve are used (0 to 1 with a step size of 0.01). More precisely, AP@0.5 calculated with a fixed IoU threshold of 0.5 is used, while AP is obtained by averaging AP@α for IoU thresholds α from 0.5 to 0.95 with a step size of 0.05.

4.2. Evaluation Results

In this section, the evaluation results for the trained olive tree detectors will be presented. Tables 2 and 3 show evaluation results of all 9 models on maps Tinj03 - Flight02 and Tinj03 - Flight03. As it can be seen from Table 1, these maps weren't used for any model's training.

Combined results for both maps (Tinj03 - Flight02 and Tinj03 - Flight03) are presented in Table 4.

This maps present parts of a large orchard (each map corresponds to one drone flight). Since the implementation of computer-based tree counting and labeling is particularly interesting for large orchards (small orchards are economically less significant), results for that type of orchards are the focus of our interest and basis for a future applications.

As expected, model (M5) for which training phase used only other parts (maps) of the same orchard (Tinj03), showed slightly better results than others but differences were not significant, moreover, two models have higher precision (M2 and M8) and one (M8) has higher AP. Recall and F1 measure, as can be seen, of the pretrained model lags significantly behind models trained on images of olive groves obtained by drone.

Table 2: Results for map Tinj03 – Flight02

Model	Precision	Recall	F1	AP@0.5	AP
pretrained	0.1212	0.0258	0.0426	0.0069	0.0044
M1	0.3333	0.1935	0.2449	0.1034	0.0354
M2	0.9434	0.9677	0.9554	0.9585	0.6484
M3	0.9212	0.9806	0.9500	**0.9754**	**0.6652**
M4	0.9375	0.9677	0.9524	0.9567	0.6037
M5	**0.9500**	**0.9806**	**0.9651**	0.9750	0.6441
M6	0.9379	0.9742	0.9557	0.9632	0.6408
M7	0.9157	0.9806	0.9470	0.9645	0.6367
M8	0.9487	0.9548	0.9518	0.9479	0.6512

Table 3: Results for map Tinj03 – Flight03

Model	Precision	Recall	F1	AP@0.5	AP
pretrained	0.2667	0.0500	0.0842	0.0194	0.0092
M1	0.4545	0.2188	0.2954	0.1565	0.0352
M2	**1.0000**	0.8875	0.9404	0.8812	0.5256
M3	0.9605	0.9125	0.9359	0.9094	0.5650
M4	0.9813	**0.9813**	**0.9813**	0.9787	0.6016
M5	0.9691	**0.9813**	0.9752	**0.9799**	**0.6514**
M6	0.9419	0.9125	0.9270	0.8968	0.5034

M7	0.9398	0.9750	0.9571	0.9604	0.5981
M8	0.9810	0.9688	0.9748	0.9601	0.6474

Table 4: Mean metrics value on maps Tinj03 – Flight02 and Tinj03 – Flight03

Model	Precision	Recall	F1	AP@0.5	AP
pretrained	0.1939	0.0379	0.0634	0.0131	0.0068
M1	0.3939	0.2061	0.2701	0.1300	0.0353
M2	**0.9717**	0.9276	0.9479	0.9199	0.5870
M3	0.9409	0.9466	0.9429	0.9424	0.6151
M4	0.9594	0.9745	0.9668	0.9677	0.6026
M5	0.9596	**0.9809**	**0.9701**	**0.9774**	0.6478
M6	0.9399	0.9433	0.9413	0.9300	0.5721
M7	0.9277	0.9778	0.9520	0.9625	0.6174
M8	0.9649	0.9618	0.9633	0.9540	**0.6493**

According to the results from table 4, it can be commented that models M5 and M8 stand out as the best possible options in this case.

Since majority of models used some maps from Tinj03 orchard (7 out of 9), more objective detection results may be the ones obtained for Tinj02 orchard (5 out of 9). Results for Tinj02 are presented in Table 5.

Again, the best model (M4) in this case, used map of the orchard in the training phase. The best model for previous case (M5) showed rather low precision result (0.7941) while one of the best models for detections on Tinj03 maps – M8 showed rather high precision (0.9552) and recall (0.9275) values on this map, also. Both models, M5 and M8 have not used Tinj02 maps in training phase. Detections obtained by applying these two models to two evaluation maps and the Tinj02 map (neither of these models used Tinj02 map for the training) compared to detections of the pretrained model are presented in Figure 7.

Table 5: Results for map Tinj02 (flight from May 2021).

Model	Precision	Recall	F1	AP@0.5	AP
pretrained	0.6585	0.7826	0.7152	0.669	0.2795
M1	0.9296	0.9565	0.9429	0.9465	0.5526
M2	0.873	0.7971	0.8333	0.7472	0.2496
M3	0.8481	**0.971**	0.9054	**0.9669**	0.5378
M4	**0.9571**	**0.971**	**0.964**	0.9661	**0.5688**
M5	0.7941	0.7826	0.7883	0.7358	0.1567
M6	0.9286	0.942	0.9353	0.9374	0.5684
M7	0.9167	0.9565	0.9362	0.9291	0.512
M8	0.9552	0.9275	0.9412	0.9146	0.4582

Corresponding to this, Table 6 presents the exact number of detected olive trees, ground-truth labels, TP, FP, and FN. It can be noted that, on map Tinj03 - Flight 02 (Figure 7), in both models (M5 and M8), some of the fig trees planted between olives were mistaken for olive trees.

Next evaluation was done on the detection results for an extensive orchard that has not been used for training of any model (Mravinci01 – Figure 3.a.). Mravinci01 can be seen as a special case of the olive grove because it is characterized by irregular pruning, overlapping tree crowns, and the non-uniform shaping of trees. Moreover, in this aerial image, various types of trees are also present, apart from olives. In the case of the map Mravinci01, even human annotators have a hard time labeling olive trees. Precision and recall values are significantly lower than in previous cases.

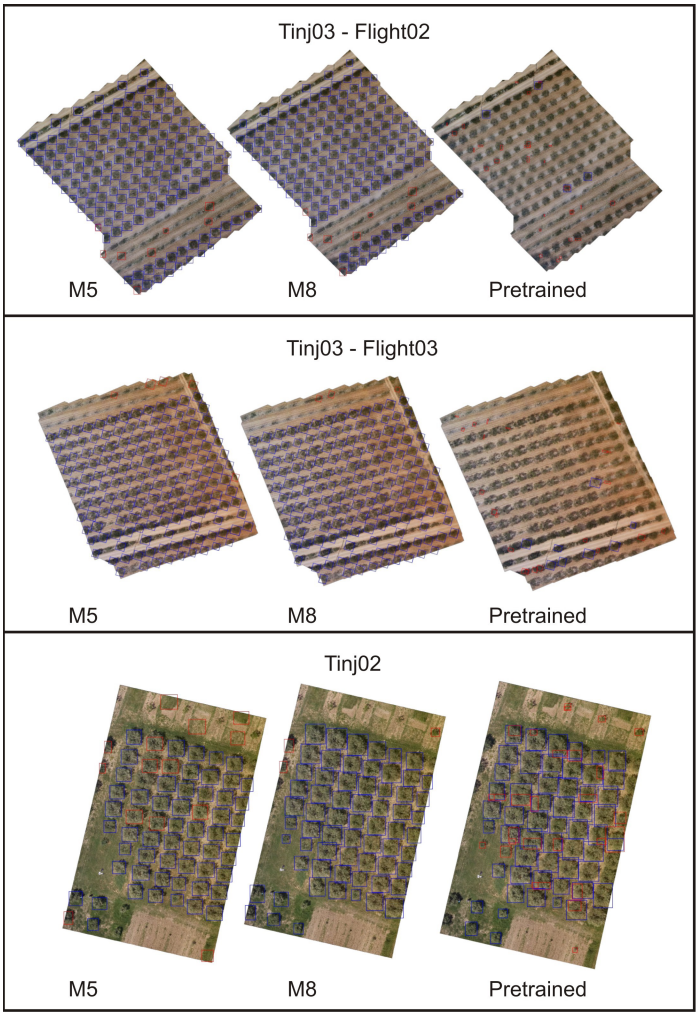


Figure 7: Detections made by model M5, model M8 and pretrained models. True positive (TP) detections are shown in blue, false positive (FP) detections are shown in red.

Table 6: Model M5, model M8 and pretrained model detections on maps Tinj03 – Flight02, Tinj03 – Flight03 and Tinj02

map	true boxes	model	detections	TP	FP	FN
Tinj03 – Flight02	155	M5	160	152	8	3
		M8	156	148	8	7
		pretrained	33	4	29	151
Tinj03 – Flight03	160	M5	162	157	5	3
		M8	158	155	3	5
		pretrained	30	8	22	152
Tinj02	69	M5	68	54	14	15
		M8	67	64	3	5
		pretrained	82	54	28	15

Best precision was achieved with model M6 (0.5068) and highest recall was achieved with model M3 (0.6138). This is expected due to aforementioned reasons, as well as lack of proper training examples for model generation (only Mravinci02 map can be considered as extensive orchard but with larger distances between trees). However, even here, significant improvement of

generated models over pretrained model can be confirmed - precision for pretrained model was 0.1205 and recall 0.0516. The pretrained model produced drastically more false positives, especially on the part of the map with the pine trees (Figure 8).

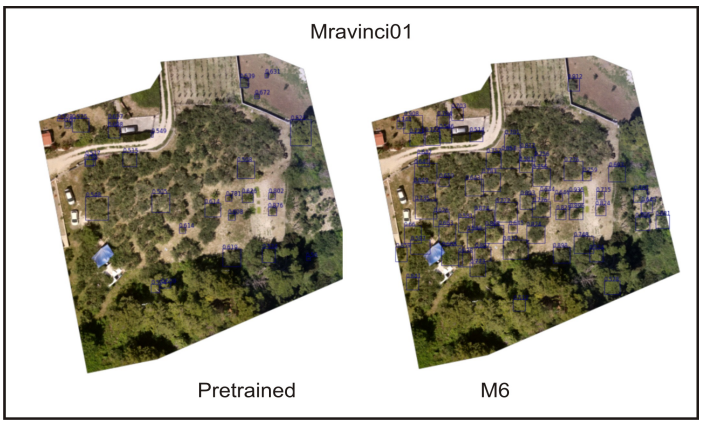


Figure 8: Special case: M6 vs Pretrained model comparison for olive grove Mravinci01.

Finally, we present analysis results of differences in detection performance for the same orchard surveilled at different seasons and times of day. One of the test orchards (Tinj02) has been mapped in May (and used for generating some of models presented in Table 1) and later in December of the same year (2021). Detection results for second flight (map) are given in Table 7.

Table 7: Results for map Tinj02 (flight from December 2021).

Model	Precision	Recall	F1	AP@0.5	AP
pretrained	0.6706	0.7215	0.6951	0.5425	0.1995
M1	0.8462	0.8354	0.8408	0.8217	0.367
M2	0.75	0.6456	0.6939	0.4926	0.1676
M3	0.8222	**0.9367**	0.8757	0.9039	0.4976
M4	0.9605	0.9241	0.9419	0.92	0.5043
M5	0.9016	0.6962	0.7857	0.6863	0.2983
M6	**0.9737**	**0.9367**	**0.9548**	**0.9302**	**0.5083**
M7	0.9012	0.9241	0.9125	0.9124	0.4457
M8	0.9733	0.9241	0.9481	0.9196	0.4576

As opposite to results presented in Table 5, for the second flight, the best performance has been achieved with model M6. Interesting, performance of some models was better for December flight than for May flight (M6, M8). Precision and recall comparison for all models is presented in Figures 9 and 10.

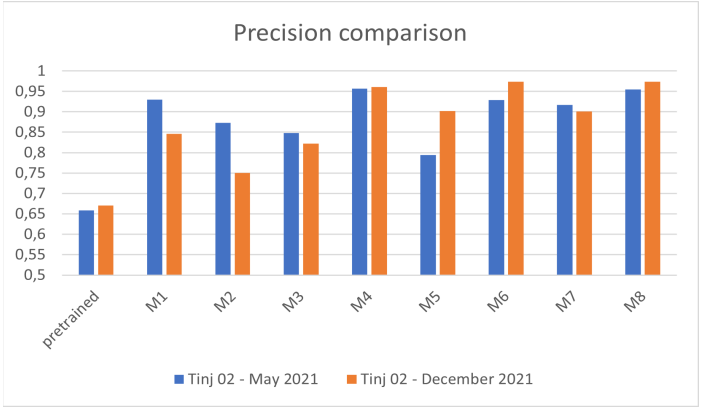


Figure 9: Precision comparison for Tinj02 maps.

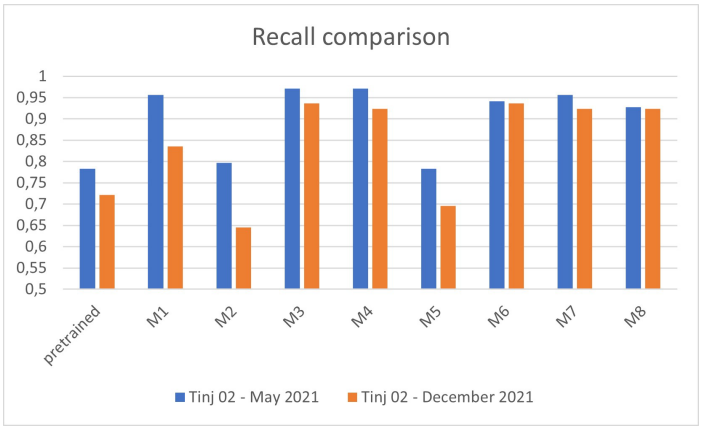


Figure 10: Recall comparison for Tinj02 maps.

When compared to variations (absolute differences) in results for all models between two similar maps taken the same day in the same large orchard, absolute differences were similar. For instance, average absolute difference for precision results between maps generated for Tinj02 flights is 0.048 while the average absolute difference for precision results between maps generated for Tinj03-Flight02 and Tinj03-Flight03 is 0.054. This implies that detection results for the same olive orchard are not to be significantly degraded during period of several months. However, this should be confirmed on a larger number of test cases.

5. Discussion and conclusion

Automatic olive tree detection is a task with many challenges. Acquired images of the olive groves can vary significantly due to different types of soil and vegetation in orchards, changes in vegetation during seasons, the age of the orchard and tree sizes, irregular pruning and pruning types, the non-uniform shaping of trees, changes in weather conditions and illumination. Also, trees in the orchard can be planted very closely, forming joint crowns, and making it very difficult, even for human annotator, to label individual tree. In this paper, we presented a procedure for development of a deep learning object detector for detecting individual olive trees from aerial RGB images by fine-tuning the prebuilt RetinaNet model on local data.

During development of the olive trees object detector, several models were trained using a different training set of images - different olive grove maps. Maps of five diverse olive groves (small and large) were generated but the focus was on automatization of monitoring a large olive grove such as Tinj03. Comparison of model performance for the olive tree detection in different times of the year was presented. As it can be seen from the obtained results, there was generally no degradation in detection. For some particular tests such as evaluation of detections from diverse parts of the orchard in Tinj03, the best performing models were the model M5 which uses only other parts of the same orchard as the training data, and model M8, which expands that data with images from two other olive groves (Tinj01 and Mravinci02). As already said, even though the olive groves such as Tinj03 will be the focus of future research, a trained detector should be generalizable to orchards with diverse vegetations and various-sized olive trees. In this context, for further use, we propose the model M4 and, alternatively, M8. Although there is no clear winner between tested models, perhaps M4 could be

considered as the most reliable. Proposed model (M4) has been trained on 356 ground truth olive trees while the runner-up (M8) has been trained on 381 ground truth olive trees which classifies them in top 3 models according to the number of trees used for training. This indicates that further improvements can be expected with additional training examples.

Comparing to the prebuilt model that showed very poor performance in olive trees detection, experimental results have shown the dominance in performance of fine-tuned models. Achieved precision and recall even with the relatively small training dataset (generally > 95% for heavy pruned orchards) makes this approach useful for the implementation.

Findings related to the fairly stable detection results of the same olive orchard taken several months apart are certainly interesting because, to the best of our knowledge, there has been no such analysis in the literature so far.

There are several directions for future research. The most imminent one should be utilizing the olive-tree detector for olive groves analysis, such as crop yield estimation and monitoring plant health and fruit status using vegetation indices. Automated collection of images of individual olive trees will greatly speed up and facilitate the analysis process. Image processing procedures aimed at obtaining information about plant health (e.g. from NDVI index values) or plant water status (e.g. from thermal camera data) will be able to be automated, because in this way only the part of the image related to an individual tree (output from the detector) will be brought to the input.

Moreover, as noted, there is still room for further improvement of the obtained detection model(s) by expanding the existing training datasets with more aerial images of olive trees from different localities, types of pruning and in different seasons to obtain more robust olive detectors.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] National Research Council, Division on Earth and Life Studies, Board on Earth Sciences and Resources, Committee on Strategic Directions for the Geographical Sciences in the Next Decade, Understanding the Changing Planet: Strategic Directions for the Geographical Sciences, National Academies Press, 2010.

[2] C.I. Gan, R. Soukoutou, D.M. Conroy, Sustainability Framing of Controlled Environment Agriculture and Consumer Perceptions: A Review. Sustainability 2023, **15**(1), 304. https://doi.org/10.3390/su15010304.

[3] M. Dijk, T. Morley, M.L. Rau, S. Yashar, A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050, Nature Food, **2**, 2021, 494–501, https://doi.org/10.1038/s43016-021-00322-9.

[4] M. Elferink, F. Schierhorn, Global Demand for Food is Rising. Harvard Business Review April 07, 2016.

[5] D. Niklis, G. Baourakis, B. Thabet, G. Manthoulis, "Trade and logistics: the case of the olive oil sector," in MediTERRA 2014. Presses de Sciences Po, 203 – 226, 2014, doi : 10.3917/scpo.cihea.2014.02.0203.

[6] F. B. Insights, "Olive oil market size, share & covid-19 impact analysis, by type (refined olive oil, virgin olive oil, olive pomace oil, and others), end-user (household/retail, food service/horeca, food manufacturing, and others), and regional forecast, 2020–2027", 2021.

[7] S. Mili, M. Bouhaddane, "Forecasting Global Developments and Challenges in Olive Oil Supply and Demand: A Delphi Survey from Spain". Agriculture, 2021, **11**(3), 191. https://doi.org/10.3390/agriculture11030191.

[8] A. Kamilaris, A., Gao, F., Prenafeta-Boldú, F.X., Ali, M.I., "Agri-IoT: A Semantic Framework for Internet of Things-Enabled Smart Farming Applications". 3rd World Forum on Internet of Things (WF-IoT) IEEE, Reston, VA, USA, 442–447, 2016, doi: 10.1109/WF-IoT.2016.7845467.

[9] W. Bastiaanssen, D. Molden, I. Makin, "Remote sensing for irrigated agriculture: examples from research and possible applications". Agric. Water Manag. **46** (2), 137–155, 2000, doi: 10.1016/S0378-3774(00)00080-9.

[10] P. Nevavuori, N. Narra, T. Lipping, "Crop yield prediction with deep convolutional neural networks", Computers and Electronics in Agriculture, vol. **163**, 2019, https://doi.org/10.1016/j.compag.2019.104859.

[11] A. Matese, P. Toscano, S. F. Di Gennaro, L. Genesio, F. P. Vaccari, J. Primicerio, C. Belli, A. Zaldei, R. Bianconi, B. Gioli, "Intercomparison of UAV, aircraft and satellite remote sensing platforms for precision viticulture". Remote Sensing, **7**(3):2971-2990, 2015, https://doi.org/10.3390/rs70302971.

[12] M. Waleed, T. -W. Um, A. Khan and Z. Ahmad, "An Automated Method for Detection and Enumeration of Olive Trees Through Remote Sensing," in IEEE Access, vol. **8**, 108592-108601, 2020, doi: 10.1109/ACCESS.2020.2999078.

[13] J. M. Ponce, A. Aquino, B. Millan, J. M. Andújar, "Automatic Counting and Individual Size and Mass Estimation of Olive-Fruits Through Computer Vision Techniques," in IEEE Access, vol. **7**, pp. 59451-59465, 2019, doi: 10.1109/ACCESS.2019.2915169.

[14] S. Benalia, B. Bernardi, J. Blasco, A. Fazari, G. Zimbalatti, "Assessment of the Ripening of Olives Using Computer Vision", Chemical Engineering Transactions. **58**, 355-360, 2017, https://doi.org/10.3303/CET1758060.

[15] Petteri Nevavuori, Nathaniel Narra, Tarmo Lipping, "Crop yield prediction with deep convolutional neural networks", Computers and Electronics in Agriculture, **163**, 2019, https://doi.org/10.1016/j.compag.2019.104859.

[16] X. Liu, W. Min, S. Mei, L. Wang, S. Jiang, "Plant Disease Recognition: A Large-Scale Benchmark Dataset and a Visual Region and Loss Reweighting Approach", IEEE Transactions on Image Processing, **30**, 2003-2015, 2021, doi: 10.1109/TIP.2021.3049334.

[17] N. T. Waskitho, "Unmanned aerial vehicle technology in irrigation monitoring", Advances in Environmental Biology, **9**(23), 7–10, 2015.

[18] C. Albornoz, L. F. Giraldo, "Trajectory design for efficient crop irrigation with a UAV," 2017 IEEE 3rd Colombian Conference on Automatic Control (CCAC), 2017, 1-6, doi: 10.1109/CCAC.2017.8276401.

[19] Z. Zhen, L. J. Quackenbush, and L. Zhang, "Trends in automatic individual tree crown detection and delineation—evolution of lidar data," Remote Sensing, **8**(4): 333. https://doi.org/10.3390/rs8040333.

[20] M. Dalponte, H. O. Orka, L. T. Ene, T. Gobakken, and E. Nasset, "Tree crown delineation and tree species classification in boreal forests using hyperspectral and als data", Remote Sensing of Environment, **140**, 306 – 317, 2014. https:// doi.org/10.1016/j.rse.2013.09.006.

[21] G. Avola, S.F. Di Gennaro, C. Cantini, E. Riggi, F. Muratore, C. Tornambè, and A. Matese, "Remotely Sensed Vegetation Indices to Discriminate Field-Grown Olive Cultivars", Remote Sensing, **11**, 1242, 2019, https://doi.org/10.3390/rs11101242.

[22] I. N. Daliakopoulos, E. G. Grillakis, A. G. Koutroulis, I. K. Tsanis, "Tree crown detection on multispectral vhr satellite imagery", Photogrammetric Engineering & Remote Sensing, **75**(10), 1201 – 1211, 2009, DOI: 10.14358/PERS.75.10.1201.

[23] J. Peters, F. Van Coillie, T. Westra, R. De Wulf, "Synergy of very high resolution optical and radar data for object-based olive grove mapping", International Journal of Geographical Information Science, **25**(6), 971 – 989, 2011, https://doi.org/10.1080/13658816.2010.515946.

[24] R. Sarabia, A. Aquino, J. M. Ponce, G. Lopez, J. M. Andújar, "Automated identification of crop tree crowns from uav multispectral imagery by means of morphological image analysis", Remote Sensing, **12**(5), 748, 2020, https://doi.org/10.3390/rs12050748.

[25] L. Saxena, L. Armstrong, "A survey of image processing techniques for agriculture". Proceedings of Asian Federation for Information Technology in Agriculture, Australian Society of Information and Communication Technologies in Agriculture. Perth, Australia, 401-413, 2014.

[26] E. Hamuda, M. Glavin, E. Jones, "A survey of image processing techniques for plant extraction and segmentation in the field", Computers and Electronics in Agriculture, **125**, 184–199, 2016, doi:10.1016/j.compag.2016.04.024.

[27] A. Singh, B. Ganapathysubramanian, A.K. Singh, S. Sarkar, "Machine learning for high-throughput stress phenotyping in plants", Trends Plant Sci. **21** (2), 110–124, 2016, DOI:https://doi.org/10.1016/j.tplants.2015.10.015.

[28] A. Khan, U. Khan, M. Waleed, A. Khan, T. Kamal, S. N. K. Marwat, M.

Maqsood, F. Aadil, "Remote sensing: An automated methodology for olive tree detection and counting in satellite images," IEEE Access, **6**, 77 816–77 828, 2018, doi: 10.1109/ACCESS.2018.2884199.

[29] M. Waleed, T.-W. Um, A. Khan, and U. Khan, "Automatic detection system of olive trees using improved k-means algorithm", Remote Sensing, **12**(5), 2020, https://doi.org/10.3390/rs12050760.

[30] A. Kamilaris, F.X. Prenafeta-Boldú, "Deep learning in agriculture: a survey", Computers and Electronics in Agriculture, **147**, 70–90, 2018, https://doi.org/10.1016/j.compag.2018.02.016.

[31] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, T., C. McCool, "Deepfruits: a fruit detection system using deep neural networks". Sensors, **16**(8) , 2016, https://doi.org/10.3390/s16081222.

[32] M. Onishi, T. Ise, "Explainable identification and mapping of trees using UAV RGB image and deep learning", Scientific reports, **11**(1), 903, 2021, https://doi.org/10.1038/s41598-020-79653-9.

[33] A. Safonova, E. Guirado, Y. Maglinets, D. Alcaraz-Segura, S. Tabik, "Olive tree biovolume from uav multi-resolution image segmentation with mask r-cnn", Sensors, **21**(5), 2021, https://doi.org/10.3390/s21051617.

[34] B. G. Weinstein, S. Marconi, M. Aubry-Kientz, G. Vincent, H. Senyondo, E. P. White, "Deepforest: A python package for RGB deep learning tree crown delineation", Methods in Ecology and Evolution, **11**(12), 1743 – 1751, 2020, https://doi.org/10.1111/2041-210X.13472.

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, "Focal loss for dense object detection", 2017 IEEE International Conference on Computer Vision (ICCV), 2999 – 3007, 2017, doi: 10.1109/ICCV.2017.324.

[36] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, **42**(2), 318-327, 1 Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[37] S. Marcel, Y. Rodriguez, „Torchvision the machine-vision package of torch", Proceedings of the 18th International Conference on Multimedea 2010, Firenze, Italy, October 25-29, 2010, DOI: 10.1145/1873951.1874254.

[38] I. Marin, S. Gotovac, V. Papić, "Individual Olive Tree Detection in RGB Images," 2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), 2022, 1-6, doi: 10.23919/SoftCOM55329.2022.9911397.

[39] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge", International Journal of Computer Vision, **88** (2), 303-338 2010, https://doi.org/10.1007/s11263-009-0275-4.

[40] C. Coelho, M. F. P. Costa, L. L. Ferras, A. J. Soares, "Object detection with retinanet on aerial imagery: The algarve landscape", International Conference on Computational Science and Its Applications. Springer, 2021, 501 – 516, DOI: 10.1007/978-3-030-86960-1_35.

[41] T.-Y. Lin, P. Doll´ar, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature pyramid networks for object detection", 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 936-944, 2017, DOI: 10.1109/CVPR.2017.106.

[42] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, 770-778, doi: 10.1109/CVPR.2016.90.

[43] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, E. A. B. da Silva, "A comparative analysis of object detection metrics with a companion open-source toolkit", Electronics,, 2021, **10**(3), 279, https://doi.org/10.3390/electronics10030279.

[44] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollar, „Microsoft COCO: Common objects in context,", Computer Vision – ECCV 2014. ECCV 2014, Lecture Notes in Computer Science, **8693**. Springer, Cham, https://doi.org/10.1007/978-3-319-10602-1_48.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 89-99 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

HistoChain: Improving Consortium Blockchain Scalability using Historical Blockchains

Marcos Felipe, Haiping Xu*

Computer and Information Science Department, University of Massachusetts Dartmouth, Dartmouth, 02747, USA

ARTICLE INFO

Article history:
Received: 21 February, 2023
Accepted: 10 May, 2023
Online: 21 May, 2023

Keywords:
Consortium blockchain
Historical blockchain
On-chain big data
Scalable storage
Dynamic load balancing
Healthcare data

ABSTRACT

Blockchain technology has been successfully applied in many fields for immutable and secure data storage. However, for applications with on-chain big data, blockchain scalability remains to be a main concern. In this paper, we propose a novel scalable storage scheme, called HistoChain, for a consortium blockchain network to manage blockchain data. We use a current blockchain and historical blockchains to store on-chain big data, where the current blockchain and the historical blockchains store data from recent years and earlier years, respectively. Both the current blockchain and the historical blockchains are maintained by super peers in the network; while regular peers manage only the current blockchain and can retrieve historical data by making queries to the super peers. We present procedures for generating historical blockchains, dynamically balancing the data retrieval workload of super peers, and concurrently retrieving historical blockchain data in response to queries. We further provide a case study of healthcare data storage using a consortium blockchain, and the simulation results show that our scalable HistoChain storage scheme supports efficient access and sharing of big data on the blockchain.

1. Introduction

In recent years, the use of blockchain technology in many fields has gained increasing interest and popularity [1]. As a distributed and decentralized ledger, blockchain technology allows for the protection of transactions and data while maintaining the data sharing and reliability of a peer-to-peer network [2]. Peers maintain "chains" of blocks consisting of various types of data stored as transactions. Each block contains the hash value of the previous block in the chain, so any attempt to modify one block has a ripple effect on all subsequent blocks in the blockchain. These altered hash values can be easily identified because peers in the network maintain copies of the chain and can independently verify transactions and blocks. Permissioned blockchains allow peers in the network with the required permissions to access recorded transactions, while the key benefits of security, immutability, integrity, and transparency are preserved for transaction records [3]. The reliability and ease of securing and accessing data may explain the growing prevalence of blockchain technology worldwide. Bitcoin, a digital currency that utilizes public blockchain technology, had over 100 million users in 2022. The Bitcoin blockchain grew by more than 400 gigabytes from January 2012 to July 2022, and has even doubled since February 2019. In the face of this incredible growth, the cost of becoming a full-fledged node in a blockchain network is daunting and could become completely impractical. Similar to public networks like Bitcoin and Ethereum, consortium networks also run into storage problems [4]. In general, applications that require big data storage pose such problems, even if these networks do not consist of many peers or transactions. A wide range of domains, such as healthcare, real estate, insurance, and the Internet of Things (IoT), have adopted blockchain technology, resulting in a variety of data types and applications. While these applications typically use consortium blockchain networks, data-rich applications inevitably face storage issues, which raise significant concern about blockchain scalability.

The concern for blockchain scalability is the main reason for many studies on consortium blockchain storage management [5], [6]. However, most of the proposed solutions employ various off-chain storage strategies such as InterPlanetary File System (IPFS) and cloud storage, where IPFS is a decentralized, secure, verifiable, distributed storage system that can be integrated with blockchain networks [7]. Although off-chain approaches can alleviate the scalability issues of blockchain storage, the benefits of using blockchain technology are lost as the data is moved off the chain and new issues regarding the security and maintainability of off-chain data can be introduced. In this paper, we propose an on-chain approach, called *HistoChain*, to reduce the storage burden on most peers in a blockchain network by

*Corresponding Author: Haiping Xu, University of Massachusetts Dartmouth, Dartmouth, MA 02747, Email: hxu@umassd.edu

splitting the current blockchain (*CB*) and transferring the old data to a historical blockchain (*HB*), thereby reducing the size of the *CB* by half. In the *HistoChain* approach, *HB*s are immutable blockchains containing historical data separate from the *CB*, while the *CB* contains only the most recent years of blockchain data. After a set period of time, the *CB* will have grown further, and it will then be split again, generating another *HB*. In our approach, the nodes in the network are set up as either super peers or regular peers, with a smaller but substantial number of nodes forming a group of super peers, each of which maintains a copy of the *CB* and all *HB*s. Regular peers, which comprise most of the nodes in the network, need only retain the *CB*. This greatly reduces the storage burden on regular peers, which can then access data from the historical blockchains by making queries to the super peer group. In our approach, we use a time-based partitioning method to split the *CB* when it reaches a certain age. For example, if this age is 10 years, an *HB* will be created containing the first 5 years of data, leaving only the most recent 5 years of data in the *CB*. This splitting process can continue over time, resulting in the creation of multiple *HB*s.

Since regular peers are not required to store *HB*s, making query requests to the super peer group is their means of accessing historical data from the blockchain. When a super peer receives a request to search for historical data, it retrieves the requested data from the historical blockchains, and sends a summary report containing all retrieved information back to the requesting regular peer. In our approach, we introduce a meta-block, a mutable block attached to the beginning of the *CB* or each of the *HB*s, which contains index information for all transactions stored in the corresponding blockchain. This index information can facilitate fast and efficient data retrieval from a large blockchain that contains many years of data; therefore, the search time for historical data can be significantly reduced.

This work significantly extends the scalable storage scheme we previously proposed for on-chain big data using historical blockchains, originally presented at the IEEE International Workshop on Blockchain and Smart Contracts in 2022 (IEEE BSC 2022). In our previous work [8], we defined a primary super peer, called *PSP*, as an elected super peer who plays a role in efficiently facilitating access to data in *HB*s by regular peers. However, this approach introduces centralization and requires the necessary trust in a particular super peer (i.e., the *PSP*), which shall be best avoided in a blockchain architecture. In this paper, we allow a query to be sent to any super peer, which is responsible for collecting retrieved historical data and returning a summary report. To ensure temporal efficiency in query execution, query delegation will be performed within the super peer group. We design a dynamic load balancing algorithm to support fulfilling a request in a timely and concurrent manner. Each request for historical data sent to the super peer group is divided into subqueries with a search time of no more than 5 years, which are assigned to super peers based on their current workload. For this purpose, each super peer maintains a Shared Assignment Table (*SAT*) that keeps a record of assignments for all super peers and their completion times. Once an assignment is accepted by a super peer, an update to the *SAT* is broadcast within the super peer group to ensure that the super peers are aware of the latest status of the blockchain network.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents the *HistoChain* framework for scalable storage using historical blockchains and describes the procedure for generating historical blockchains. Section 4 describes in detail the dynamic load balancing algorithm and the retrieval process of historical blockchain data. Section 5 presents the case studies and their analysis results. Section 6 concludes the paper and mentions future work.

2. Related Work

Scalability challenges in blockchain technology, especially in public blockchain systems, remain a persistent issue. In [9], the authors introduced the Bitcoin Lightning Network (BLN), a decentralized system where transactions can be sent off-chain for value transfer through channels. The BLN, through its ability to make micro-payments, has positively impacted the scalability of the global Bitcoin blockchain network by reducing the need to broadcast many transactions. Danksharding is a newer type of sharding architecture proposed to scale the Ethereum network [10]. In the Danksharding proposal, nodes can validate larger data volumes through distributed data sampling across blob; therefore, nodes can avoid processing all data and larger data volumes can be handled by the Ethereum network. Scalability challenges also arise in consortium blockchain networks when large amounts of data need to be stored. In the context of consortium, off-chain strategies to improve the scalability of blockchain applications are the main focus of further research. To reduce the high cost of computation and storage for blockchain-based applications, in [11], the authors investigated a series of off-chain computation and storage approaches. They proposed five off-chain models that move computation and data off the blockchain without violating the trustless property. In [12], the authors proposed an off-chain scalability solution, called ChainSplitter, for Industrial Internet of Things (IIoT) blockchain applications. The proposed approach features a hierarchical storage structure where the recent blocks are stored in an overlay network and the majority of blockchain data is stored in the cloud. Despite being structured as a decentralized cloud storage system, the blockchain data in the cloud is not maintained by peers and thus acts as an off-chain repository for blockchain data. IPFS also offers a scalable off-chain solution for blockchains. In [13], the authors presented a blockchain-based application using IPFS specifically for healthcare systems. They focused on storage of electronic health records (EHRs) and used the IPFS service to transfer data off-chain while retaining hashes of the data on the blockchain. In [14], the authors attempted to reduce the transaction size and increase the transaction throughput of an experimental consortium blockchain network by storing the hash values of encrypted data on-chain and using IPFS to store the encrypted data itself off-chain. They integrated Hyperledger Fabric [15], which is a modular blockchain framework typically using off-chain storage for big data, with IPFS services and provided a solution for secure storage and efficient access to a task-scheduling scheme. While the off-chain approach provides a viable way to mitigate the scalability problem of blockchains, as noted in [11], the fundamental properties of blockchains can be compromised to varying degrees when using the off-chain approach. In contrast, our *HistoChain* approach stores big data in historical blockchains and does not rely on off-chain storage; therefore, all essential

properties of the blockchain data can be strictly maintained using our on-chain storage mechanism.

There are very few on-chain based approaches that address the scalability issues in blockchain networks. In [16], the authors proposed to use Hyperledger Fabric to implement a consortium blockchain for patient access and management of personal health records (PHRs). Although scalability issues remain a major challenge, they concluded that Hyperledger Fabric for on-chain data storage could offer a more practical solution to ensure the privacy of PHRs than the Ethereum public blockchain. In [17], the author introduced the concept of section-blockchain, an on-chain approach for reducing the storage cost of blockchain networks with under-stored devices. In their approach, all nodes store a portion of the complete blockchain and provide incentives for upgrading their local storage. Furthermore, they proposed segmented blockchains to enable nodes to store a blockchain segment [18]. They showed that their approach can help reduce the storage cost of a blockchain without compromising the security requirement of the blockchain. In [19], the authors proposed a framework for cloud-based blockchains to store medical multimedia files on-chain securely and reliably. They used a cloud-based blockchain to store all blockchain data to support data accessibility, redundancy, and security, while a lite blockchain allows local storage of text-based information and metadata for multimedia files. Although the above methods allow for big data storage, data retrieval can be slow because portions of transactions are stored in different blockchains. Conversely, our *HistoChain* approach divides a complete blockchain into a current blockchain and multiple historical blockchains, each of which are full-fledged blockchains containing complete transaction information. Regular peers can then access their local current blockchain and request historical blockchain data from super peers concurrently, making the data retrieval process much more efficient.

One of the main advantages of the *HistoChain* approach is that it supports dynamic load balancing, so requests for historical blockchain data can be retrieved in a timely and concurrent manner. There is a great deal of research efforts in developing dynamic load balancing algorithms in the context of cloud computing and P2P systems. In [20], the authors proposed a load balancing scheduling algorithm for virtual server clusters applied to storage systems to ensure uniform load distribution of virtual server clusters. Their approach is based on the state of the server clusters and periodically sends collected feedback to the load balancer to bring the internal load performance of the system to a more balanced state. In [21], the authors introduced a strategy to use a dynamic hashing scheme to locate data keys based on a structured P2P architecture and maintain the load balance among the peers. They showed that the load balancing of P2P systems can be significantly improved using their proposed method. In [22], the authors proposed a dynamic load management algorithm for cloud computing based on the current state of virtual machines (VM). In their approach, the allocation table is parsed to find each idle and available VM, from which the active load of all VMs under consideration is calculated. Similarly, in our *HistoChain* approach, we utilize a shared assignment table to achieve dynamic load balancing within the super peer group, where the assignment is determined on the basis of the lowest total workload of the super peers. In this sense, our approach complements existing dynamic load balancing mechanisms in cloud computing and P2P systems and provides a simple yet efficient solution to support concurrent processing of complex query requests for current and historical blockchain data.

3. Scalable Storage Using Historical Blockchains

3.1. A Framework for Scalable Blockchain Networks

Data storage technologies such as physical storage and cloud storage each have their inherent advantages, but this meteoric rise in blockchain-enabled applications has led to a great deal of research focused on decentralized storage for managing large amounts of data while maintaining its viability for nodes and networks. To demonstrate this storage requirement, we examine an example of blockchain applications in healthcare. A patient visiting a hospital may generate a certain amount of data, especially in the case of multimedia files such as X-rays or CT scans. If a hospital is to consider adopting blockchain technology for data storage, it must remain scalable because a large number of patients will generate large amounts of data over a long period of time. This issue is further complicated for an entire network of hospitals that utilize a consortium blockchain as a means of sharing medical data. While viable techniques do exist to store off-chain medical data, the benefits offered by using blockchain storage are compromised in this use. In this paper, we propose the *HistoChain* approach that supports the maintenance and sharing of medical data on the chain, with the burden being borne by a smaller group of well-equipped super peers representing large and resourceful hospitals in a local area. Such large hospitals will be able to dedicate more resources to the network to maintain older on-chain data stored in historical blockchains. This makes it feasible for regular peers to participate in the network to maintain the benefits and convenience offered by blockchain technology while having a much lower storage burden without moving their data off-chain. Figure 1 shows the *HistoChain* framework for a scalable consortium blockchain network.

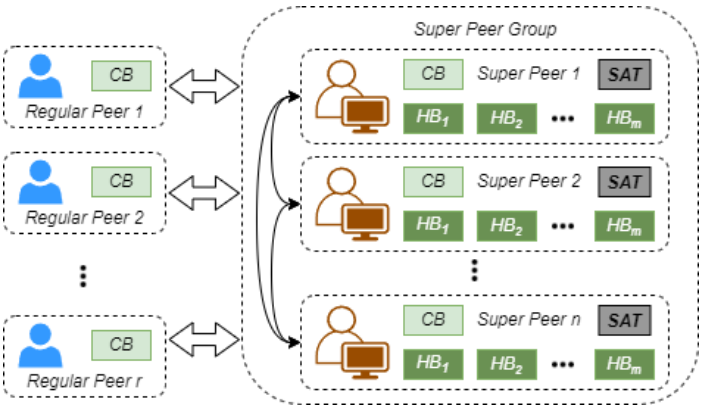


Figure 1: A Framework for a Scalable Consortium Blockchain Network

As shown in Figure 1, a consortium blockchain network consists of *n* super peers and *r* regular peers. The super peers are tasked with maintaining the current blockchain *CB* and all historical blockchains *HBs*, as well as creating and verifying new blocks and transactions using a consensus process. Shifting the burden of historical data storage and freeing regular peers from participating in the consensus process allows the introduction of highly lightweight regular peers. Regular peers maintain only the *CB*, but can access historical data stored in *HBs* through queries

to the super peer group. Upon receiving a query, a super peer splits it into subqueries and assign them to super peers based on the shared assignment table *SAT* as a means of dynamic load balancing to ensure that access to the data remains timely. More importantly, as described in Section 3.4, when the current blockchain reaches a certain age, a super peer can split it into a chain of historical blocks and a reduced chain of current blocks.

3.2. The Block Structure

A block, as a building block of a blockchain, can be defined by three parts: the block header, the list of transactions, and the verification section. Figure 2 shows the structure of a block with a list of *m* transactions.

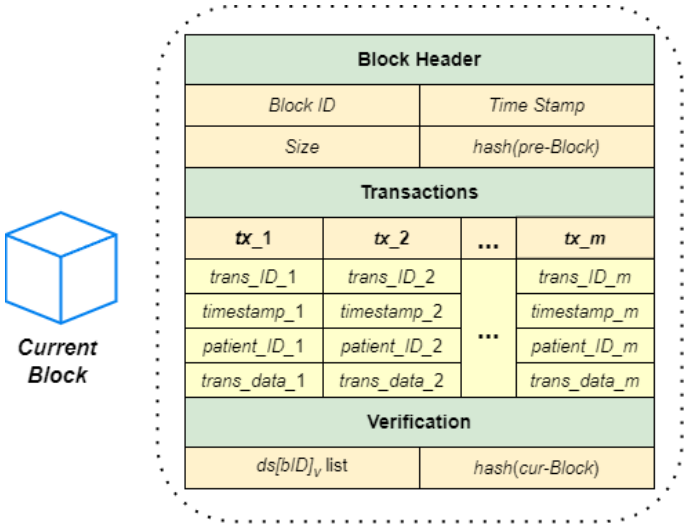


Figure 2: The Structure of a Block with a List of Transactions

As shown in Figure 2, the block header is defined as a 4-tuple (*B*, *T*, *S*, *H*), where *B* is the block ID, *T* is the timestamp when the block is created, *S* is the size of the list of transactions recorded in the block, and *H* is the hash value of the previous block. In the context of healthcare, each transaction in the transaction list is defined as a 4-tuple (*TI*, *TS*, *PI*, *TD*), where *TI* is the transaction ID, *TS* is the timestamp when the transaction is created, *PI* is the patient ID, and *TD* is the transaction data, including text-based messages and images files. The verification section is essential for the integrity of the blockchain storage, which includes a list of digital signatures, $ds[bID]_v$, for a block with ID *bID*, where *v* is a super peer that approves it as a new block in the consensus process. Any pending block must be approved by the majority of the super peers before it can be added to the blockchain, at which point the hash of the block is computed by applying a hash function to the block file containing all the above components excluding the verification section, and the hash value *hash*(*cur-Block*) is attached to the end of the block file. Note that in order to limit the block size, each block contains no more than 500 transactions and only contains transactions created during the same day. Therefore, the last block created at the end of a day may contain less than 500 transactions.

3.3. The Structure of a Meta-Block

To support efficient data retrieval in a blockchain, we define a *meta-block* as a special block that stores metadata for the current blockchain or each of the historical blockchains. A meta-block is the only mutable block in a blockchain and is attached at the beginning of the blockchain. Figure 3 shows the structure of a meta-block. As shown in the figure, a meta-block consists of two parts: the block header and a HashMap *HM*. The block header is defined as a 4-tuple (*SD*, *ED*, *SB*, *EB*), where *SD* is the timestamp of the first transaction in the first block of the blockchain; *ED* is the timestamp of the last transaction in the last block of the blockchain; *SB* and *EB* are the block IDs of the first block and the last block of the blockchain, respectively. In the second part, the HashMap *HM* contains a list of <*key*, *value*> pairs, where the key is a patient ID and the value is a list of locations where the patient transactions are stored. Each location is defined as a triple (*B*, *A*, *O*), where *B* is the block ID, *A* is the address of the transaction in the block, and *O* is the offset of the transaction size.

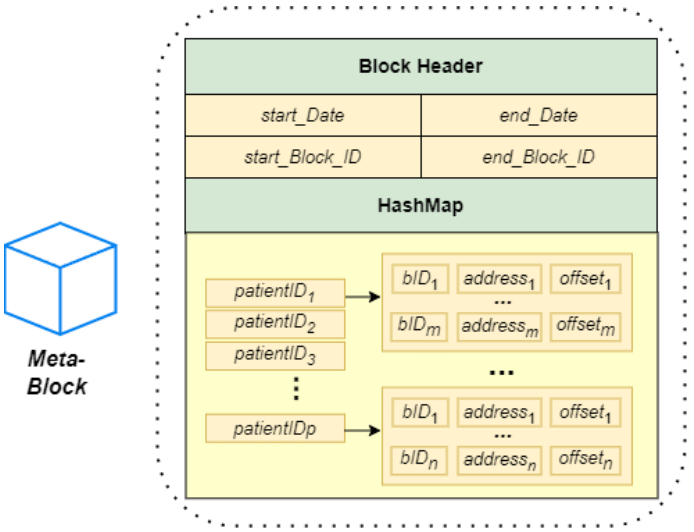


Figure 3: The Structure of a Meta-Block

The use of meta-blocks in a blockchain network provides an additional layer of organization and structure. By placing metadata in a separate block attached to the beginning of a blockchain, searching for information in the blockchain becomes much easier. This metadata allows peers to determine the exact location of transactions in the blockchain that need to be extracted to complete queries on current and historical blockchain data. Thus, the search space is much reduced and the time it takes to execute a query can be minimized. Note that to ensure the integrity of the blockchain metadata, a meta-block can be reviewed, validated and refreshed at any point in time by reading data from the relevant part of the blockchain.

3.4. Generation of a Historical Blockchain

At its inception, the current blockchain is the only blockchain in the network. When the current blockchain reaches a certain age, say 10 years, a split occurs. The oldest 5 years of data are transferred to a new blockchain, called a historical blockchain, while the most recent 5 years of data remain in the current blockchain. When a new historical blockchain is generated, a new meta-block containing its metadata is appended to the beginning of the historical blockchain, and the current blockchain's meta-block is refreshed to reflect the movement of that data. This process is repeated 5 years later when the current blockchain again

contains 10 years of data. Figure 4 shows how the current blockchain *CB* is split into a historical blockchain and a new current blockchain. Let the block IDs of the first and last block in *CB* be *m* and *n*, respectively. Note that *m* = 1 if the current blockchain has never been split before. Let block *k* be the most recent block in *CB* that is at least 6 years old. We establish blocks *m* through *k* as a historical blockchain *HB* and generate a new meta-block MB_{HB} for it. Blocks *k*+1 through *n* persist as the updated current blockchain, while blocks *m* through *k* are removed. The meta-block MB_{CB} associated with the current blockchain is refreshed by scanning the data in the new current blockchain (i.e., blocks *k*+1 through *n*). We now have an updated current blockchain and a historical blockchain, each containing 5 years of data.

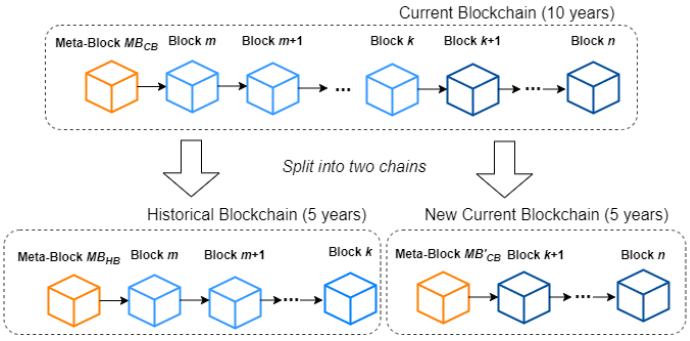


Figure 4: A Blockchain Split into a Historical and a Current Blockchain

A super peer is responsible for splitting a current blockchain with a certain age into a reduced current blockchain and a historical blockchain. When a super peer completes this task, it broadcast the updated current blockchain to all peers and the new historical blockchain to all super peers for updating. Algorithm 1 shows the process of splitting the current blockchain *CB* with 10 years of data into a historical blockchain *HB* and an updated current blockchain *CB*.

Algorithm 1: Splitting of a Current Blockchain

Input: A current blockchain *CB* containing 10 years of data
Output: Historical blockchain *HB* with 5 years of old data and an updated *CB* with the most recent 5 years of data

1. Let *m* and *n* be the IDs of the first and the last block in *CB*
2. Let *k* be the most recent block at least 6 years old, where *n* > *k*
3. Extract blocks *m* through *k* from *CB* and create a new historical blockchain *HB* with the *k*-*m*+1 blocks
4. Create an empty meta-block MB_{HB} associated with *HB*
5. Set *SD* in MB_{HB} as the date of the first transaction in block *m*
6. Set *ED* in MB_{HB} as the date of the last transaction in block *k*
7. Set *SB* and *EB* in MB_{HB} to *m* and *k*, respectively
8. **for** each block β in *HB*
9. Scan block β and add each triple (*B*, *A*, *O*) associated with *patientID* α to a list LS_α
10. Create a HashMap in MB_{HB} and add all pairs of $<\alpha, LS_\alpha>$ to it
11. Attach MB_{HB} to the beginning of *HB*
12. Remove blocks *m* through *k* from *CB*
13. Update *CB*'s meta-block MB_{CB} accordingly, as with MB_{HB}
14. **return** *HB* and *CB*

As shown in Algorithm 1, the meta-block of *HB*, MB_{HB}, contains the date of the first transaction in the first block of *HB*, the date of the last transaction in the last block of *HB,* and the block IDs of the first and last block of *HB*. To create a HashMap that contains all <*key*, *value*> pairs, each block in *HB* is scanned, and each triple (*B*, *A*, *O*) associated with the patient ID α is added to a list LT_α. Once the scanning process is complete, all pairs of $<\alpha, LT_\alpha>$ are added to the HashMap in MB_{HB}. Now in *CB*, all blocks that have been recorded in *HB* are deleted, and the meta-block of the updated *CB* must be refreshed by removing all triples that reference transactions that have been transferred to *HB*. Finally, the new *HB* and the updated *CB* are returned for broadcasting.

4. Retrieval of Historical Blockchain Data

4.1. Load Balancing Data Retrieval Requests

In the context of blockchain applications in the healthcare domain, suppose a regular peer (e.g., a doctor) queries patient information from blockchains for multiples of 5 years. When the data to be searched is for the most recent 5 years, the regular peer can search directly from its local current blockchain. When the data to be searched is for the past *sLen* years, where $sLen \in \{5n \mid n \geq 2\}$, the regular peer can search for patient information for the most recent *c* years directly from its local blockchain, where *c* is the age of the current blockchain; while the remaining (*sLen* - *c*) years of data must be retrieved from the historical blockchains by making a query to any of the super peers. The request for such a query involves a patient ID (for which data is collected) and the number of years of data being search, called the *search length*. Figure 5 shows the querying process for accessing historical blockchain data.

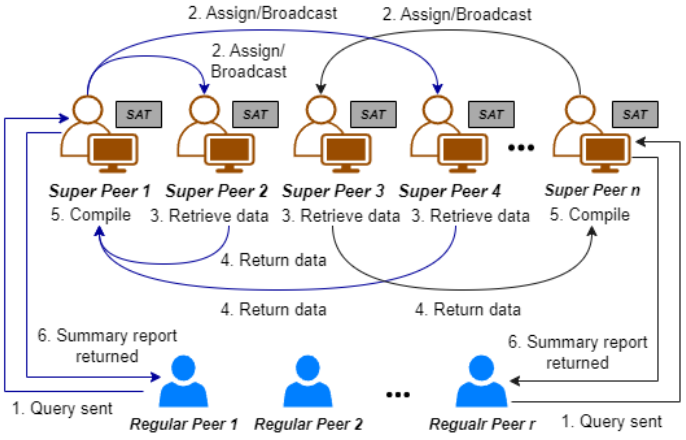


Figure 5: Querying Process for Accessing Historical Blockchain Data

From Figure 5, we can see that when a super peer receives a query from a regular peer, it acts as a director, dividing the query into subqueries and distributing them evenly based on the weights of queries to be completed by the super peers. Each query receives a weight based on the search length. For example, for a blockchain with a current blockchain of 7 years, a query with a search length of 20 years can be split into three subqueries, a 3-year search and two 5-year searches with weights of 3/5 and 1, respectively. Data retrieved from all subqueries are returned to the assigning super peer (if not completed by the assigning super peer) and compiled into a single summary report that is returned to the requesting regular peer. Note that the search for the most recent 7 years of blockchain data must be performed locally by the requesting

regular peer, who is responsible for combining its local report with the summary report received from the assigning super peer into a single summary report.

A new subquery with a 5-year search length is always assigned first to the super peer with the lowest total weight. Subqueries sent to super peers are stored in their query queues, and the total weight of the queries assigned to each super peer must be approximately equal. A super peer processes subqueries in its query queue on a first-come, first-served basis. When a super peer retrieves relevant historical data for a subquery, it compiles the results and returns a response to the assigning super peer, including a summary report of the relevant transactions with links to associated files that a regular peer can download. Note that each super peer maintains its own copy of the historical blockchains; therefore, searches assigned to multiple super peers can be performed by them simultaneously.

4.2. The Structure of Shared Assignment Table

In our proposed *HistoChain* approach, the lightweight nature of regular peers allows for a scalable architecture with the super peers facilitating access to historical data. To this end, queries for patient data are sent from regular peers to the super peers and the summary reports are returned upon completion. To ensure efficient execution of this approach, *dynamic load balancing* is employed, where each super peer retains a copy of *SAT* and broadcasts an updated *SAT* with any assignment changes made by the super peer to the super peer group. These broadcasts contain the current workload and assignment of each super peer as well as estimated time when the subquery must be completed. In case no response is received by the end of the estimated completion time, the task must be completed at the highest priority by the assigning super peer to avoid further delay. Figure 6 illustrates the structure of an *SAT* shared by a group of *n* super peers.

Shared Assignment Table (*SAT*)		
Time Stamp	*Publisher ID*	
Super Peer ID	**Query Queue**	**Total Weight**
Super Peer 1	{ [Query_ID.SubQuery_ID $q_1.sq_1$, Assigning_ID SP_1 , Receiving_ID SP_1 , time_estimate te_1] ... }	tw_1
Super Peer 2	{ [Query_ID.SubQuery_ID $q_2.sq_1$, Assigning_ID SP_2 , Receiving_ID SP_2 , time_estimate te_2] ... }	tw_2
•••	•••	•••
Super Peer n	{ [Query_ID.SubQuery_ID $q_n.sq_1$, Assigning_ID SP_n , Receiving_ID SP_n , time_estimate te_n] ... }	tw_n

Figure 6: The Structure of a Shared Assignment Table (SAT)

When a query is received by an assigning super peer Ψ, it will be split into multiple subqueries, each of which can be assigned to a super peer based on the lowest total weight. This ensures the uniform distribution of weights among the super peers and the timely completion of the subqueries. Each subquery consists of a query ID, a patient ID, a requested start date (*SD*), and an end date (*ED*), defined as a 4-tuple (*qID*, *pID, SD, ED*). Once the assignment is recorded into *SAT*, the updated *SAT* is broadcast within the super peer group. To prevent conflicts, a super peer always uses the latest version of *SAT* for the assignment by checking the publishing timestamp. A super peer may reject a subquery request due to various reasons. When this happens, Ψ must update *SAT* and broadcast it again. Algorithm 2 shows the query assignment process done by Ψ. Let the blockchain be of age 5 or more. Since the age of the current blockchain ranges from 5 to 10 years, the search length of subquery sq_1 can be less than 5 years. To avoid adding network time to the data retrieval time of short subqueries with a search length less than 5 years, the assigning super peer always completes such a subquery by itself rather than assigning it to another super peer.

Algorithm 2: Query Assignment by Assigning Super Peer Ψ

Input: Query q with a search length *sLen* in $5x$ years, $x \in [1, 10]$, current blockchain age c, shared assignment table *SAT*
Output: Updated shared assignment table *SAT*

1. **if** $sLen \leq c$ **return** *SAT* // only local search is needed
2. Split q into subqueries $sq_1 ... sq_m$, where $m = (sLen - c)/5$, search length $|sq_1| = 10 - c$, and $|sq_i| = 5$, $2 \leq i \leq m$.
3. **if** $|sq_1| < 5$
4. Assign sq_1 to Ψ *//self-assign sq_1 for less than 5-year search*
5. **else** // when $|sq_1| == 5$
6. Assign sq_1 to the super peer with the lowest total weight in *SAT*
7. **for each** subquery ρ in $sq_2 ... sq_m$
8. Assign ρ to the super peer with the lowest total weight in *SAT*
9. Broadcast updated *SAT* to all super peers
10. **return** updated *SAT*

4.3. Retrieval of Historical Blockchain Data

We now define the procedure for the retrieval of historical data by a super peer *SP*. Let subquery ρ, defined as a 4-tuple (*qID*, *pID, SD, ED*), be a subquery assigned to *SP* by an assigning super peer Ψ, then the search length of the subquery $|\rho|$ must be no more than 5 years that is covered by one of the historical blockchains. To identify the historical blockchain to be searched, *SP* needs to compare the start date *SD* and end date *ED* of the subquery with those of the historical blockchains by examining their meta-blocks. Once the historical blockchain is identified, the search is facilitated by investigating again its meta-block, which contains indices specifying the exact location of transactions in the identified historical blockchain. Algorithm 3 shows how historical data can be retrieved from historical blockchains by super peer *SP*.

Algorithm 3: Historical Blockchain Data Retrieval (Subquery)

Input: Subquery ρ as a 4-tuple (*qID*, *pID, SD, ED*)
Output: A summary report with retrieved historical data for ρ

1. Create an empty summary report *SR_ρ.qID*
2. **for each** historical blockchain Π
3. Read $MB_{\Pi}.SD$ and $MB_{\Pi}.ED$ from meta-block MB_{Π}
4. **if** $MB_{\Pi}.SD > \rho.ED$ || $MB_{\Pi}.ED < \rho.SD$
5. **continue** // outside of the search period, search next Π
6. Get a list of triples *LTX* from $MB_{\Pi}.HM$ with *pID* as the key
7. **for each** triple (B, A, O) in *LTX*
8. Read transaction *tx* from block B at address $[A, A + O]$
9. **if** $tx.TS \geq \rho.SD$ && $ts.TS \leq \rho.ED$
10. Add retrieved *tx* and links to relevant files to *SR_ρ.qID*
11. **break** // only one Π needs to be searched for subquery ρ
12. **return** summary report *SR_ρ.qID*

As shown in Algorithm 3, the patient ID in the subquery ρ is used as the key in the meta-block's HashMap to access the exact locations of relevant transactions in the associated historical blockchain. Super peer *SP* then reads the relevant transactions and record them in a summary report *SR_ρ.qID*. The summary report may contain links to multimedia files, which are hosted by *SP*. Finally, the summary report *SR_ρ.qID* is returned to the assigning super peer Ψ.

Based on Algorithm 2 and Algorithm 3, we now define the entire process by which the assigning super peer Ψ completes a summary report for a query q with a search length *sLen*, made by a requesting regular peer. This query completion process done by Ψ is described in Algorithm 4.

Algorithm 4: Query Completion by Assigning Super Peer Ψ

Input: Query q with a search length *sLen*, shared assignment table *SAT*
Output: A completed summary report *SR_Ψ*

1. Invoke Algorithm 2 on q to create and assign subqueries
2. **if** any assigned subquery *sq* is rejected by a super peer
3. Assign *sq* to Ψ itself
4. Broadcast the updated *SAT*
5. **for each** super peer *SP* with an assigned subquery ρ
6. Wait summary report *SR_ρ.qID* to be received from *SP* after *SP* invokes Algorithm 3
7. **if** time estimate of ρ is exceeded and *SR_ρ.qID* is not received
8. Assign ρ to Ψ itself and invoke Algorithm 3
9. Remove subquery assignment for *SP* from *SAT*
10. Broadcast updated *SAT*
11. Compile each *summary report* into a complete report *SR* Ψ
12. **return** summary report *SR_Ψ*

As shown in Algorithm 4, upon receiving query q, the assigning super peer Ψ splits it into subqueries and assigns them to super peers by invoking Algorithm 2. If any assigned subquery ρ is rejected by an assigned super peer *SP*, ρ is reassigned to Ψ itself and an updated *SAT* is broadcast. Each super peer *SP* receiving an assignment then retrieves the historical data requested in the assignment by invoking Algorithm 3. The assigning super peer Ψ then awaits the summary report from each *SP*. When a summary report for a subquery ρ is returned, Ψ removes the corresponding assignment in its *SAT* and broadcasts this update. If any subquery is not completed by the time estimate, Ψ assigns the subquery to itself, broadcasts an updated *SAT*, and completes the subquery. When all summary reports for the subqueries become available, Ψ compiles them into a complete final summary report (in chronological order of the subquery start date) and returns it to the requesting regular peer.

Note that a regular peer can perform a local search for a query whose search length is equal to the age of the current blockchain in a similar manner. Remote searches of historical blockchain data by super peers are conducted concurrently with the local search of the current blockchain by a regular peer. The historical data returned from a remote search is then merged with the local search data by the requesting regular peer.

5. Case Study

In this section, we present a series of simulations in the context of healthcare to demonstrate the feasibility and effectiveness of the *HistoChain* approach. In our experiments, we assume that there are 10 large local hospitals participating in a consortium blockchain network. There are also 30 small and medium medical facilities in the network. A consortium blockchain may have a 50-year lifespan, which is enough time to aggregate a substantial amount of data to be useful for experiments. We limit the total number of transactions in each block to 500, where each transaction may contain medical data in the form of image and text files. For simulation purposes, the number of visits per day is between [200, 500] and [50, 200] for large hospitals and small/medium sized medical facilities, respectively.

5.1. Estimation of Blockchain Size

To estimate the blockchain sizes along years, we use a time-based partitioning method to generate historical blockchains. A time-based partitioning occurs in the 10th year of the current blockchain; the earliest 5 years of data make up an historical blockchain, while the most recent 5 years of data are retained by the current blockchain. Using this method, super peers representing large local hospitals retain all historical blockchains as well as the current blockchain, while regular peers representing small/medium sized medical facilities store only the current blockchain. Table 1 lists the parameters used in our experiments.

Table 1: Parameters Used for Blockchain Size Estimate

Image occurrence (%)	Image size	Image count	Text occurrence (%)	Text size	File size growth rate (%)	Time to split (year)
5%	1 ~ 3 MB*	1 ~ 5	100%	0.003 ~ 0.007 MB*	0, 1, 3, 5	10

* File sizes are subject to increase by a 5-year file size growth rate.

As shown in Table 1, for a hospital visit, we assume that there is a 5% probability of including images, such as x-rays, in the doctor's notes. The size of the images is typically between [1MB, 3MB] and the number of images attached is limited to 5. The sizes of text-based medical records are also listed in Table 1. Note that in our experiments, we consider 5-year file size growth rates of 0%, 1%, 3% and 5%, with the file size bound increasing uniformly each year over the 5-year period. For example, when the 5-year growth rate is 3%, the image size increases by 0.6% per year and the maximum image size can reach 4.89 MB in 50 years, which is usually large enough for medical image files.

We now simulate the creation of a 50-year blockchain to estimate the storage burden of regular and super peers in the network. On each day, a large hospital or a small/medium sized medical facility in the network generates a random number of visits within the given range [200, 500] or [50, 200], respectively. A transaction is generated for each visit and stored in a block that can hold up to 500 transactions, independent of the transaction size. Each transaction has a 5% chance of including at least one image file. If a transaction does include image files, the number of image files is chosen randomly within the given range [1, 5]. In addition, the size of each image file or text file is also randomly generated within the certain ranges, as defined in Table 1.

To address the possible growth of image and text file sizes along years, we consider 5-year file size growth rates of 0%, 1%, 3% and 5% in our experiments. For each growth rate, we collected data from a sample of 10 simulations to establish the mean of the evaluation. The 0% growth rate is included as a baseline; while

not a realistic assumption, this establishes the minimum size of the blockchain against which the other growth rates can be considered. Figure 7 shows the change of blockchain storage along the years for the entire blockchain (including both current and all historical blockchains). The experimental results show that the effectiveness of using a historical blockchain structure is evident. After 50 years, the storage volume of the entire blockchain exceeds 33 TB at 0% growth rate, 35 TB at 1%, 38 TB at 3%, and 43 TB at 5%. Due to the storage burden, this would not be a viable solution for regular peers to store the entire blockchain.

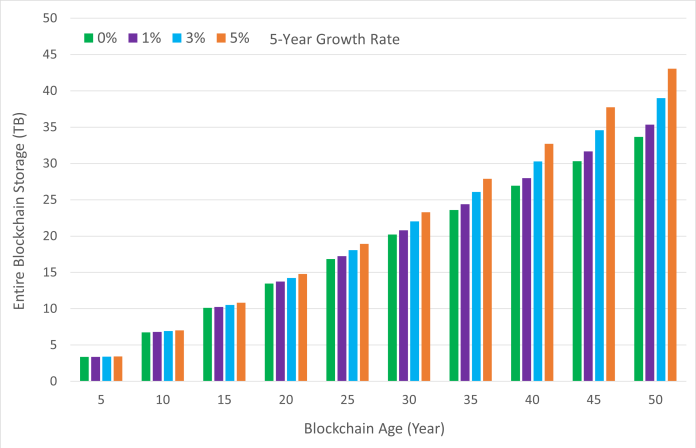


Figure 7: Total Blockchain Size by Year with 5-Year Growth Rates

Now, with the introduction of the historical blockchain structure, the storage load for regular peers can be greatly reduced, as regular peers no longer need to store the entire blockchain. Figure 8 shows the change of blockchain storage along the years for the current blockchain. The experiment records the size of the current blockchain in the year before the current blockchain split (e.g., year 4, year 9, year 14, etc.) to show the approximate maximum size of the current blockchain along the way.

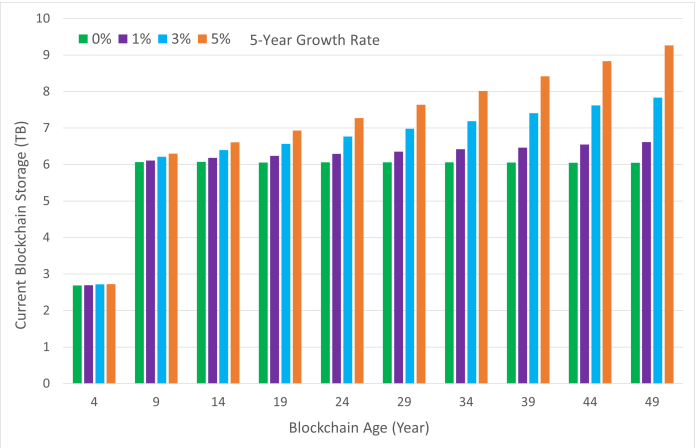


Figure 8: Current Blockchain Size by Year with 5-Year Growth Rates

From Figure 8, we can see that the size of the current blockchain is much smaller than the size of the entire blockchain. At 0% growth rate, the current blockchain size is at most 6.05 TB; at 1%, 6.62 TB; at 3%, 7.83 TB; and at 5%, 9.26 TB. The results show that for a growth rate of 0%, the size of the current blockchain is consistent regardless of the age of the blockchain. At a growth rate of 5%, the size of the current blockchain increases with the year but remains manageable for regular peers. Note that the size of the current blockchain doubled from year 4 to year 9 because the current blockchain did not need to split during those 9 years.

5.2. Data Retrieval Time for a Single Request

In this experiment, we measure the data retrieval time for a single query request for blockchain historical data by a regular peer. The data retrieval request is a search for a patient's medical records within a specified number of years. For any search within the current blockchain age, the data can be readily retrieved from the current blockchain; however, when the search length is greater than the current blockchain age, a query needs to be sent to a super peer to identify the relevant data and retrieve them from the historical blockchain(s). In this experiment, we let the age of the entire blockchain be 50 years old; therefore, up to 50 years of data can be retrieved from the blockchain. Table 2 lists additional parameters used for data retrieval in the simulations.

Table 2: Parameters for Data Retrieval Used in the Simulations

Search length (year)	Annual patient visits	File size growth rate (%)	Network latency time	Data extraction time	Data export time	Average meta-block size
5, 10, 15, …, 50	1 ~ 7	3	0.5 seconds	0.02s /MB	0.017s /MB	100MB

Since one of the important factors affecting the search time is in reading meta-blocks of the historical blockchains that contain 5-years of data, we consider search lengths in 5-year intervals up to 50 years. A 50-year blockchain also means that the current blockchain has just been split, so the current blockchain contains only the most recent 5 years of data. This setting helps to show the data retrieval time for the maximum amount of historical data. For a 5-year search, it will only be processed by a regular peer. For any search length of 10 years or more, the most recent years of data will be retrieved by a regular peer and the rest of data must be retrieved by super peers.

We assume a maximum of 7 hospital visits per patient per year and set a file size growth rate of 3% for 5 years, which allows for a reasonable increase in the size of medical image files and text files. Parameters such as image size bounds, image count bounds, text size bounds, and probability of occurrence of images in medical records can be found in Table 1. For search length of 10 years or more, measuring data retrieval time requires consideration of the *network latency time* for searching data in the historical blockchain(s), *data extraction time* for extracting index information from the relevant meta-blocks and the data from relevant blocks, and *data export time* for writing the extracted historical transaction data to a summary file. While the exact location of a transaction in a historical blockchain can be determined in constant time from the index information stored in a meta-block, opening a meta block file and reading the data from the file takes nontrivial time. Based on the average size of the meta-blocks, retrieving the index information from a meta-block can take up to several seconds. Since in our experiments, transactions are generated randomly, the extraction time is dependent upon the size of the transactions. For historical blockchain data, a super peer needs to write the extracted transaction data to a summary file. If a request is split by an

assigning super peer and completed by multiple assigned super peers in parallel, the summary reports returned must be compiled by the assigning super peer and returned to the requesting regular peer. This amount of time is included in the data export time, where a longer query adds more compilation time as it can be split into more subtasks and more reports need to be compiled.

We call our approach *decentralized, fine-grained* because there is no single trusted peer for load balancing; instead, dynamic load balancing is utilized by each super peer in the group based on the *SAT*. We now compare our decentralized, fine-grained approach to a *centralized, coarse-grained* approach [8], where a search query is processed by a single super peer, regardless of the search length. Figure 9 shows the results of 30 simulations for the centralized and decentralized approaches for each given search length up to 50 years.

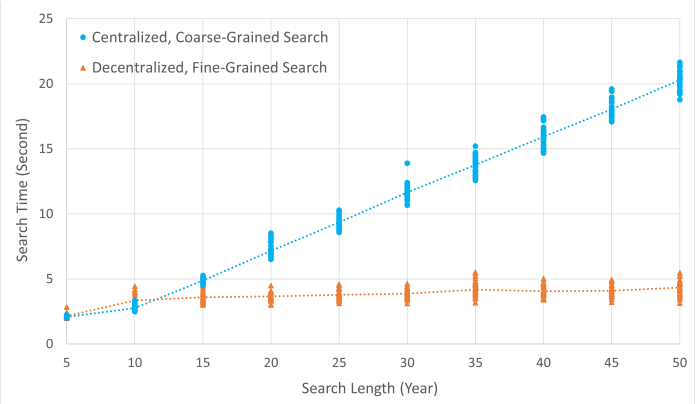


Figure 9: Retrieval Time for Individual Request with Varying Search Length

From Figure 9, we can see that for 5-year search, the average search time for both approaches is about 2 to 3 seconds. This is because the 5-year search can be processed locally by a regular peer and does not require remote data retrieval by super peers. The average 10-year search time with the decentralized approach is slightly larger, which can be attributed mainly to the increase in network time; the assigning super peer may need to delegate a remote 5-year search to another super peer and await its response. Otherwise, the search time would be the same, since the remote search for 5-year data is handled by one super peer in both methods. As the search length increases, the data retrieval time increases accordingly, with a maximum of about 20 seconds in the centralized approach for a 50-year search length. We see this growth is approximately linear, which is expected because the searches in multiple historical blockchains are performed sequentially by a single super peer, rather than in parallel by multiple super peers. In contrast, in the decentralized approach, there is a slight initial increase in search time for a 10-year search, but this increase is flat for longer searches. We see that a 50-year parallel search takes just over 4 seconds on average. The very small increase in time from a 10-year search to a 50-year search can be explained by the time it takes to compile summary reports received from multiple super peers.

Note that the 10-year search time does not increase significantly over the 5-year search time in both approaches because the 10-year search consists of a local search by a regular peer in the current blockchain and a remote search of the remaining data by a super peer, both of which are performed concurrently. The insignificant increase in the average data retrieval time in the 10-year search in both approaches is due to the additional network time and export time caused by the remote search of the historical data.

5.3. Data Retrieval Time for Concurrent Requests

Our approach supports simultaneous processing of multiple query requests. In this experiment, we compute the distribution of weights across a group of 10 super peers for 10, 25, and 50 concurrent requests. Since requests are expected to be received at 5-minute intervals and up to 50 concurrent requests can be processed in this interval (at times of high workload), there is no overflow. Weights are assigned in proportion to the number of years involved in the search. We again consider searches involving up to 50 years of data at 5-year intervals. A 5-year search is not considered, as it can be retrieved locally from the current blockchain by a regular peer. We assume the probability of each search length occurring is equal, forming a uniform distribution. Figure 10 shows the variance of the weights assigned to each super peer in this strategy, where a number of simulations are generated for each number of concurrent query requests.

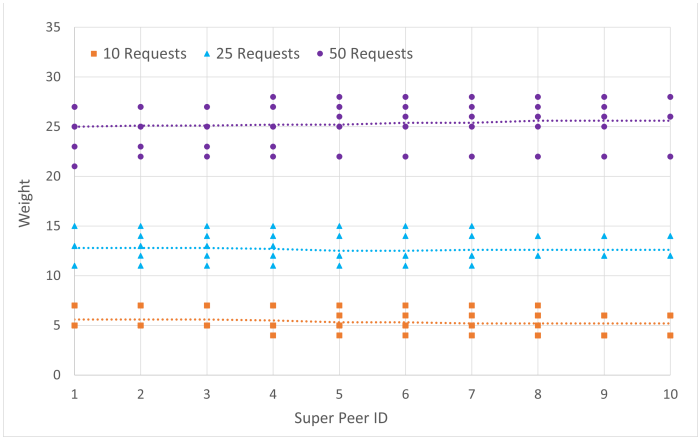


Figure 10: Distribution of Weights for Varying Numbers of Concurrent Requests

In a group of 10 super peers, queries are randomly sent to super peers who split the queries and assign subqueries to others to ensure even load balancing among the super peers. In this way, simultaneous historical blockchain data retrieval requests can be processed concurrently by the super peers. To examine the search time of concurrent data retrieval requests, the requests of regular peers for 10 to 50 years of data are measured. From Figure 10, we can see that the distribution of weights among the super peers is approximately uniform. For 10 concurrent queries, the average weight of the super peers is 5.37; for 25 queries, it is 12.65; and for 50 queries, it is 25.32. This demonstrates the effectiveness of the dynamic load balancing algorithm, which allows for even workload distribution in the super peer group and leads to efficient concurrent data retrieval by the super peers.

We further compare the centralized and decentralized approaches to demonstrate the efficiency of the decentralized, fine-grained approach. Load balancing can also be incorporated in the centralized approach, so weights are assigned to each request according to the length of the request [8]. We assume search lengths of up to 50 years and simulate 10, 20, 30, 40, and 50 concurrent searches at 5-minute intervals to calculate the total

data retrieval time. Note that 50 concurrent requests represent a very high volume of requests in a 5-minute interval, this may occur at certain times of the year, such as a flu season. We calculate the average data retrieval time for completion of all concurrent requests in a 50-year blockchain, which we refer to as the *completion time*. Figure 11 shows the average and individual completion time for the specified numbers of concurrent requests by running 30 simulations of each approach.

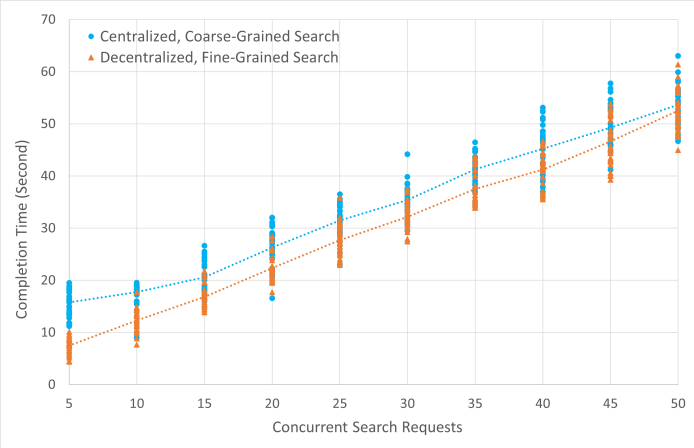


Figure 11: Completion Time for Varying Number of Concurrent Requests

From Figure 11, we can see that employing a decentralized, fine-grained approach is superior to a centralized, coarse-grained load balancing mechanism. Since in the decentralized approach, long queries are split into 5-year subqueries, the dynamic load balancing would result in more even workload distribution among super peers than in the centralized approach. On the other hand, although the average completion time of the centralized approach is higher than that of the decentralized approach for each number of concurrent requests, we note that as the number of concurrent requests increases, the queue completion times of the centralized and decentralized approaches start to converge and are almost equal at 50 concurrent requests. This is because when there are more concurrent requests, the weight distribution of the centralized approach can become more uniform and may approach the performance of the decentralized approach. This finding suggests that the decentralized, fine-grained dynamic load-balancing algorithm could be more effective in the off-season or normal season than in the peak season, although it performs better than the centralized, coarse-grained load-balancing mechanism in general.

6. Conclusions and Future Work

To address the scalability issues of consortium blockchains, recent solutions have focused on transferring data off-chain by using IPFS and cloud-based storage structures. In this paper, we propose a novel approach, called *HistoChain*, to improve consortium blockchain scalability using historical blockchains and dynamic load balancing. We introduce a time-based partitioning strategy to generate a historical blockchain, where older sections of the current blockchain are transferred to the historical blockchain after a specified time interval (e.g., 5 years). This approach allows the current blockchain to contain a useful amount of the up-to-date data, while freeing regular peers with limited resources or storage from maintaining the entire data-intensive blockchain. The historical blockchains are maintained by a group of super peers with greater resources and computing power. In addition, we introduce a meta-block, attached to a historical or the current blockchain, which serves as an index file for facilitating efficient data retrieval. To support concurrent processing of queries, we split a query into subqueries and employ a dynamic load balancing algorithm to assign the subqueries to a group of super peers. This assignment is based on a shared assignment table that records the current workload of each super peer. Once the relevant data for the query has been collected, the assigning super peer sends a summary report of the retrieved data to the requesting regular peer. Finally, we provide a case study of healthcare data storage using a consortium blockchain. The experimental results show that our *HistoChain* approach can effectively reduce the storage burden of data-intensive blockchain applications on regular peers while providing efficient access to historical data through a group of super peers.

In future work, we will implement *HistoChain* and conduct more experiments to illustrate the effectiveness of using historical blockchains to efficiently retrieve historical blockchain data in real scenarios. We will further investigate effective methods to improve the performance of concurrent data retrieval by super peers. One such method to be developed is to analyze the efficiency of parallel searches by a super peer across multiple historical blockchains. This parallelization should allow a super peer to reduce and optimize the search time if the historical blockchains are stored on different hard drives. Furthermore, a hierarchical architecture can be considered to orchestrate multiple consortium blockchains to support blockchain data sharing across cities and states. Finally, to ensure strong data privacy, it is necessary to design access control policies so that users with different roles can access blockchain data with the required permissions [23]. This is especially necessary in applications with multilevel security requirements [24], such as healthcare blockchain applications.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We thank the editors and all anonymous referees for the careful review of this paper and the many suggestions for improvements they provided. We also thank the University of Massachusetts Dartmouth for their financial support to the first author in completing this work.

References

[1] H. Guo and X. Yu, "A survey on blockchain technology and its security," Blockchain: Research and Applications, **3(2)**, February 2022, doi: 10.1016/j.bcra.2022.100067

[2] S. Nakamoto, "Bitcoin: a peer-to-peer electronic cash system," October 2008. Retrieved on January 15, 2022 from https://bitcoin.org/bitcoin.pdf.

[3] M. J. Amiri, D. Agrawal, and A. El Abbadi, "Permissioned blockchains: properties, techniques and applications," In Proceedings of the 2021 International Conference on Management of Data (SIGMOD'21), 2813-2820, Virtual Event China, June 2021, doi: 10.1145/3448016.3457539

[4] O. Dib, K.-L. Brousmiche, A. Durand, E. Thea, and E. B. Hamida, "Consortium blockchains: overview, applications and challenges," International Journal on Advances in Telecommunications, **11**(1&2), 51-64, 2018.

[5] S. Liu and H. Tang, "A consortium medical blockchain data storage and sharing model based on IPFS," In Proceedings of the 4th International Conference on Computers in Management and Business (ICCMB 2021), 147-153, Singapore, January 30 - February 1, 2021, doi: 10.1145/3450588.3450944

[6] X. Chen, K. Zhang, X. Liang, W. Qiu, Z. Zhang, and D. Tu, "HyperBSA: A high-performance consortium blockchain storage architecture for massive data," IEEE Access, **8**, 178402-178413, September 2020, doi: 10.1109/ACCESS.2020.3027610.

[7] D. P. Bauer, "InterPlanetary File System," In Getting Started with Ethereum: A Step-by-Step Guide to Becoming a Blockchain Developer, 83-96, Apress, Berkeley, CA, July 2022, doi: 10.1007/978-1-4842-8045-4_7.

[8] M. Felipe and H. Xu, "A scalable storage scheme for on-chain big data using historical blockchains," In 2022 IEEE 22nd International Conference on Software Quality, Reliability and Security Companion (QRS-C), 54-61, IEEE BSC 2022, Guangzhou, China, December 5-9, 2022, doi: 10.1109/QRS-C57518.2022.00017.

[9] J. Poon and T. Dryja, "The Bitcoin lightning network: scalable off-chain instant payments," White Paper, 2016. Retrived on September 1, 2022 from https://lightning.network/lightning-network-paper.pdf

[10] Ethereum Foundation, "DankSharding," White Paper, 2023. Retrieved on May 12, 2023 from https://ethereum.org/en/roadmap/danksharding/

[11] J. Eberhardt and S. Tai, "On or off the blockchain? insights on off-chaining computation and data," In: De Paoli, F., Schulte, S., Broch Johnsen, E. (eds) Service-Oriented and Cloud Computing, ESOCC 2017, Lecture Notes in Computer Science (LNCS), **10465**, 3-15, Springer, Cham, 2017, doi: 10.1007/978-3-319-67262-5_1.

[12] G. Wang, Z. Shi, M. Nixon, and S. Han, "ChainSplitter: towards blockchain-based industrial IoT architecture for supporting hierarchical storage," In Proceedings of the 2019 IEEE International Conference on Blockchain (Blockchain), 166-175, Atlanta, GA, USA, July 14-17, 2019, doi: 10.1109/Blockchain.2019.00030.

[13] J. Jayabalan and N.Jeyanthi, "Scalable blockchain model using off-chain IPFS storage for healthcare data security and privacy," Journal of Parallel and Distributed Computing, **164**, 152-167, June 2022, doi: 10.1016/j.jpdc.2022.03.009.

[14] D. Li, W. E. Wong, M. Zhao, and Q. Hou, "Secure storage and access for task-scheduling schemes on consortium blockchain and interplanetary file system," IEEE 20th International Conference on Software Quality, Reliability and Security Companion (QRS-C), 153-159, IEEE BSC 2020, Macau, China, December 2020, doi: 10.1109/QRS-C51114.2020.00035.

[15] E. Androulaki, A. Barger, V. Bortnikov, C. Cachin, K. Christidis, A. De Caro, D. Enyeart, C. Ferris, G. Laventman, Y. Manevich, S. Muralidharan, C. Murthy, B. Nguyen, M. Sethi, G. Singh, K. Smith, A. Sorniotti, C. Stathakopoulou, M. Vukolic, S. Cocco, and J. Yellick, "Hyperledger Fabric: a distributed operating system for permissioned blockchains," In Proceedings of the Thirteenth EuroSys Conference (EuroSys'18), Article No. 30, 1-15, Porto Portugal, April 23-26, 2018, doi: 10.1145/3190508.3190538.

[16] H. Im, K. H. Kim, and J. H. Kim, "Privacy and ledger size analysis for healthcare blockchain," In Proceedings of the 2020 International Conference on Information Networking (ICOIN), 825-829, Barcelona, Spain, 2020, doi: 10.1109/ICOIN48656.2020.9016624.

[17] Y. Xu, "Section-Blockchain: A storage reduced blockchain protocol, the foundation of an autotrophic decentralized storage architecture," In Proceddings of the 23rd International Conference on Engineering of Complex Computer Systems (ICECCS), 115-125, Melbourne, VIC, Australia, December 12-14, 2018, doi: 10.1109/ICECCS2018.2018.00020.

[18] Y. Xu and Y. Huang, "Segment blockchain: a size reduced storage mechanism for blockchain," IEEE Access, **8**, 17434-17441, 2020, doi: 10.1109/ACCESS.2020.2966464.

[19] A. Thamrin and H. Xu, "Cloud-based blockchains for secure and reliable big data storage service in healthcare systems," In Proceedings of the 15th IEEE International Conference on Service-Oriented System Engineering (IEEE SOSE 2021), 81-89, Oxford Brookes University, UK, August 23-26, 2021, doi: 10.1109/SOSE52839.2021.00015.

[20] X. Yang, H. Shi, S. Yang and Z. Lin, "Load balancing scheduling algorithm for storage system based on state acquisition and dynamic feedback," In Proceedings of the 2016 IEEE International Conference on Information and Automation (ICIA), 1737-1742, Ningbo, China, 2016, doi: 10.1109/ICInfA.2016.7832098.

[21] Y. Chang, H. Chen, S. Li and H. Liu, "A dynamic hashing approach to supporting load balance in P2P Systems," The 28th International Conference on Distributed Computing Systems Workshops, 429-434, Beijing, China, June 17-20, 2008, doi: 10.1109/ICDCS.Workshops.2008.109.

[22] R. Panwar and B. Mallick, "Load balancing in cloud computing using dynamic load management algorithm," 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), 773-778, Greater Noida, India, 2015, doi: 10.1109/ICGCIoT.2015.7380567.

[23] H. Guo, W. Li, M. Nejad, and C. Shen, "Access control for electronic health records with hybrid blockchain-edge architecture," In Proceedings of the 2019 IEEE International Conference on Blockchain (Blockchain-2019), 44-51, Atlanta, GA, USA, July 14-17, 2019, doi: 10.1109/Blockchain.2019.00015.

[24] R. Anderson, Security engineering: a guide to building dependable distributed systems, 3rd Edition, John Wiley & Sons, Indianapolis, Indiana, USA, December 2020.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 2, 97-107 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

The First Application of the Multistage One-Shot Decision-Making Approach to Reevaluate a Technology Project Decision Problem

Mohammed Al-Shanfari*

Graduate School of International Social Sciences, Yokohama National University, Yokohama, Hodogayaku, 79-4 Tokiwadai, 240-8501, Japan

ARTICLE INFO

Article history:
Received: 26 October, 2022
Accepted: 02 March, 2023
Online: 24 March, 2023

Keywords:
Multistage decision-making
One-shot decision theory
Scenario-based decision theory
Decision tree
IT project

ABSTRACT

Decision-makers must make a suitable sequence of decisions under uncertainty in a relatively long period for particular projects and situations. Conventional decision-making approaches under uncertainty are based on expected utility theory and do not sufficiently reflect the one-time nature of decisions. Similarly, the conventional approaches do not adequately incorporate the decision-maker's intuitions in the decision-analysis process. Numerous studies have demonstrated that salience information (attention-grabbing) is crucial in human decision-making exercises. However, there is limited information on the decision-making approaches incorporating the salience information and the applications of such approaches in actual practice. This study applies an approach called the multistage one-shot decision-making approach (MOSDMA) to reevaluate a previous decision problem related to a department technology project from the sultanate of Oman. Unlike traditional lottery-based approaches, MOSDMA is scenario-based, introducing an essential alternative for multistage decision-making under uncertainty. The paper is the first contribution to using the passive focus point introduced in MOSDMA in actual applications. The aim is to verify the explicability and effectiveness of the suggested method for solving decision-making under uncertainty problems in actual practice. The paper exhibits positive findings and promising potential of the approach advocating further future studies in theory and application aspects.

1. The Introduction

The case presented in this paper is the first application of the new multistage one-shot decision-making approach (MOSDMA) in reevaluating a former decision problem. The paper is an extension of work originally presented at the 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)[1]. In decision-making, decisions are typically made with a certain level of uncertainty. Uncertainty is principally deemed inherent in decision-making and significantly influences the decision alternatives. Uncertainty can be generally defined as the lack of knowledge about the probabilities of the future state of events that cannot be entirely eliminated [2]. Numerous theories have been suggested to cope with decision-making under uncertainty (e.g., [3–14]). Most existing theories adhere to the Bernoullian framework of the weighted average. Nevertheless, some decisions under uncertainty are irreversible and can be made only once, where the probability distribution is partial or insufficient. These types of decision problems are known as one-shot decision problems that could lead to significant gains or losses. In such problems, a decision-maker has only one chance to make a decision under uncertainty. Typical examples are private real-estate investments, new technology innovations, product developments, and emergency management for abnormal events. The accelerated industry clock speed environment makes one-shot decision problems extremely applicable in the technology project management fields.

Psychological experimentations studies have demonstrated that individuals systematically disregard the axioms for the expected utility and for the subjective expected utility (e.g.,[15,16]) and do not perform a summing process and weighting process (e.g.,[17–19]). Empirical studies have revealed that salience (attention-grabbing) information is crucial in human decision-making (e.g.,[20,21]). Accordingly, in [22,23] the author discusses that a decision-maker assesses alternatives based on some associated event or scenario called (the focus of a decision), which is most salient to the decision-maker because of its consequent payoff and probability, thus offering a one-shot decision theory (OSDT) [22]. In place of conventional (lottery-

*Corresponding Author: Mohammed Al-Shanfari, Yokohama, Hodogayaku, 79-4 Tokiwadai, 240-8501, Japan, mohammedshanfari@gmail.com

based) methods, the author reasons that the OSDT is needed to solve one-shot decision problems because it is scenario-based, appealing to common phenomena and intuition. When making such a one-shot decision with little or partial information, it is most common to take on one scenario, which is crucial to the decision-maker and the decision-maker's basis for reaching the desired conclusion. The OSDT presents twelve focus points that describe the decision-maker's attitude towards the possibility, satisfaction, and optimality criteria. The OSDT is generalized to the focus theory of choice (FTC) in [24–26], employing (positive and negative) evaluation systems and relative likelihood. Relative likelihood is used to measure probabilities by the highest probability event in a subset of events. Hence, as the FTC is event-based, it offers a model for practical rationality.

Detailed comparisons are offered in [23,27], to explain the advantages of OSDT and to address the differences between other decision theories based on optimistic and pessimistic utilities such as SEU. In SEU, for example, if the optimal alternative reappears many times, the total payoff gained almost confidently attains the maximum. In contrast, OSDT provides a clear answer to why an alternative is optimal when only one decision chance is left to a decision-maker. In brief, as OSDT is close to the human way of thinking, the OSDT appeals to intuition, ease of application, and explicability. A decision with OSDT results directly from human-centric decision-making, involving the decision-maker rather than just the decision analyst. This is because the decision analyst usually develops decision models based on non-human-centric methods such as the SEU. The OSDT has been successfully applied to production planning problems [28], auction problems [29], newsvendor problems for innovative products [30–32], duopoly markets of innovative products [27,33], and private real estate investment [34].

Founded on the OSDT success, the multistage one-shot decision-making approach (MOSDMA) is proposed in [35] as an extension of OSDT to cope with multistage decision-making under uncertainty problems, where decision-making can be performed only once for each stage. Extending the advantages of OSDT, MOSDMA is an essential option for multistage decision-making under uncertainty because it is scenario-based and different from other lottery-based approaches. In multistage problems, decisions are made only once at each stage to reach a final result in a series of interdependent decisions. In [36], the authors have proposed a decision model for individual multi-period consumption–investment problems utilizing the MOSDMA. In MOSDMA, according to the decision-maker's attitude towards satisfaction and likelihood, one state (focus point) is chosen at each stage. The indicated backward induction determines the sequence of optimal decisions. In such problems, the obtained sequence of optimal decisions is suitable for making a final decision. However, studies on MOSDMA are still at an early stage, particularly from the applied aspects.

Uncertainty oversight and risk management fields have evolved as essential to decision-making and project management science [37–39]. Nonetheless, studies need to gain a mutual comprehension of the portrayal of risk and uncertainty in various fields and sufficient ways to handle it [40]. For example, managerial decision-making research discussed the significance of practical and applicable models to assist decision-making under uncertainty [41], as decision-makers will be compelled to make critical decisions based on appropriate assessments. Correspondingly, recent research [41–43] established difficulties in employing mathematical models and scientific approaches in practice. For example, some challenges include limited evidence on the approaches' efficiencies, not reflecting past experiences, practicality, and lack of capabilities to apply them.

In this paper, the MOSDMA is applied to reevaluate a former information technology (IT) project decision problem. This is the first time utilizing the MOSDMA to solve a decision-making problem in actual practice. The aim is to verify the explicability and effectiveness of the proposed approach to solving decision-making under uncertainty problems in actual practice. The MOSDMA is relatively newer than OSDT; the research can contribute to closing the gap between the theory and application aspects. In the theoretical contribution, this paper extends MOSDMA for a multiple-criteria evaluation problem concerning qualitative and quantitative data [44]. Consequently, research can offer real-life applications for further improvement in the approach, alternatives evaluation stage, and decision-making process in similar fields such as IT project decision-making, decision governance, and activities related to former decision evaluation. For example, evaluations of the former decisions can be relevant to lesson-learned activities, assurance, consulting, and governance-related activities.

The remainder of the research is arranged in the following structure. Section 2 presents the case study decision problem. Then, in section 3, the problem is solved by applying the approach. Finally, sections 4 and 5 present the discussion and conclusion of this research.

2. The case study

The case study is a former Information Technology (IT) project which went through a sequence of decisions in an IT system lifecycle within a financial institution in Oman. The institution is developing and incorporating best practices in corporate governance and decision-making. The non-routine decision problems related to such projects are normally raised to a dedicated project committee for group consensus. The institution is committed to employing and improving decision-making governance practices.

An assurance function (AF) decided to implement a Department Management System (DMS) to improve and automate the department workflow, which could have been inspired by the department's needs and the country's encouragement to enhance efficiency through technology in all sectors around 2008. The DMS is a technology solution that streamline and automates the department's operations and assignments, such as planning, reporting, monitoring, and follow-up . Consequently, a vendor was chosen to deliver one of the best systems in the international market.

Although the first implementation of DMS was concluded, users could only partially utilize the system because of flows in the implementation, such as process compatibility, system reliability, and user adoption. In addition, users found that the implemented version could have been more user-friendly and sufficiently aligned with the practiced workflow. Various efforts were made to solve the identified challenges through a series of patches and customization—still, some issues needed to be fixed satisfactorily. After an extensive debate with the solution provider, the DMS was

decided to be upgraded to a newer version. Considerable person-hours were spent in revising and implementing the new version from both sides.

A time came to review the entire project as a part of the department review and the system lifecycle. Though the system may have introduced new benefits, the absence of DMS was not causing a significant hindrance to their workflow and not yielding the best-desired outcome. Therefore, a view was to present this experience and information before the decision-makers and seek a decision to abandon the project. However, the previous decision-makers felt abandoning the system would be a waste after spending a considerable amount of the contract, the experience gained with this competitive product, and the remaining retention fees. Given this rationale, the directives were to evaluate other alternatives or make an additional effort to utilize the DMS for fair use of investment.

In reconsidering the circumstances, the most recent version from the existing DMS provider could be more reliable and user-friendly. Nevertheless, the latest version will add an additional cost to the contract. In a separate endeavor for other alternatives, it was determined by an organization functioning in a similar sector that they had developed a customized in-house system for their Department. The expense was less than the current DMS, and the experience with their vendor was satisfactory. However, their locally customized system has limited features and scope compared to the international DMS product upgrade in the discussion. Moreover, details about implementation feasibility, additional costs, and future capacity are not accessible yet at the time of making the decision.

Until this point of system lifecycle, the decision-making process was mainly based on similar discussions and intuitions with limited use of scientific decision analysis tools and related mathematical decision-making approaches in the alternatives evaluation stage. However, not using these approaches may not hinder making an informed decision but can provide more context and improvements to the decision-making process for better judgment and justification. Next, the above-introduced case will be defined and reevaluated using the MOSDMA.

3. The solution

3.1. Problem description

The study employed a decision-making simulation with a focus group of mainly three participants involved in the project and aware of the decision made. The participants assist in supplying, designing, and harvesting qualitative and quantitative data sources, including interviews, discussions, documents, and workshops.

The data-gathering methodology is arranged in three main steps, as summarized in Figure 1. In step 1, an initial case review was performed to understand the case and collect foundational information for the subsequent steps. The foundational information is collected through a short questionnaire, discussions, and examination of relevant project documents. The goal is to construct the case decision story by determining the system's objective, the decision-making process, the previous alternatives, risk appetite, the type of decision-makers, and the satisfaction of the decisions made in the case. Step 1 main result is manifested in the summarized case study in section 2. Building on step 1, the research can proceed by tailoring a decision-making tool kit to harvest data related to solving this case study in step 2. The decision-making tool kit consists of a decision tree, a probability scale card, and a weighted sum scorecard. Through collaboration, three inputs of the participants are captured using the consented decision-making tool kit. The probabilities and the weight of each considered objective are donated following the decision tree in Figure 2. After the final values are placed in the finalized decision tree, the described problem can be solved. All data are detailed in the following sections.

The alternatives, in this case, are evaluated by considering three objectives: payoff as cost and benefit (CB), social impact (SI), and user satisfaction (US). First, to find the payoff (CB), the savings are obtained, as shown in Table 1. Then the CB of the three potential options is computed, as displayed in Table 2. Next, a tailored weighted sum scorecard and probability scale card are developed, as shown in Tables 3 and 4. The main weight assigned for CB, SI, and US objectives are 0.4,0.3, and 0.3, respectively. The objectives and weight are subjective to the participants' experience and agreement. Then using this kit, appropriate values are selected for each scenario. Therefore, the decision tree and the final values are visualized in Figure 2.

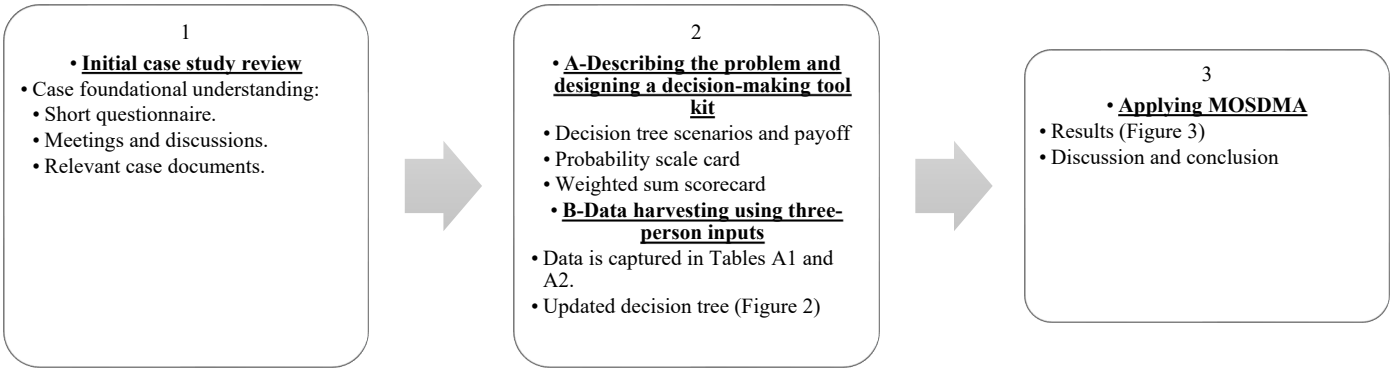


Figure 1:The data-gathering methodology steps

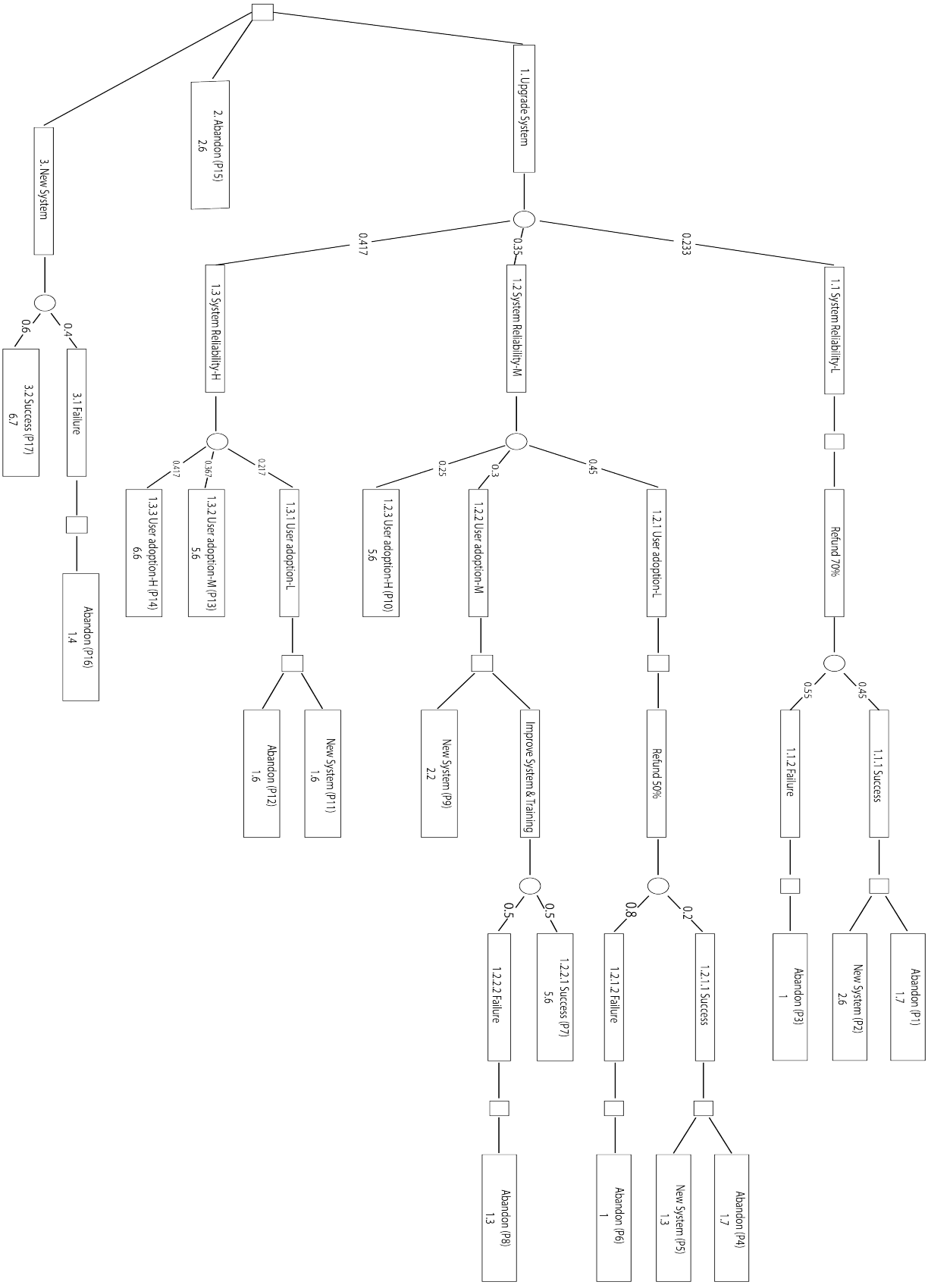


Figure 2: The decision tree showing final values.

Table 1: Estimated savings

Hours Before Automation (HBA)	Hours After Automation (HAA)	Hours Savings per assignments [a] (HS)	Total Hours Saved per year [b] (THS)	Total saved cost per year [c]	Total saved cost in 5 years
200 Hours	113 Hours	87 Hours	4350 Hours	121,800 USD[d]	609,000 USD

The above data is based on the solution forecasted benefit analysis provided by the vendor and applied to AF's annual average functions.

[a] (HS = HBA – HAA)

[b] (Average 50 assignments per year) × HS

[c] (Average hour costs 28 USD) × THS

[d] USD (United States dollars).

Table 2: Cost and Benefits (CB) estimations

Alternatives	Annual total savings cost [a]	5 years of total savings cost [a]	Previous payments (PP)	1st Year Cost (YC)	Total Setup Cost 1st Year (TYC)	4 years total annual maintenance cost (TAMC)	5 years Cumulative Cost (CC)	5 Years Net Benefits (NB)
Upgrade Current System	121,800	609,000	PP1	YC1	TYC1	TAMC1	CC1	NB1 (Highest Value)
New System	121,800	609,000	PP2	YC2	TYC2	TAMC2	CC2	NB2
Abandon	0	0	PP3	0	0	0	0	0 (Lowest Value)

[a] Values are estimated in table 1.

Values are shown in USD (United States dollars).

Table 3: The weighted sum method

Objectives	Aim	1	2	3	4	5	6	7
Cost/Benefits (CB) in USD	Max.	-300,000 or less	-299, 000 to -200,000	-199, 000 to -1	0 to 49,000	50,000 to 149,000	150,000 to 249,000	250, 000 or more
User satisfaction (US)	Max.	VL	L		M		H	VH
Social impact (SI)	Min.	VH	H		M		L	VL

(1) indicates the least favorite outcome, while (7) indicates the most favorable outcome.

VL = (Very High) , L= (Low), M = (Medium), H = (High) , VH = (Very High)

USD = United States dollars

Table 4: Probability scale card tool.

Levels	Impossible	Nearly Impossible	Very Low	Low	Moderate Low	Moderate	Moderate high	High	Very High	Extremely High	Certain
Probabilities	0	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

The CB objective is achieved for all the outcomes based on a simple premise: Payoff is equal to potential benefits after subtracting applicable costs or losses. After finalizing the payoff details, the team decides on the values for each path's objective. Then, the values are normalized based on the weighted sum scorecard described in Table 3. Finally, the values are normalized by multiplying them by the corresponding main weight of the objective. For example, in path 1, the CB, SI, and US values are normalized to the scores 2, 2, and 1 and then to the scores multiplied by the weights of the objectives 0.4, 0.3, and 0.3, which yield a final outcome of 1.7. This process of collecting and normalizing the outcome with weight scores is detailed in Table A2 in the appendix.

Based on the case study review and previous undesired outcomes, two major uncertainties were identified: system reliability and user adoption. Subsequently, the event branches low (L), medium (M), and high (H) are assigned to the parent uncertainties. The alternative (Upgrade System) has 14 possible scenario paths (numbered from P1 to P14), and the alternative (New system) has two. While the alternative (Abandon) has one certain value (P15), as expressed in the decision tree in Figure 1 and Table A2 in the appendix.

To find the probability of each scenario, the average of the three responses is taken using a customized probability scale card with 11 scales corresponding to a level and a probability, as shown in Table 4. The lowest level, "impossible," denotes a probability of 0 for the event. Comparably, the highest level, "Certain," denotes a probability of 1. Subsequently, the final outcomes and the final probabilities are positioned in the decision tree presented in Figure 2 to solve the decision problem using the MOSDMA.

3.2. Applying the approach

The multistage one-shot decision-making approach (MOSDMA) offers in which twelve types of focus points to harmonize with different types of decision-makers [36]. Out of twelve focus points, four examples of focus points characteristics are described in Table 5 by considering combinations of likelihood and satisfaction. In type (I), both likelihood and satisfaction are higher, which appears appropriate for an active decision-maker. In contrast, in type (III), it appears appropriate for an apprehensive decision-maker as although some scenario has a lower likelihood, it is still considered can induce more significant losses (as shown by the lower satisfaction level). Purchasing insurance can exemplify a type (III) focus point behavior. Type (II) looks appropriate for passive decision-makers with lower satisfaction levels and higher likelihood. Whereas in type (IV), the focus point appears proper for daring personalities. Because though the likelihood of some scenarios is lower, higher gains (higher satisfaction level) could tempt individuals to contemplate such a scenario (for example, purchasing a lottery).

Based on the focus point characteristics, the types (I), (II), (III), and (IV) are named active focus point, passive focus point, apprehensive focus point, and daring focus point, respectively.

Table 5: Characteristics of four focus points (types I-IV)

Four types of focus points	Satisfaction	Likelihood
(I) Active Focus Point (AFP)	higher	higher
(II) Passive Focus Point (PFP)	lower	higher
(III) Apprehensive Focus Point (APFP)	lower	lower
IV. Daring Focus Point (DFP)	higher	lower

Taking into account the stakeholders in this study case, only the Passive Focus Point (PFP) is considered to incorporate the overall intuition and feelings at this point of the project. First, consider a decision $a \in A$ on a decision node A at the initial stage (stage 1). Then, the outcomes and probabilities are normalized using the satisfaction function $u(x,a)$ and the relative likelihood function $\pi(x)$, as per (1) and (2) below.

$$u(x,a) = r(x,a)/\max_{x \in X} r(x,a), \quad (1)$$

where $x \in X$ stands for a state.

$$\pi(x) = p(x)/\max_{x \in X} p(x). \quad (2)$$

Then the passive focus point (PFP) of $a \in A$ is given as

$$x(a) = \operatorname{argmax}_{x \in X} \min\{\pi(x), 1 - u(x,a)\} \quad (3)$$

which means that for $a \in A$ $x(a)$ is a state that can obtain a relatively low outcome with a relatively high probability (an unfavorite scenario of a). This state mirrors the pessimistic mentality of decision-makers. Following computing all the PFPs of $a \in A$, the final optimal alternative on the decision node A denoted as $a(A)$ is chosen by

$$a(A) = \operatorname{argmax}_{a \in A} u(x(a), a), \quad (4)$$

indicates that decision-makers select the alternative with the highest outcome among the unfavorite scenarios.

In MOSDMA, the PFP of each alternative is found from the last stage (stage 4 in Figure 2), compared by their outcomes fitting to the focus point, and rolled back until the initial stage (stage 1) is reached to make the final selection. Rather than computing the expected utility of each alternative, comparing each other on a decision node and then rolling back in stochastic dynamic programming. Figure 3 has been designed to resemble the decision tree in Figure 2 to translate the results computed by applying (1), (2), (3), and (4). In this case, there are four stages: stage 4 is the last, and stage 1 is the initial stage. Stages 4, 3, and 2 are condensed to the chance nodes, and stage 1 is the primary decision node. First, (1) and (2) are employed to normalize the outcomes and probabilities to find satisfaction and likelihood values. Then, (3) is used to get PFP between siblings' branches in each stage, starting from the last stage (stage 4) to stage 2. The PFPs are highlighted in gray in Figure 3. Finally, following the migration of the outcome values corresponding to the highlighted PFPs in stage 2 to stage 1, the final decision can be selected using (4) in stage 1.

For example, starting from stage 4, to acquire the PFP between the sibling branches 1.2.2.1 and 1.2.2.2, first by applying (3), the minimum value of {π(x) , 1-u(x,a) } at each branch is found. The minimums for branches 1.2.2.1 and 1.2.2.2 are 0 and 0.768, respectively. From these two minimums, the maximum value is 0.768, which indicates that the focus point between these two siblings is branch 1.2.2.2. Subsequently, the outcome and probability values of branch 1.2.2.2 are migrated to parent branch 1.2.2. The outcome r(x, a) migrates with the same value 1.3; however, probability p(x) is multiplied by the parent's branch 1.2.2 probability (0.5´0.30 = 0.15). The same rolling-back process employed in stage 4 is replicated in stage 3. Similarly, by applying (3), the minimum values for the branches 1.2.1, 1.2.2, and 1.2.3 are 0.821, 0.424, and 0, respectively. As a result, branch 1.2.1 is the focus point with the maximum value among its sibling. Likewise, the rest of the PFPs in stage 3 are found, and their outcomes and probabilities are migrated to the parent's branches in stage 2 as in the previous stage. The outcome and probability values for parent 1.2 after migration from the child branch 1.2.1 are 1 and 0.126 (0.35 ´ 0.36), respectively. Duplicating the same process in previous stages, the PFP in stage 2 are the branches 1.1 and 3.1. For the initial stage (stage 1), only the outcomes of branches 1.1 and 3.1 are migrated to stage 1. In stage 1, as shown in Figure 3, the final outcomes 1, 2.6, and 1.4 are compared using (4) to make the final decision. Accordingly, alternative number 2 (Abandon the system) is the highest outcome among the unfavorite scenarios, which resembles a pessimistic mentality as per the PFP type.

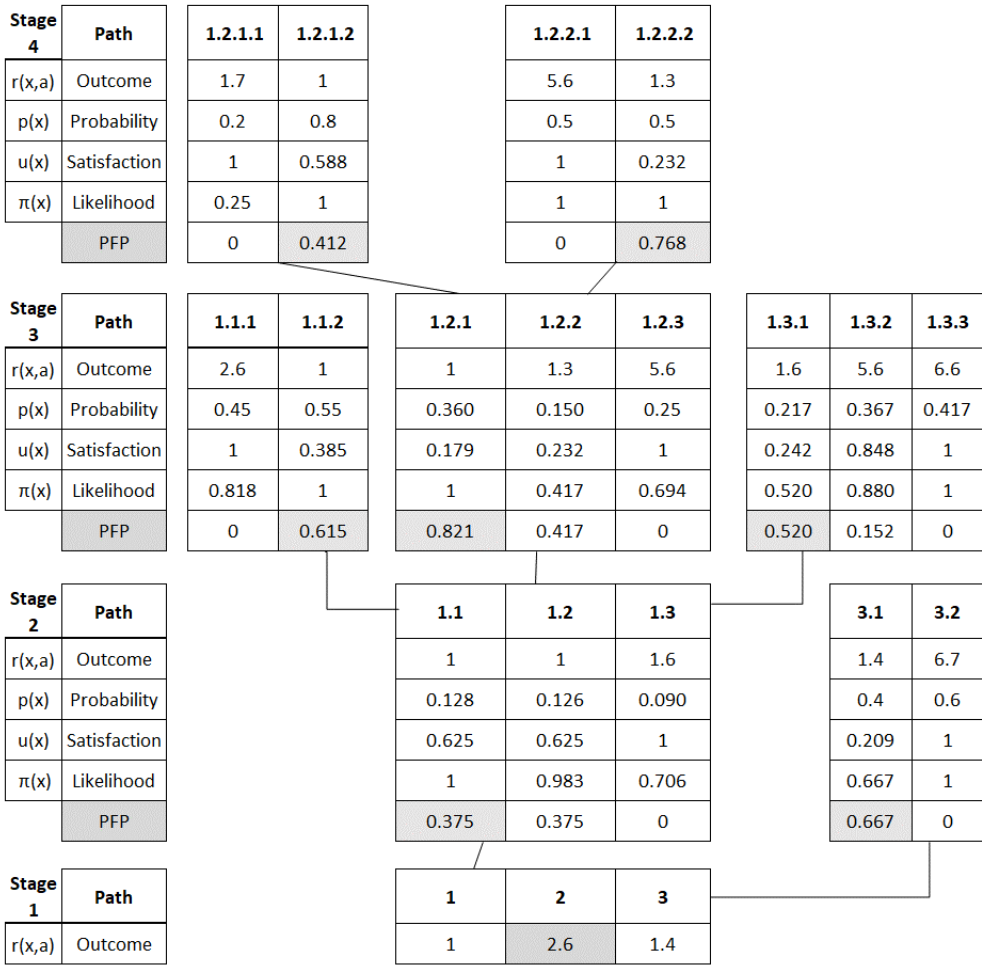


Figure 3: Applying the Passive Focus Point (PFP).

4. Discussion

In this paper, the pessimistic mentality was considered to apply the passive focus point (PFP) following the MOSDMA. The PFP brings a relatively low outcome with a relatively high probability. In the PFP type of the MOSDMA approach, the decision-maker chooses one decision that can get the highest outcome amongst the unfavorite scenarios from the decision alternatives in the initial decision stage (stage 1 in Figure 3). This paper considered the pessimistic mentality to apply the passive focus point (PFP), one of the twelve focus points introduced by the MOSDMA. The PFP obtains a moderately low outcome with a relatively high probability. The decision-makers select one conclusion to obtain the highest outcome among the unfavorite scenarios proposed in the initial decision stage, as represented by stage 1 findings in Figure 3. The endorsed alternative is the alternative (Abandon the system) with the outcome of 2.6. This decision-making mirrors the pessimistic mental set in using the PFP. The acquired empirical reevaluation of alternatives is intuitively acceptable and comparable to the actual feelings of the individuals concerned. No major difficulties were noticed in applying and understanding the approach. Participants found it uncomplicated and valuable for future use.

Eventually, despite the pessimistic feelings when deciding back around 2014, the chosen alternative was "upgrade the system," while other alternatives were discounted. The motivation was founded on the discussion of not demolishing the consumed resources; the experience gained, and the time devoured in this solution and the initiative implementation. More in-depth evaluations could give more systematic explanations and justification for the decisions made. In this case, the decision-makers did not include comparable decision analysis approaches in the alternative evaluation stage and leaned mainly on deliberating the available information and intuition. They may have believed that other alternatives may bring the lowest outcomes and satisfaction levels if unsuccessful, which could be supported using such a scenario-based method. This reveals opportunities for improvements in the alternatives evaluation stage and fills the gaps in acknowledging the value of such decision-making approaches in actual practice.

If the case is solved assuming the decision-maker is optimistic using the same approach but with an active focus point (AFP) type, the result would be "3. New System". Later, it was found that the project did not produce the best-desired outcomes. The new management is reviewing alternative 3, "discard the current product and implement a new system." However, there is a high degree of attention to reviewing the problem and improving the decision-making process. The new direction could be based on the undesired project outcomes and immaculate corporate governance improvements.

5. Conclusion

This study employs the multistage one-shot decision-making approach (MOSDMA) to revisit an actual decision problem for the first time. More studies are required from both theoretical and applied aspects, as MOSDMA is considered a new approach at an early stage. Nevertheless, this research is the first contribution involving the passive focus point (PFP) suggested by MOSDMA to reexamine a real multistage decision-making-under-uncertainty problem.

The obtained empirical reevaluation of alternatives is intuitively acceptable to the contributors. The study showed that MOSDMA could assist in reevaluating previous decisions and its capability to make an informed one with proper analysis and justifications aligned to stakeholders' satisfaction levels and intuition. This establishes the effectiveness of the MOSDMA and the promising capability of the introduced workflow to reevaluate such decision problems in similar environments. Furthermore, the approach was found reasonably explicable, practical, and systematically considering the decision-makers' intuitions.

The quality of the assessed scenarios and the gathered data is restricted to the experience and commitment of the participants in this decision-making analysis exercise. The knowledge of the future outcome and collaborating in a group setup or open disclosed style may have influenced the participant's inputs since this is a reevaluation of a past problem with currently known outcomes compared to the pressure confronted in real-time decision-making or undisclosed inputs of each participant. Since 2014, considerable restructuring has occurred in the financial institution. Accordingly, a number of applicable people involved in this case were unreachable to participate and to add more inputs to this study. Nevertheless, this limitation could be mediated since this is a decision review of a recently known outcome and a well-documented project.

Further future research is needed to improve MOSDMA theoretically and for more practical applications. For instance, forthcoming approach applications may consider more contemporary decision problems, complex stages, more data and comparisons with other approaches, other focus groups and organizations, and various details in capturing the participant inputs and reactions. Nevertheless, this study demonstrates that the approach could be employed to reexamine a former decision problem which can contribute to analyzing lessons learned and areas for improvement. Likewise, MOSDMA has the prospect of being utilized in the areas of crucial new unresolved problems, auditing, governance practices, and consulting assignments.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgment

Appreciation is extended to the immense support of the participants. The views and opinions expressed in this paper do not necessarily reflect the official policy or position of the Omani institutions.

References

[1] M. Aishanfari, P. Guo, "Application of the Multistage One-shot Decision-making Approach to an IT Project in the Central Bank of Oman," in 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 552–557, 2021, doi:10.1109/IEEM50564.2021.9672939.

[2] F.H. Knight, Risk, uncertainty and profit, Houghton Mifflin, 1921.

[3] S. Cerreia-Vioglio, D. Dillenberger, P. Ortoleva, "Cautious Expected Utility and the Certainty Effect," Econometrica, **83**(2), 693–728, 2015, doi:https://doi.org/10.3982/ECTA11733.

[4] D. Dubois, H. Prade, R. Sabbadin, "Decision-theoretic foundations of qualitative possibility theory," European Journal of Operational Research, **128**(3), 459–478, 2001, doi:10.1016/S0377-2217(99)00473-7.

[5] T. Galaabaatar, E. Karni, "Subjective Expected Utility With Incomplete Preferences," Econometrica, **81**(1), 255–284, 2013, doi:https://doi.org/10.3982/ECTA9621.

[6] I. Gilboa, "Expected utility with purely subjective non-additive probabilities," Journal of Mathematical Economics, **16**(1), 65–88, 1987, doi:10.1016/0304-4068(87)90022-X.

[7] P. Guo, Y. Wang, "Eliciting dual interval probabilities from interval comparison matrices," Information Sciences, **190**, 17–26, 2012, doi:10.1016/j.ins.2011.12.014.

[8] D. Kahneman, A. Tversky, "Prospect Theory: An Analysis of Decision under Risk," Econometrica, **47**(2), 263–291, 1979, doi:10.2307/1914185.

[9] J. von Neumann, O. Morgenstern, A. Rubinstein, Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition), Princeton University Press, 1944.

[10] J. Quiggin, "A theory of anticipated utility," Journal of Economic Behavior & Organization, **3**(4), 323–343, 1982, doi:10.1016/0167-2681(82)90008-7.

[11] L.J. Savage, "The foundations of statistics.," Naval Research Logistics Quarterly, **1**(3), 236–236, 1954, doi:https://doi.org/10.1002/nav.3800010316.

[12] D. Schmeidler, "Subjective Probability and Expected Utility without Additivity," Econometrica, **57**(3), 571–587, 1989, doi:10.2307/1911053.

[13] F. Gul, W. Pesendorfer, "Expected Uncertain Utility Theory," Econometrica, **82**(1), 1–39, 2014, doi:https://doi.org/10.3982/ECTA9188.

[14] P. Guo, H. Tanaka, "Decision making with interval probabilities," European Journal of Operational Research, **203**(2), 444–454, 2010, doi:10.1016/j.ejor.2009.07.020.

[15] M. Allais, "Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine," Econometrica: Journal of the Econometric Society, 503–546, 1953.

[16] D. Ellsberg, "Risk, ambiguity, and the Savage axioms," The Quarterly Journal of Economics, 643–669, 1961.

[17] Z. Zhou, W. Zhao, X. Chen, H. Zeng, "MFCA extension from a circular economy perspective: Model modifications and case study," Journal of Cleaner Production, **149**, 110–125, 2017, doi:10.1016/j.jclepro.2017.02.049.

[18] N. Stewart, F. Hermens, W.J. Matthews, "Eye Movements in Risky Choice," Journal of Behavioral Decision Making, **29**(2–3), 116–136, 2016, doi:https://doi.org/10.1002/bdm.1854.

[19] L. Zhou, Y.-Y. Zhang, Z.-J. Wang, L.-L. Rao, W. Wang, S. Li, X. Li, Z.-Y. Liang, "A Scanpath Analysis of the Risky Decision-Making Process," Journal of Behavioral Decision Making, **29**(2–3), 169–182, 2016, doi:https://doi.org/10.1002/bdm.1943.

[20] N. Lacetera, D.G. Pope, J.R. Sydnor, "Heuristic Thinking and Limited Attention in the Car Market," American Economic Review, **102**(5), 2206–2236, 2012, doi:10.1257/aer.102.5.2206.

[21] M.R. Busse, N. Lacetera, D.G. Pope, J. Silva-Risso, J.R. Sydnor, "Estimating the Effect of Salience in Wholesale and Retail Car Markets," American Economic Review, **103**(3), 575–579, 2013, doi:10.1257/aer.103.3.575.

[22] P. Guo, "One-Shot Decision Theory," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, **41**(5), 917–926, 2011, doi:10.1109/TSMCA.2010.2093891.

[23] P. Guo, One-Shot Decision Theory: A Fundamental Alternative for Decision Under Uncertainty, Springer, Berlin, Heidelberg: 33–55, 2014, doi:10.1007/978-3-642-39307-5_2.

[24] X. Zhu, K.W. Li, P. Guo, "A bilevel optimization model for the newsvendor problem with the focus theory of choice," 4OR, 2022, doi:10.1007/s10288-022-00520-6.

[25] P. Guo, "Focus theory of choice and its application to resolving the St. Petersburg, Allais, and Ellsberg paradoxes and other anomalies," European Journal of Operational Research, **276**(3), 1034–1043, 2019, doi:10.1016/j.ejor.2019.01.019.

[26] P. Guo, "Dynamic focus programming: A new approach to sequential decision problems under uncertainty," European Journal of Operational Research, 2022, doi:10.1016/j.ejor.2022.02.044.

[27] P. Guo, "One-shot decision approach and its application to duopoly market," International Journal of Information and Decision Sciences, **2**(3), 213, 2010, doi:10.1504/IJIDS.2010.033449.

[28] X. Zhu, P. Guo, "Bilevel programming approaches to production planning for multiple products with short life cycles," 4OR, **18**(2), 151–175, 2020, doi:10.1007/s10288-019-00407-z.

[29] C. Wang, P. Guo, "Behavioral models for first-price sealed-bid auctions with the one-shot decision theory," European Journal of Operational Research, **261**(3), 994–1000, 2017, doi:10.1016/j.ejor.2017.03.024.

[30] P. Guo, X. Ma, "Newsvendor models for innovative products with one-shot decision theory," European Journal of Operational Research, **239**(2), 523–536, 2014, doi:10.1016/j.ejor.2014.05.028.

[31] X. Zhu, P. Guo, "Single-level reformulations of a specific non-smooth bilevel programming problem and their applications," Optimization Letters, **14**(6), 1393–1406, 2020, doi:10.1007/s11590-019-01444-7.

[32] X. Zhu, P. Guo, "Approaches to four types of bilevel programming problems with nonconvex nonsmooth lower level programs and their applications to newsvendor problems," Mathematical Methods of Operations Research, **86**(2), 255–275, 2017, doi:10.1007/s00186-017-0592-2.

[33] P. Guo, R. Yan, J. Wang, "Duopoly Market Analysis within One-Shot Decision Framework with Asymmetric Possibilistic Information," International Journal of Computational Intelligence Systems, **3**(6), 786–796, 2010, doi:10.2991/ijcis.2010.3.6.9.

[34] P. Guo, "Private Real Estate Investment Analysis within a One-Shot Decision Framework," International Real Estate Review, **13**(3), 238-260., 2010.

[35] Y. Li, P. Guo, "Possibilistic individual multi-period consumption–investment models," Fuzzy Sets and Systems, **274**, 47–61, 2015, doi:10.1016/j.fss.2015.01.005.

[36] P. Guo, Y. Li, "Approaches to multistage one-shot decision making," European Journal of Operational Research, **236**(2), 612–623, 2014, doi:10.1016/j.ejor.2013.12.038.

[37] S. Ward, C. Chapman, "Transforming project risk management into project uncertainty management," International Journal of Project Management, **21**(2), 97–105, 2003, doi:10.1016/S0263-7863(01)00080-1.

[38] A. Jaafari, "Management of risks, uncertainties and opportunities on projects: time for a fundamental shift," International Journal of Project Management, **19**(2), 89–101, 2001, doi:10.1016/S0263-7863(99)00047-2.

[39] S.D. Green, "Towards an integrated script for risk and value management," Project Management, **7**(1), 52–58, 2001.

[40] O. Perminova, M. Gustafsson, K. Wikström, "Defining uncertainty in projects – a new perspective," International Journal of Project Management, **26**(1), 73–79, 2008, doi:10.1016/j.ijproman.2007.08.005.

[41] B.J. Galli, "The Future of Economic Decision Making in Project Management," IEEE Transactions on Engineering Management, **67**(2), 396–413, 2020, doi:10.1109/TEM.2018.2875931.

[42] W.G. Meyer, "The Effect of Optimism Bias on the Decision to Terminate Failing Projects," Project Management Journal, **45**(4), 7–20, 2014, doi:10.1002/pmj.21435.

[43] B.J. Galli, "Effective Decision-Making in Project Based Environments: A Reflection of Best Practices," International Journal of Applied Industrial Engineering (IJAIE), **5**(1), 50–62, 2018, doi:10.4018/IJAIE.2018010103.

[44] M. Aruldoss, T.M. Lakshmi, V.P. Venkatesan, "A Survey on Multi Criteria Decision Making Methods and Its Applications," American Journal of Information Systems, **1**(1), 31–43, 2013, doi:10.12691/ajis-1-1-5.

Appendix

Table A1: Probabilities inputs following Table 4 and the decision tree in Figure 2

Alternatives	Event ID	Probabilities			
		Participant 1	Participant 2	Participant 3	Final (Average)
1	1.1	0.25	0.25	0.2	0.233
	1.2	0.35	0.4	0.3	0.350
	1.3	Residual Probability			0.417
	1.1.1	0.35	0.4	0.6	0.450
	1.1.2	Residual Probability			0.550
	1.2.1	0.45	0.5	0.4	0.450
	1.2.2	0.25	0.35	0.3	0.300
	1.2.3	Residual Probability			0.250
	1.2.1.1	0.15	0.25	0.2	0.200
	1.2.1.2	Residual Probability			0.800
	1.2.2.1	0.45	0.55	0.5	0.500
	1.2.2.2	Residual Probability			0.500
	1.3.1	0.2	0.15	0.3	0.217
	1.3.2	0.35	0.35	0.4	0.367
	1.3.3	Residual Probability			0.417
2					1.000
3	3.1	0.45	0.45	0.3	0.400
	3.2	Residual Probability			0.600

Table A2. Collecting and normalizing the outcome with weight scores.

Final Outcomes Results				Final Outcomes Scores				Scores with Weight			
Path ID[a]	CB[b]	SI	US	CB	SI	US	Total	CB (0.4)	SI (0.3)	US (0.3)	Total
P1	******	H	VL	2	2	1	**5**	0.8	0.6	0.3	**1.7**
P2	******	H	M	2	2	4	**8**	0.8	0.6	1.2	**2.6**
P3	******	VH	VL	1	1	1	**3**	0.4	0.3	0.3	**1**
P4	******	H	VL	2	2	1	**5**	0.8	0.6	0.3	**1.7**
P5	******	H	VL	1	2	1	**4**	0.4	0.6	0.3	**1.3**
P6	******	VH	VL	1	1	1	**3**	0.4	0.3	0.3	**1**
P7	******	L	H	5	6	6	**17**	2	1.8	1.8	**5.6**
P8	******	H	VL	1	2	1	**4**	0.4	0.6	0.3	**1.3**

P9	******	M	L	1	4	2	**7**	0.4	1.2	0.6	**2.2**
P10	******	L	H	5	6	6	**17**	2	1.8	1.8	**5.6**
P11	******	H	L	1	2	2	**5**	0.4	0.6	0.6	**1.6**
P12	******	H	L	1	2	2	**5**	0.4	0.6	0.6	**1.6**
P13	******	L	H	5	6	6	**17**	2	1.8	1.8	**5.6**
P14	******	VL	VH	6	7	7	**20**	2.4	2.1	2.1	**6.6**
P15	******	H	M	2	2	4	**8**	0.8	0.6	1.2	**2.6**
P16	******	VH	VL	2	1	1	**4**	0.8	0.3	0.3	**1.4**
P17	******	VL	H	7	7	6	**20**	2.8	2.1	1.8	**6.7**

[a] Each Path ID (P1 to P17) corresponds to a path in Figure 2 with the same naming convention.

[b] The CB values are masked for the easiness, readability, and privacy reasons.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 2, 108-115 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Hybrid Machine Learning Model Performance in IT Project Cost and Duration Prediction

Der-Jiun Pang*

International University of Malaya-Wales (IUMW), Faculty of Arts and Science, Kuala Lumpur, 50480, Malaysia

ARTICLE INFO

Article history:
Received: 26 October, 2022
Accepted: 02 March, 2023
Online: 24 March, 2023

Keywords:
Machine Learning
Project Cost and Time Estimation
Budget and Duration Prediction
Hybridization

ABSTRACT

Traditional project planning in effort and duration estimation techniques remain low to medium accurate. This study seeks to develop a highly reliable and efficient hybrid Machine Learning model that can improve cost and duration prediction accuracy. This experiment compared the performance of five machine learning models across three different datasets and six performance indicators. Then the best model was verified with three other types of live project data. The results indicated that the MLR-DNN is a highly reliable, effective, consistent, and accurate machine learning model with a significant increase in accuracy over conventional predictive project management tools. The finding pointed out a potential gap in the relationship between dataset quality and the Machine Learning model's performance.

1. Introduction

This paper is an extension of work initially presented at the 2022 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE2022) [1]. Planning and estimation are imperative for any Information Technology (IT) project. Estimation aids in tracking progress and delivery velocity. However, due to the close relationship between cost and time factors, any project delay might result in cost overruns.

The investigators [2], [3] revealed that the top-ranked IT project risk is "*Underestimated Costs and Time*". According to the authors [4], 60% of IT projects have cost and time problems. Budget and timeline underestimation seems to occur at various stages of the project lifecycle. The most undesirable scenario happens when the budget and duration are underestimated at the beginning of the project lifecycle.

Artificial intelligence (AI) can improve decision-making in complex environments with clear objectives. A study concluded that, in terms of accuracy, artificial intelligence tools outperform traditional tools [5]. The value of AI can only be activated as humans and machines function complementarily integrated.

Hybridizing Machine Learning (ML) models are getting their popularity recently. According to researchers [6], hybridization effectively advances prediction models. This article focuses on the performance of various hybrid ML models in prediction accuracy enhancement to improve cost and duration estimation to address the critical IT failure problem.

*Corresponding Author: Der-Jiun Pang, Email: djpang@gmail.com

2. Methodology

2.1. The Machine Learning Model Evaluation

This study was designed to demonstrate to the research community that the evaluations are comprehensive and can explain their significance. Five hybrid ML models were developed using Python and evaluated using three different datasets, including two public datasets. These models were trained and tested on three different datasets to reduce bias caused by data quality. The best-performing ML model was selected based on the performance measured by six different metrics. It was then put forward for live project verification to determine its performance in predicting project cost and duration.

These five hybrid ML models were: Hybrid Multiple Linear Regression Deep Neural Network (MLR-DNN), Particle Swarm Optimised DNN (PSO-DNN), Hybrid Gradient Boosting Regression DNN (GBR-DNN), Hybrid Random Forest Regression DNN (RFR-DNN), and Hybrid eXtreme Gradient Boosting DNN (XGB-DNN).

Controlled experiments play a vital role in applied machine learning, and the behaviour of algorithms on specific problems must be learned empirically. A machine learning experiment procedure involves a series of steps, 1. Data collection. 2. Data pre-processing: cleaning and manipulating acquired data to prepare it for modelling. 3. Model training: the model is trained on a training dataset, usually a subset of the data collected. 4. Model tuning: change in hyperparameters to optimize the model's performance. ML performance is measured by the defined performance metrics indicated in section 2.2. 5. Model evaluation: determine the

model's performance on a testing dataset or another subset of the data collected. 6. Model deployment: the best model is then used to make predictions on live project data.

2.2. Performance Metrics

Evaluating the performance of ML models is essential to ensure their effectiveness. The choice of the performance metric is an important factor in this evaluation process. It depends on the specific ML problem being solved and the project's goals. The performance parameter used in this study is accuracy, which evaluates the number of correct predictions made as a percentage of all predictions made. The associated "accuracy" performance metrics used were $RMSE$, MAE, $RMSLE$, $MMRE$, $MdMRE$ and $Pred(m)$.

The **Root Mean Square Error (*RMSE*)** acts as a heuristic model for testing and training measures differences between predicted values and actual values from 0 to ∞. The smaller the $RMSE$, the better the model [7]. $\hat{y}_i$ is predicted output or forecasted values and y_i is the actual or observational values.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (1)$$

The **Root Mean Squared Log Error** (***RMSLE***) is a logarithmically calculated $RMSE$ commonly used metric or loss function in the regression-based machine learning model. The lesser error, the better the model is.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[log(y_i + 1) - log(\hat{y}_i + 1)]^2} \quad (2)$$

The **Mean Absolute Error** (***MAE***) measures the magnitude of errors regardless of their direction in a series of estimates. MAE is superior to $RMSE$ in terms of explanation-ability. $RMSE$ has a distinct advantage over MAE using absolute values, which is undesirable in many mathematical calculations. The smaller value, the better the model is.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (3)$$

The **Mean Magnitude of Relative Error** (***MMRE***) and **Median Magnitude of Relative Error** (***MdMRE***) are two important performance metrics derived from the overall mean and median errors. The primary function of $MMRE$ is to serve as an indicator for differentiating between prediction models. The model with the lowest $MMRE$ typically being chosen typically implies low uncertainty or inaccuracy. The better the model, the smaller the values are.

$$MMRE = \frac{1}{n}\sum_{i=1}^{n} MRE_i \quad (4)$$

$$MdMRE = \frac{1}{n}\sum_{i=1}^{n} MdMRE_i \quad (5)$$

Percentage of Estimate, $\boldsymbol{Pred(m)}$, is an alternative to the $MMRE$ that is a commonly used prediction quality metric. It simply measures the proportion of forecasts within m% the actual value. The bigger the m, the less information and confidence in a prediction's accuracy [8].

$$Pred(m) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases}1, & if\ MRE_i \le \frac{m}{100}\\ 0, & otherwise\end{cases} \quad (6)$$

2.3. Degree of Augmentation

The degree of augmentation (DOA), χ, is a prediction enhancement measurement in error reduction to measure a hybrid model. A dual-layer hybrid cascaded ML model comprises two ML models represented as layers one and two (Figure 1). In stage one, the layer one ML model makes a prediction value $\hat{y}_{t-1}$ as inputs to stage two (y_t) to be processed by the layer two ML model with prediction output $\hat{y}_t$. The difference (or error) in the predicted result versus the actual result at stage one is denoted as Δ_{t-1}.

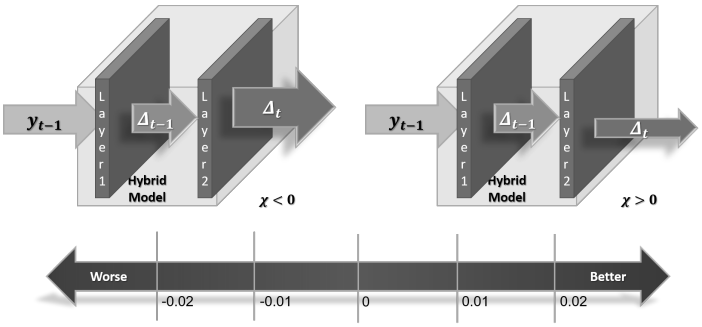


Figure 1: The Degree of Augmentation Scale

$$y_t = \hat{y}_{t-1} \quad (7)$$

$$\Delta_{t-1} = |y_{t-1} - \hat{y}_{t-1}| = |y_{t-1} - y_t| \quad (8)$$

Similarly, the difference in stage 2 is represented as Δ_t.

$$\Delta_t = |y_t - \hat{y}_t| = |\hat{y}_{t-1} - \hat{y}_t| \quad (9)$$

The assumption of difference in stage two is more diminutive than in stage one. The effect of convergence resulted in MAE reduction; therefore, augmentation occurred.

$$\Delta_t < \Delta_{t-1} \quad (10)$$

$$\hat{y}_{t-1} < \frac{y_{t-1} + \hat{y}_t}{2} \quad (11)$$

$$\chi = \Delta_{t-1} - \Delta_t = y_{t-1} - 2y_t + \hat{y}_t \quad (12)$$

By using equation (12), the $MMRE$ for stage one (Δ_{t-1}) and stage two (Δ_t) enables to calculate of the degree of augmentation, χ, for each of the hybrid models. The degree of augmentation, χ is bi-directional. A negative value indicates $MMRE$ increases or diverging, whereas a positive value specifies $MMRE$ decrease or

converges. The positive magnitude of χ shows the strength of augmentation. The higher the χ means the better the hybrid model. The more significant negative value of χ means the hybrid model is ineffective. $\chi > .01$ is considered effective, $\chi \leq 0$ is ineffective. For $0 < \chi \leq .01$ is marginally effective, which means its augmentation is not significant enough to remain effective.

In an optimistic augmentation scenario, the Interquartile Range (IQR) becomes narrower, whereas the range becomes wider in an adverse augmentation scenario. This convergent phenomenon indicates the $MMRE$ decreases in positive augmentation. Contrary, in a divergent case, $MMRE$ increases in negative boost.

2.4. Data Collection

Figure 2 illustrates the data collection procedure. Each dataset was randomly split into two groups in a 70:30 ratio, 70% for training and 30% for testing. The relevant dataset was acquired online or gathered from previous project material. The collected data was then converted (if necessary) and pre-processed using scaling (for example, the *scikit-learn* scaling package) to prepare for ML assessment.

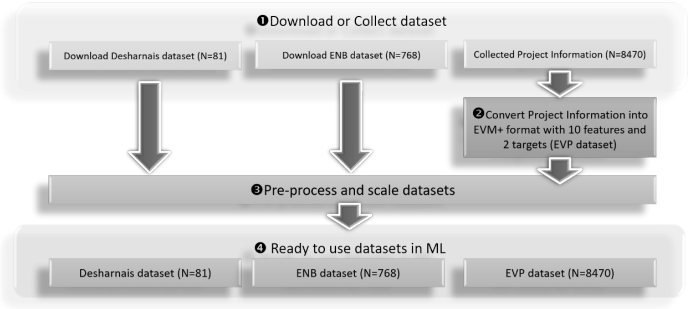


Figure 2: The data collection procedure

2.5. ML Evaluation

These ML models were evaluated in three steps depending on their algorithm settings. First, the respective models were trained using historical data in the learning or training step. Later in the testing step, these ML models were tested based on a peer comparison of their performance indicators. Each ML model was optimized through hyperparameter tuning until the best results were obtained (Figure 3).

2.6. Dataset Descriptions

A study concurs that the model may poorly correlate with a dataset that makes learning "incomplete" [9]. This evaluation used three dataset sources to minimize potential bias due to the dataset's influences. Two are publicly available, and the third dataset is a collection of actual historical project data named EVP. Both Desharnais and ENB datasets were selected in this study because of their multi-target attributes.

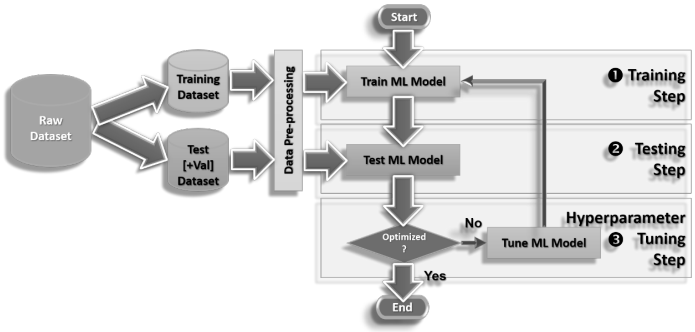


Figure 3: The ML evaluation workflow

It is challenging to ensure the quality of an ML dataset, mainly because the relationship between the qualities of the data and their effect on the ML system's compliance with its requirements is infamously complex and hard to establish [10]. In this study, dataset quality was defined as its appropriateness in terms of accuracy and value.

1) **Desharnais Dataset**

Jean-Marc Desharnais gathered the Desharnais dataset from ten organizations in Canada between 1983 and 1988. There are 81 projects (records) and 12 attributes [11], a relatively small public dataset of which four nominal fields are considered redundant in ML model evaluation. Table 1 provides statistical information about this dataset. Four entries have missing data. Most studies that use this dataset use 77 of the 81 records [12]. This study backfilled the missing fields with a "-1" value. Small dataset size issues could be compensated by adopting data-efficient learning or data augmentation strategies [13]. Desharnais datasets were used in many research. Therefore, it can benchmark the investigation against other published results.

2) **ENB Dataset**

The Energy Building Dataset [14] contains 768 instances of eight measured building parameters as feature variables. The dataset includes the two corresponding target heating load and cooling load attributes. A nominal field is considered redundant in this dataset.

Table *2* provides statistical information about this public dataset. The data comes from real-world applications and reflects real-world events with a multi-target. ENB is another popular dataset being used by many studies. The data size is deemed appropriate with more than 300 samples [15]. The ENB dataset is interesting, with only two targets closely associated, while the features have no interdependency, making prediction more complicated.

Table 1: Descriptive Statistics for Desharnais Dataset

Descriptive Statistics	id	Proj	Team Exp	Mgr Exp	Year End	LEN	Effort	TRXN	Entities	Points Non Adjust	Adjust	Points Adjust	LANG
Valid	81	81	81	81	81	81	81	81	81	81	81	81	81
Missing	0	0	0	0	0	0	0	0	0	0	0	0	0
Mean	41.00	41.00	2.19	2.53	85.74	11.67	5046.31	182.12	122.33	304.46	27.63	289.23	1.56
Std. Deviation	23.53	23.53	1.42	1.64	1.22	7.43	4418.77	144.04	84.88	180.21	10.59	185.76	.71
IQR	40.00	40.00	3.00	3.00	2.00	8.00	3570.00	136.00	112.00	208.00	15.00	199.00	1.00
Minimum	1.00	1.00	-1.00	-1.00	82.00	1.00	546.00	9.00	7.00	73.00	5.00	62.00	1.00

Descriptive Statistics	id	Proj	Team Exp	Mgr Exp	Year End	LEN	Effort	TRXN	Entities	Points Non Adjust	Adjust	Points Adjust	LANG
Maximum	81.00	81.00	4.00	7.00	88.00	39.00	23940.00	886.00	387.00	1127.00	52.00	1116.00	3.00

Table 2: Descriptive Statistics for ENB Dataset

Descriptive Statistics	id	Relative compactness	X1	X3	X4	X5	X6	X7	X8	Y1	Y2
Valid	768	768	768	768	768	768	768	768	768	768	768
Missing	0	0	0	0	0	0	0	0	0	0	0
Mean	384.500	.764	671.708	318.500	176.604	5.250	3.500	.234	2.813	22.307	24.588
Std. Deviation	221.847	.106	88.086	43.626	45.166	1.751	1.119	.133	1.551	10.090	9.513
IQR	383.500	.147	134.750	49.000	79.625	3.500	1.500	.300	2.250	18.675	17.513
Minimum	1.000	.620	514.500	245.000	110.250	3.500	2.000	.000	.000	6.010	10.900
Maximum	768.000	.980	808.500	416.500	220.500	7.000	5.000	.400	5.000	43.100	48.030

Table 3: Descriptive Statistics for EVP Dataset

Descriptive Statistics	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Y1	Y2
Valid	8470	8470	8470	8470	8470	8470	8470	8470	8470	8470	8470	8470
Missing	0	0	0	0	0	0	0	0	0	0	0	0
Mean	.500	.053	.642	.633	.804	.791	3.057	1.162	.170	.013	1.002	.838
Std. Deviation	.006	.139	.276	.318	.272	.318	18.152	2.354	.316	.203	.205	.251
IQR	.000	.035	.446	.554	.297	.314	.926	.079	.375	.059	.042	.245
Minimum ($x10^{-3}$)	.500	141.9	3.000	34.55	8.000	7.000	460.0	99.00	-1524	-3953	.000	35.15
Maximum	1.000	1.000	1.611	3.976	3.774	4.757	1461.738	136.935	2.864	1.068	4.700	2.656

3) EVP Dataset

Earned Value Management (EVM) is widely acknowledged as the most reliable contemporary project management instrument or cost and timeline forecasting technique. EVM calculates the amount of work performed to measure project performance and progress. The Earned Value Plus dataset is based on the conventional EVM attributes and added two new attributes related to the project management and size indexes. It contains 8,470 (more than 8000 records) instances from more than 600 historical project data in EVM format was deemed sufficient to train the ML model effectively (Table 3).

3. Experimental Results

Each optimized model was tested in four cycles. Evaluation results were obtained through each testing cycle and tabulated for each performance indicator. Each performance metric was calculated based on the average performance. The following subsections describe how the ML model performed, illustrated by graphical presentation in two graphs. The first graph shows performance results in $RMSE$, MAE and $RMSLE$. The second graph shows the performance results in $MMRE$, $MdMRE$ and $Pred(0.25)$.

3.1. Desharnais Dataset

MLR-DNN was the most optimal model for predicting the probability of a given experiment, while PSO-DNN appeared as the worst. MLR-NNN had the highest $Pred(0.25)$ value and the best $RMSLE$ and MAE values among all models tested in this study (Figure 4 and Figure 5). MLR-DNN is a hybrid cascaded ML model comprising MLR (Multiple Linear Regressor) and cascading with DNN (Deep Neural Network) embedded with four hidden layers and 64 neurons in each hidden layer.

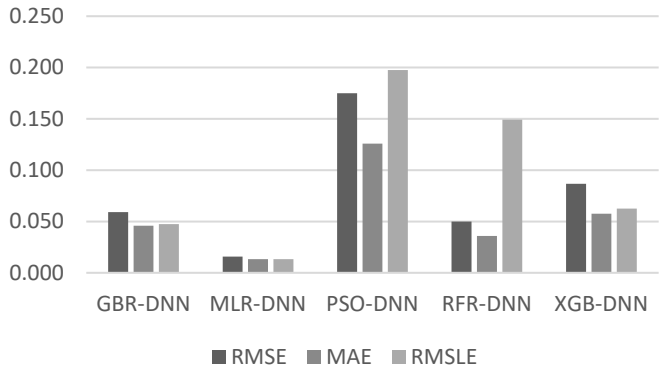


Figure 4: The $RMSE$, MAE, and $RMSLE$ results in the Desharnais dataset

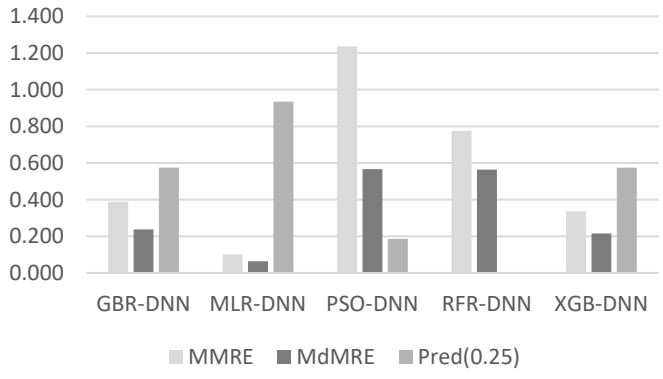


Figure 5: The $MMRE$, $MdMRE$, $Pred(25)$ results in the Desharnais dataset

3.2. ENB Dataset

MLR-DNN outplayed all other performance metrics, with the lowest $MMRE$ value being the least desirable model. The optimum $MdMRE$ value was .011, and the highest $Pred(0.25)$

value was .492, according to the most favourable $RMSE$ value. The most accurate MAE value was .004 (Figure 6 and Figure 7).

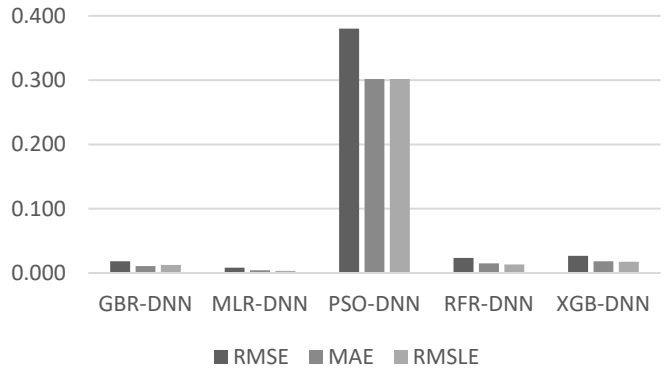


Figure 6: The $RMSE$, MAE, and $RMSLE$ results in the ENB dataset

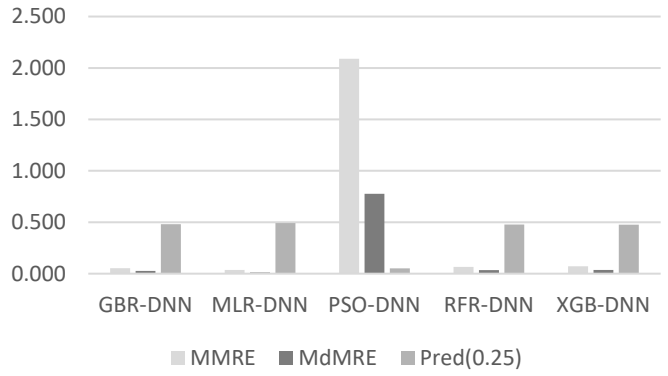


Figure 7: The $MMRE$, $MdMRE$, $Pred(25)$ results in the ENB dataset

3.3. EVP Dataset

MLR-DNN ranked as the top-performing ML model, with the lowest $MdMRE$ value and highest $Pred(0.25)$ value. The most favourable $RMSE$ value was .003, the best $RMSLE$ value was .003 and the most accurate MAE value of <.001 (Figure 8 and Figure 9).

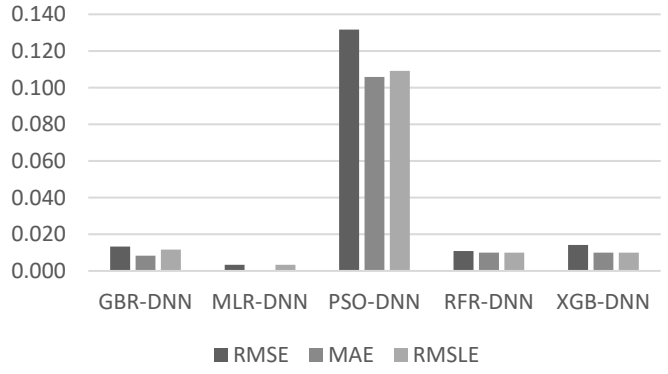


Figure 8: The $RMSE$, MAE, and $RMSLE$ results in the EVP dataset

3.4. Degree of Augmentation

The degree of augmentation, χ, is used as an error reduction indicator in a cascaded hybrid ML model using equation (12). The $MMRE$ for stage one (Δ_{t-1}) and stage two (Δ_t) enables us to calculate the degree of augmentation, χ, for each of the hybrids cascaded ML models (Figure 1). The hybrid model MLR-DNN demonstrated an average error reduction of .026 compared to the MLR model alone. PSO-DNN was excluded from the DOA comparison because PSD-DNN is not a cascaded standalone ML model but part of DNN with Particle Swarm Optimization (PSO) backpropagation. Overall results revealed that MLR-DNN outperformed all three other hybrids cascading DNN models, suggesting that cascading two different ML models may not produce positive results. Both GBR-DNN and XGB-DNN did not improve prediction accuracy, whereas the RFR-DNN model performed worse than RFR or DNN alone.

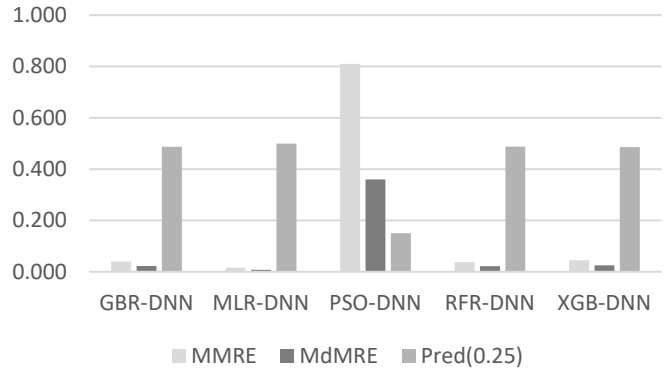


Figure 9: The $MMRE$, $MdMRE$, $Pred(25)$ results in the EVP dataset

Based on the performance results, the MLR-DNN model performed exceptionally well on all three datasets. The dependency on the quality of the dataset remains significant. This finding indicated that the PSO-DNN model was the most underwhelming performer in ENB and EVP datasets. However, for all three datasets, the least compelling performer was PSO-DNN. The runner-up position for both ENB and EVP datasets was GBR-DNN. However, the runner-up for the Dasharnais dataset was XGB-DNN.

The results also indicated that hybrid cascaded ML models such as GBR-DNN & XGB-DNN do not guarantee a positive gain and may sometimes have detrimental effects, for example, the RFR-DNN model. GBR-DNN performed relatively well in Desharnais and ENB datasets. However, it performed poorly in the EVP dataset. The result indicated that the quality of the dataset remains significant. This finding opens the door for future research.

The interquartile range (IQR) is a reliable measure of variability representing the dispersion of the middle 50% of the data [16]. The IQR is calculated as IQR = Q3 − Q1 statistically; the smaller IQR indicates the error range is relatively small. MLR-DNN showed the narrowest IQR and largest Mann-Whitney U effect size to strengthen its position as the most accurate ML model among the other models in this study. MLR-DNN enhanced the overall prediction accuracy compared to other models with a significant magnitude of error reduction.

From observation of the statistical value in

Table 4 for the degree of augmentation χ and Mann-Whitney U test effect size r, it seems like there is some form of proportion.

The investigators [17] explained that effect size is the difference between the variable's value in the control and test groups. The magnitude of χ increases and r increases, $|\chi| \propto r$. The significant difference between χ and r is that the effect size does not cater to attributes of positive or negative augmentation. This finding reflects that the degree of augmentation is a more appropriate performance indicator for measuring cascaded hybrid ML models.

Table 4: Degree of Augmentation Statistical Data

	GBR-DNN		MLR-DNN		RFR-DNN		XGB-DNN	
Descriptive statistics	Δ_{t-1}	Δ_t	Δ_{t-1}	Δ_t	Δ_{t-1}	Δ_t	Δ_{t-1}	Δ_t
Valid	11857	1779	11857	1779	11857	1779	11857	1779
Missing	0	0	0	0	0	0	0	0
Mean	.007	.006	.028	.002	.006	.009	.006	.006
Std. Deviation	.019	.018	.035	.003	.018	.011	.018	.014
IQR	.006	.006	.024	.001	.005	.007	.005	.005
Minimum (x10-6)	.105	8.492	12.13	1.059	.001	51.88	.002	.083
Maximum	.615	.115	.578	.061	.648	.188	.659	.397
p-value of Shapiro-Wilk		<.001		<.001		<.001		<.001
Degree of Augmentation χ		.001		.026		-.003		.000
Mann-Whitney U		9205868		809668		6171934.5		8978979
Wilcoxon W		79517879		2394758		76472087.5		79290990
(z) score		-8.701		-62.910		-28.286		-10.166
p-value		.000		.000		.000		.000
Effect Size r		.074		.538		.239		.061

4. Verification Results

Three types of live project data (Waterfall, Hybrid, and Agile) were used to verify MLR-DNN performance. The live performance results explained how effective MLR-DNN could be used practically in project management.

4.1. Waterfall Project

XYZ is one of the largest telecommunications operators in South East Asia. Due to exponential growth in customer demand, XYZ decided to enhance its operations support capability. MLR-DNN was used during the live project verification stage to forecast the budget and duration. Two EVM data samples were collected at 43% and 53% completion points. Table 5 displays the results.

MLR-DNN outperformed traditional EVM by 8.4% and 54.1% in average cost at Estimate At Completion (EAC) and average schedule prediction at Estimate Duration At Completion (EDAC), respectively. These findings align with a study which indicates CPI (cost) accuracy is relatively better than SPI (time) accuracy in EVM calculation [18].

Table 5: Waterfall Project Verification

% Complete	**Actual**		**ML Prediction**		***MRE***	
	EAC	EDAC	EAC	EDAC	EAC	EDAC
43%	.70	.67	.80	.65	.1	.02
53%	.70	.67	.74	.65	.04	.02
				MMRE	.07	.02

The MLR-DNN model improved and significantly enhanced the performance of project effort and duration estimation. Work Breakdown Structure (WBS) and EVM remain moderately accurate despite being less dependent on humans. The result indicated that the dataset's quality continues to have a significant impact, opening future research opportunities.

4.2. Hybrid Waterfall-Agile Project

Hybrid Agile-Waterfall projects combine agile approaches with waterfall methodologies to deliver projects. The waterfall method to record specific requirements and the agile methodology to deliver gradually in sprints are examples of hybrid projects. Another hybrid agile-waterfall model is software development teams adopting the agile methodology, while hardware implementation teams stick to the waterfall approach. The amount of agile versus waterfall project technique adoption in scope coverage determines the blending ratio.

STU is a major telecommunications operator in South East Asia with millions of customers. It would like to optimize and enhance its operations support and telemarketing capability. The project cost is moderately high: hardware, commercial out-of-shelf products, software customization, system integration, consulting, and professional services.

Table 6: Hybrid Waterfall-Agile Project Verification

% Complete	**Actual**		**ML Prediction**		***MRE***	
	EAC	EDAC	EAC	EDAC	EAC	EDAC
31%	.86	.96	1.23	.82	.37	.14
38%	.86	.96	.88	.73	.02	.23
54%	.86	.96	.84	.81	.02	.15
70%	.86	.96	.88	.74	.02	.22
92%	.86	.96	.75	.92	.11	.04
				MMRE	.11	.16

Five samples were collected from the same project at different stages and times (Table 6). One noticeable phenomenon is that prediction accuracy depends on the percentage of completion points. The closer the project's end, the more accurate the forecast is. At 31% completion, it was a less accurate prediction than the 54% completion point. The characteristic of EVM is inherited and aligned with findings in [10].

The predicted EDAC was accurate enough, with an average variance of 16% compared to any existing PM techniques and tools with 35-60%. There were insufficient details as to why there was a higher variance of EDAC than compared to EAC. Nevertheless, the project details revealed many change requests initiated that might impact prediction accuracy.

4.3. Agile Project

The MLR-DNN was fed with live agile project-scaled EVP data to predict project duration and cost in this verification test. Agile projects are typically shorter in duration and use fixed-length iterations. These projects usually have a low to medium budget, fixed period, and flexible scope.

ABC is a popular online banking software offering various electronic payment services to customers and financial institutions. A backlog of enhancements was prioritized in a different sprint by adopting a 100% agile methodology for the whole software development life cycle. Project resources were relatively small, usually less than ten people.

Project size was determined by the amount of project value in USD. Project is considered "small" < 500k; 1 million > "medium" ⩾ 500k, and "large" > 1 million. The percentage of completion was defined as the average project delivery progress

Table 7: Agile Project Verification

% Complete	Actual		ML Prediction		*MRE*	
	EAC	EDAC	EAC	EDAC	EAC	EDAC
100% (Sprint 1)	1	1	.99	1.00	.01	0
100% (Sprint 2)	1	1	.99	.99	.01	.01
50% (Sprint 3)	1	1	.77	.59	.23	.41
70% (Sprint 4)	.85	1	.93	.77	.08	.23
80% (Sprint 5)	.92	1	.94	.86	.02	.14
				MMRE	.07	.16

Three project-type live data samples were collected at different stages, iterations, sprints, and releases comprised of Agile, Hybrid, and Waterfall projects (Table 7). The overall prediction accuracy comparison between traditional EVM vs MLR-DNN in three project types is illustrated in Figure 10.

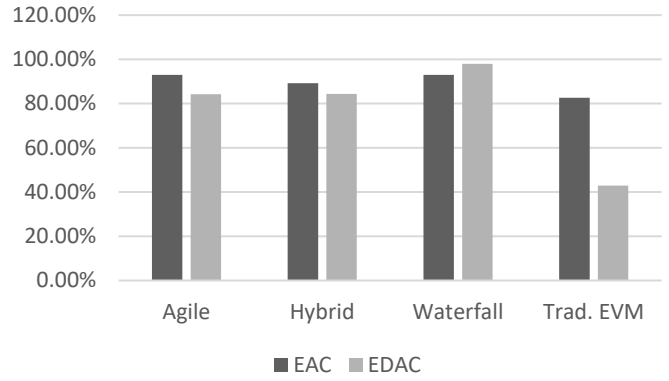


Figure 10: Performance Comparison between MLR-DNN and Traditional EVM in both Schedule and Cost Prediction

MLR-DNN model performed well in agile projects. It accurately predicted cost and schedule dimensions for many waterfall projects. Cost forecast accuracy is relatively better than duration forecast accuracy.

5. Machine Learning Biases

Machine learning (ML) algorithms are becoming more used in various industries. These algorithms, however, are not immune to bias, which can have detrimental repercussions. Therefore, it is critical to understand and address potential ML biases in order to ensure that these algorithms are fair and equal.

Type I - Algorithmic bias refers to systematic errors or unfairness resulting from employing algorithms inherited from the ML model, including how the model was constructed or trained, leading to biased outcomes [19]. Type II – Dataset bias is another type of bias that relates to the tendency of ML models to deliver inaccurate or unreliable predictions due to flaws or inconsistencies in the data used to train them [20]. It can result from various factors, including data collection methods and pre-processing techniques. To reduce ML biases, practitioners should evaluate models and datasets for performance and choose the least biased models.

6. Conclusion and Further Research

Traditional project planning in effort and duration estimation techniques remain low to medium accurate. This study seeks to develop a highly reliable and efficient Hybrid ML model that can improve cost and duration prediction accuracy. The results of the experiments indicated that MLR-DNN was the superior, effective, and reliable machine learning model.

The verification results in Agile, Hybrid and Waterfall projects indicated that the MLR-DNN model improved and significantly enhanced project effort performance and duration estimation. Despite WBS and EVM (conventional project management tools) being less dependent on humans, they are moderately accurate.

The results indicated that hybrid cascaded ML models such as GBR-DNN & XBG-DNN do not guarantee a positive gain and may sometimes have detrimental effects, for example, the RFR-DNN model. MLR-DNN inherits other neural network flaws being computationally costly and operating in black boxes with little explanation.

The accuracy of neural networks (including MLR-DNN) depends on the volume and the quality of training data [21]. Therefore, the dataset's quality significantly impacts the ML model's performance. This finding opens the door for future research.

References

[1] D.-J. Pang, K. Shavarebi, S. Ng, "Development of Machine Learning Models for Prediction of IT project Cost and Duration," in 2022 IEEE 12th Symposium on Computer Applications & Industrial Electronics (ISCAIE), IEEE: 228–232, 2022, doi:10.1109/ISCAIE54458.2022.9794529.

[2] D.-J. Pang, K. Shavarebi, S. Ng, "Project practitioner experience in risk ranking analysis-an empirical study in Malaysia and Singapore," Operations Research and Decisions, **32**(2), 2022, doi:10.37190/ord220208.

[3] D.-J. Pang, K. Shavarebi, S. Ng, "Project Risk Ranking Based on Principal Component Analysis - An Empirical Study in Malaysia-Singapore Context," International Journal of Innovative Computing, Information and Control, **18**(06), 1857–1870, 2022, doi:10.24507/IJICIC.18.06.1857.

[4] TD. Nguyen, T.M. Nguyen, T.H. Cao, "A conceptual framework for is project success," in Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, 142–154, 2017, doi:10.1007/978-3-319-56357-2_15.

[5] D. Magaña Martínez, J.C. Fernandez-Rodriguez, "Artificial Intelligence Applied to Project Success: A Literature Review," International Journal of

Interactive Multimedia and Artificial Intelligence, **3**(5), 77, 2015, doi:10.9781/ijimai.2015.3510.

[6] A. Mosavi, M. Salimi, S.F. Ardabili, T. Rabczuk, S. Shamshirband, A.R. Varkonyi-Koczy, "State of the art of machine learning models in energy systems, a systematic review," Mdpi.Com, **12**(7), 2019, doi:10.3390/en12071301.

[7] S. Bayram, S. Al-Jibouri, "Efficacy of Estimation Methods in Forecasting Building Projects' Costs," Journal of Construction Engineering and Management, **142**(11), 05016012, 2016, doi:10.1061/(ASCE)CO.1943-7862.0001183.

[8] D. Port, M. Korte, "Comparative studies of the model evaluation criterions MMRE and PRED in software cost estimation research," in ESEM'08: Proceedings of the 2008 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ACM Press, New York, New York, USA: 51–60, 2008, doi:10.1145/1414004.1414015.

[9] E. Korneva, H. Blockeel, "Towards Better Evaluation of Multi-target Regression Models," in Communications in Computer and Information Science, Springer Science and Business Media Deutschland GmbH: 353–362, 2020, doi:10.1007/978-3-030-65965-3_23.

[10] S. Picard, C. Chapdelaine, C. Cappi, L. Gardes, E. Jenn, B. Lefevre, T. Soumarmon, "Ensuring Dataset Quality for Machine Learning Certification," in Proceedings - 2020 IEEE 31st International Symposium on Software Reliability Engineering Workshops, ISSREW 2020, 275–282, 2020, doi:10.1109/ISSREW51248.2020.00085.

[11] A.K. Bardsiri, "An intelligent model to predict the development time and budget of software projects," International Journal of Nonlinear Analysis and Applications, **11**(2), 85–102, 2020, doi:10.22075/ijnaa.2020.4384.

[12] MF Bosu, SG Macdonell, "Experience: Quality benchmarking of datasets used in software effort estimation," Journal of Data and Information Quality, **11**(4), 1–26, 2019, doi:10.1145/3328746.

[13] R.M. Thomas, W. Bruin, P. Zhutovsky, G. Van Wingen, "Dealing with missing data, small sample sizes, and heterogeneity in machine learning studies of brain disorders," Machine Learning, 249–266, 2019, doi:10.1016/B978-0-12-815739-8.00014-6.

[14] OpenML enb, May 2021.

[15] M.A. Bujang, N. Sa'at, T.M. Ikhwan, T.A.B. Sidik, "Determination of Minimum Sample Size Requirement for Multiple Linear Regression and Analysis of Covariance Based on Experimental and Non-experimental Studies," Epidemiology Biostatistics and Public Health, **14**(3), e12117-1 to e12117-9, 2017, doi:10.2427/12117.

[16] D.T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, 2005, doi:10.1002/0471687545.

[17] P. Kadam, S. Bhalerao, "Sample size calculation," International Journal of Ayurveda Research, **1**(1), 55, 2010, doi:10.4103/0974-7788.59946.

[18] M. Fasanghari, S.H. Iranmanesh, M.S. Amalnick, "Predicting the success of projects using evolutionary hybrid fuzzy neural network method in early stages," Journal of Multiple-Valued Logic and Soft Computing, **25**(2–3), 291–321, 2015.

[19] S.S. Gervasi, I.Y. Chen, A. Smith-Mclallen, D. Sontag, Z. Obermeyer, M. Vennera, R. Chawla, "The Potential For Bias In Machine Learning And Opportunities For Health Insurers To Address It," Https://Doi.Org/10.1377/Hlthaff.2021.01287, **41**(2), 212–218, 2022, doi:10.1377/HLTHAFF.2021.01287.

[20] A. Paullada, I.D. Raji, E.M. Bender, E. Denton, A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," Patterns, **2**(11), 100336, 2021, doi:10.1016/J.PATTER.2021.100336.

[21] J. Zhou, X. Li, H.S. Mitri, "Classification of rockburst in underground projects: Comparison of ten supervised learning methods," Journal of Computing in Civil Engineering, **30**(5), 04016003, 2016, doi:10.1061/(ASCE)CP.1943-5487.0000553.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 8, No. 2, 116-122 (2023)
www.astesj.com
Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTES Journal
ISSN: 2415-6698

Hybrid Discriminant Neural Networks for Performance Job Prediction

Temsamani Khallouk Yassine [*,1], Achchab Said [1], Laouami Lamia[2] , Faridi Mohammed[2]

[1] *University Mohammed V, ENSIAS, Rabat, Morocco*

[2] *University Hassan 1, ENCG, Settat, Morocco*

ARTICLE INFO

Article history:
Received: 28 September, 2022
Accepted: 16 March, 2023
Online: 11 April, 2023

Keywords:
Artificial Neural Network
Performance job prediction
Particle swarm optimization
Human talent
Variable selection

ABSTRACT

Determining the best candidates for a certain job rapidly has been one of the most interesting subjects for recruiters and companies due to high costs and times that takes the process. The accuracy of the models, particularly, is heavily influenced by the discriminant variables that are chosen for predicting the candidates scores. This study aims to develop an performance job prediction systems based on hybrid neural network and particle swarm optimisation which can improve recruitment screening by analyzing historical performances and conditions of employees. The system is built in four stages: data collection, data preprocessing, model building and optimisation and finally model evaluation. Additionally, we highlight the significance of Particle Swarm Optimization (PSO) in enhancing the performance of the models created by presenting a training algorithm that uses PSO. We conduct a study to compare the performance of each hybrid model and summarize the results.

1 Introduction

The field of human resources (HR) has undergone significant changes over the past few decades, and the rise of artificial intelligence (AI) has had a major impact on how HR functions are performed. From recruitment and employee evaluation to training and career development, the introduction of AI has led to a transformation in the way HR professionals perform their duties. This article will delve into the literature on the impact of AI in HR, analyzing the advantages and drawbacks of utilizing AI in this field, and examining the potential implications for HR professionals and organizations.

Another area where AI is having an impact is in employee evaluation and performance management. AI algorithms can analyze an employee's work history, skills, and achievements to predict their potential for growth and future success within the company. This can help HR professionals make informed decisions about employee development and career advancement. For example, in a study by Deloitte, 92% of HR professionals reported that AI has improved the accuracy of performance evaluations (Deloitte, 2019). Performance job prediction has become increasingly popular with the advent of machine learning algorithms such as decision trees, random forests, and support vector machines, which can handle vast quantities of data and discern intricate connections between various variables. These algorithms can be trained on historical performance data to make predictions about the future performance of new hires or current employees. In this study, we will focus on the application of the ANN on the candidates performance prediction [1].

Designing an ANN with the appropriate parameters can result in a powerful tool. In fact, the process of choosing selecting the architecture that will works well includes the number of input, hidden neurons and weight values, for a complex situations can pose an optimization task challenge. The training process plays a crucial role in determining the ANN topology. Firstly, the most suitable architecture needs to be chosen by assessing the problem at hand, which entails identifying the number of input, hidden, and output neurons. Secondly, the ideal weight values that enable the ANN model to perform at its best must be identified. While the ANN architecture is typically determined by experience, some researchers have started using meta-heuristic algorithms such as Particle Swarm Optimization to explore various possible architectures and select the optimal one based on a fitness criterion.

The primary objective of this study is to create a job performance prediction model that utilizes ANN, PSO, and appropriately selected variables based on the availability of data. The first step involves examining the effectiveness of variable selection models by comparing discriminant analysis and logistic regression techniques. The second step entails identifying the optimal ANN topology by proposing a training process that employs the PSO algorithm to determine the ideal neural network topology.

[*]Corresponding Author: TEMSAMANI Khallouk Yassine, Rabat, Temsamani.khallouk.yassine@gmail.com

2 Literature review

2.1 Performance job prediction

Performance job prediction is the process of using various data points, such as an employee's job performance, education and skill sets, personality traits, and even social media activity, to make predictions about an individual's potential for growth and success within a company [2]. This information can be used by organizations to make informed decisions about employee evaluation, promotion, and training. The goal of performance job prediction is to create a more productive and efficient workforce by identifying high-potential employees and providing them with the resources and support they need to succeed [3].

One of the key advantages of performance job prediction is its ability to provide actionable insights into employee performance. For example, organizations can use these predictions to identify high-performing employees and provide them with the resources and training they need to excel in their roles. In addition, predictions can help managers make informed decisions about promotions, pay increases, and other compensation-related matters.

The accuracy of performance job prediction models is dependent on the quality and quantity of data used in the model. Data sources may include things like past performance evaluations, training data, and demographic information. The use of multiple data sources allows organizations to build a more complete picture of an employee's potential, providing a more accurate prediction.

One of the most commonly used approaches for performance job prediction is regression analysis. This method uses statistical methods to model the relationship between predictor variables (e.g., past performance, training data) and the dependent variable (future performance). Regression analysis can provide valuable insights into the impact of different factors on an employee's performance, allowing organizations to make informed decisions about staffing, development, and compensation.

An alternative method involves utilizing machine learning algorithms, such as decision trees, random forests, and gradient boosting, to construct models that predict employee performance by analyzing data [4]–[7].

2.2 Classification for Prediction

Smart choices can be made by utilizing techniques such as classification and prediction. Scholars in the domain of machine learning have suggested numerous methods for classification and prediction assignments. This research, in particular, focuses on the classification methods employed in the machine learning procedure. These approaches to analyzing data are used to derive models that define significant data categories or anticipate forthcoming trends in the data [8].

The classification process is composed of two main phases: during the learning phase, the classification algorithm scrutinizes the training data to generate a classifier, which is essentially a set of guidelines for classification. In the classification phase, the accuracy of the classifier is evaluated by testing it on the test data. If the accuracy is satisfactory, the model can be utilized to make predictions on new data. There are several techniques for classification, such as Bayesian approaches, decision trees, neural networks, and numerous others.

This study will focus on the application of the artificial neural network and its parameter optimization [9].

2.3 Artificial Neural Network

Artificial neural networks, a subcategory of artificial intelligence, draw inspiration from neurobiology and entail designing machines capable of learning and accomplishing specific assignments, such as classification, prediction, or grouping. These networks comprise interlinked neurons that learn from the data they encounter to detect linear and nonlinear patterns in intricate data, resulting in dependable forecasts for new scenarios. The inaugural neuron model, which was grounded on biological neurons, was introduced in 1943 by McCulloch and Pitts.

In 1943, McCulloch and Pitts introduced the first neuron model, which proved that formal neurons are capable of performing logical functions. Later, in 1949, psychologist Donald Hebb introduced parallel and connected neural network models and proposed many rules for updating weights, including the well-known Hebbian rule [10]. Frank Rosenblatt, a psychologist, created the perceptron model in 1958. This model was able to identify simple shapes and carry out logical functions [11]. Nevertheless, in 1969, Minsky and Papert revealed the limitations of the perceptron, specifically in addressing nonlinear problems [11]. In the 1980s, interest in artificial neural networks was renewed by the introduction of Rumelhart's Back-Propagation algorithm, which enhances parameters for multilayered neural networks by transmitting errors to the hidden layers [12].

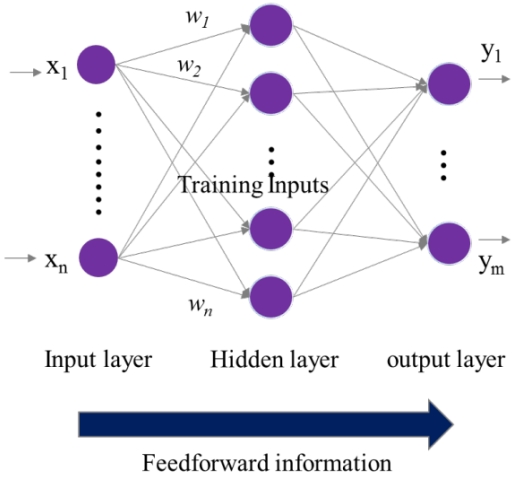


Figure 1: Artificial Neural Network

Since then, the use and application of neural networks have expanded into various fields. In fact, studies have demonstrated that a Multilayer Perceptron network with a single hidden layer has the capability to estimate any function of Rn in Rm with high accuracy [13]. The structure and settings of a neural network are crucial factors that determine its effectiveness and efficiency.The multilayer network is a commonly used structure in neural networks, consisting of an input layer, one or more hidden layers, and an output layer. The hidden layer functions as a mediator between the

input and output neurons, and the connections among the layers are represented by the weights of the connecting links. Generally, the layers in a neural network are linked together in a manner that allows data to flow solely in one direction - this is known as a feed-forward approach. There are no loops or cycles in the network. The data originates from the input layer, traverses through the hidden layers, and finally arrives at the output layer. An example of a neural network with a single hidden layer is shown in Figure 1 to provide a better understanding of its functioning. In addition, previous research has demonstrated that this architecture is optimal for solving classification problems, which is the focus of this article [13].

For a neural network with one hidden layer, we can observe that each hidden neuron (indexed $j = 1, ..., n$) takes in an input that is the result of a weighted sum of the inputs to the entire network. The transfer function f is used to process the input and convert it into an output signal.

$$z_j = f\left(w_{j0} + \sum_{i=1}^{n} w_{ji} x_i\right) \tag{1}$$

The variable n and and the variable m are the number of input neurons and the hidden ones, respectively and w_{ij} is the weight from the i_{th} input neuron to the j_{th} hidden neuron, x_i is input variable i and w_{j0} is a bias term. The hidden neuron signals are subsequently transmitted to the output neurons through weighted connections, similar to the transmission between the input and hidden layers. Consequently, the output neurons obtain the sum of all weighted hidden neurons, which is then passed through a transfer function g, based on the required output range. The output y_o of the network's output neuron o is formulated as:

$$y_o = g\left(b_{z0} + \sum_{i=1}^{m} \beta_{zj}\left(f\left(w_{j0} + \sum_{i=1}^{n} w_{ji} x_i\right)\right)\right) \tag{2}$$

With b_{zj} represent the weight from the J_{TH} hidden neuron to the O_{th} output neuron and b_{z0} is the bias. As mentioned earlier, the reason why neural networks are popular in different areas is due to their capacity to approximate linear or nonlinear functions. However, the challenge is to determine the optimal topology and weight values of the network that can closely approximate the target function. This task can be thought of as an optimization problem, where the objective is usually to minimize a cost function based on the total sum of squared errors.

2.4 Optimizing parameters for ANN

2.4.1 Input variables

Once the artificial neural network is established, the next step is to identify the necessary information required to build the network. This information is provided in the form of input variables that are used to assess the potential job performance of candidates. To permit to the ANN to accurately classify new observations, the input variables must be carefully selected to ensure the classification model performs well. Therefore, it is crucial to identify the most relevant variables for classification purposes.

2.4.2 Architecture

Tthe structure of an ANN is an input layer, output layer, and one or more hidden layers. Hence, there are other crucial factors that have an impact on the performance of the artificial neural network, and they need to be considered while designing it. These factors comprise the number of neurons present in each layer and the number of hidden layers that are incorporated into the network. These parameters determine the behavior of the neural network and vary depending on the problem to be solved.

The study employs the neural network architecture with one hidden layer for classification purposes [13] , which is widely acknowledged as the optimal structure for such problems according to existing literature.

Selecting the appropriate number of neurons for the hidden layers of an artificial neural network can be a difficult task. Having too many neurons can lead to an increase in the number of computations required by the algorithm. Conversely, selecting too few neurons in the hidden layer can result in a reduction in the model's capacity to learn [14]. So, it is crucial to choose the optimal number of neurons to achieve the highest possible performance of the neural network.

2.4.3 Learning algorithm

The process of finding the optimal weights and biases that maximize the performance of a neural network is known as learning algorithms, which consist of a set of rules. Various techniques have been used in literature to determine the best architectures and topology of weights and biases for the neural network, depending on the learning type.

Supervised learning refers to the scenario where the dataset used for training is labeled, while unsupervised learning train on unlabeled datasets. In unsupervised learning, the weights of the neural network are adjusted based on specific criteria to identify patterns or regularities in the observations.

The principal goal of this research is to improve the performance of an artificial neural network (ANN) in predicting job performance of a candidate. This is done using a supervised learning approach where the labels for the classes are already known and provided during the training stage. The learning algorithm adjusts the connection weights between inputs layers and the target ones to estimate their dependencies and minimize the error function, such as mean squared error.

The optimisation techniques can be classified into two groups:

• The first set of techniques is based on the steepest descent method and includes methods like gradient descent, Levenberg Marquardt, Backpropagation, and their variations. However, some of these algorithms require a significant amount of computational resources in terms of time and memory. Out of these, the Backpropagation algorithm is the most widely utilized, as it is a highly effective tool for determining the gradient in neural networks. However, it has its limitations, particularly with regards to the issue of getting stuck in local minima.

• The second group encompasses techniques that are inspired by the evolution of living species, such as genetic algorithms and swarm algorithms among others.

2.4.4 Transfer Function

Before training a neural network, one of the parameters that needs to be determined is the transfer function. The selection of an activation function is dependent on the specific use case. For instance, binary functions are well-suited for organization and distribution problems, whereas continuous and differentiable functions like sigmoid function are utilized to approach continuous functions. Notably, the sigmoid transfer function is commonly used because it combines nearly linear, curvilinear, and nearly constant behavior based on the input value [14]. The sigmoid transfer function's adaptability enables the artificial neural network to manage both linear and non-linear issues. It's possible to represent the sigmoid function as:

$$f(x) = \frac{1}{(1 + \exp(-x))} \tag{3}$$

The function being used as the transfer function in this study is bounded between zero and one, and it takes a real-valued input and produces an output within that range.

2.5 Particle Swarm Optimization (PSO)

PSO is an evolutionary computation method that was created by J. Kennedy, a social psychologist, and R. Eberhart, an electrical engineer, in 1995 [15]. It is a type of swarm intelligence algorithm that draws inspiration from the natural behavior of social organisms, such as birds flocking, and is employed as an optimization technique in a variety of research domains.

Social animals that live in groups, like swarms, often need to travel long distances to migrate or search for food. To do so efficiently, they optimize their movements in terms of time and energy expenditure and cooperate with one another to achieve their objective. The PSO algorithm is rooted in this behavior and is utilized to discover solutions to problems by optimizing a continuous function in a data space. Each member of the group, similar to the animals in a swarm, decides their movement based on their own experience and that of their peers, resulting in a complicated and effective process [15].

The PSO algorithm is designed around a group of individuals known as particles. At the first time, these particles are placed randomly in the solution space and move around in search of the optimal remedy to the challenge. Each particle's position represents a potential solution to the challenge. The movement of each piece is governed by specific rules. Each particle has a memory that allows it to remember the best point it has encountered so far and tends to return to that point. Additionally, each particle is informed of the best point found by its neighbors and tends to move towards that point.

The initial step for utilizing the PSO algorithm involves establishing a search area comprising of particles and a fitness function for optimization. Afterward, we commence by initializing the system with a set of haphazard solutions (particles). Each particle is allotted a positional value signifying a plausible solution data, a velocity value that denotes the extent to which the data can be modified, and a personal best value (pBest) that represents the particle's most optimal solution reached thus far.

Algorithm 1: PSO algorithm[16]

```
for Particle i in swarm S do
    Set up the particle i ;
end
while stopping condition is false do
    for Individual i of the swarm do
        Calculate the fitness f(xi(t));
        if the fitness value>p_best then
            Assign the current value as the updated personal
              best (p_best);
        end
    end
    Select the optimal fitness value among all particles and
      denote it as (g_best);
    for Individual i of the swarm do
        Adjust the velocity of the particle in accordance
          with the Eq (1);
        Adjust the velocity of the particle in accordance
          with the Eq (2);
    end
end
```

3 Research Framework

In order to define the solutions, a research framework must first be developed. The research includes four major processes that are included in this research: data collection, preprocessing of data, variable selection, model building and the evaluation of the model.

The proposed framework is shown in the Figure 2 below.

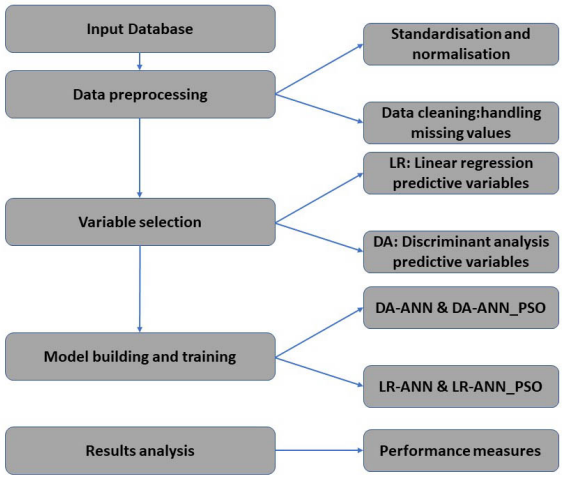


Figure 2: The proposed methodology

3.1 Discriminant variables

To determine the optimal architecture for an ANN, the first step is to identify the input variables. These variables will be used to construct mathematical models that predict job performance. However, selecting the appropriate variables can be a challenge and is crucial

for the model's accuracy. In this study, 15 variables listed in Table 2 were used, based on the availability of data.

3.2 Data preprocessing

Studies and research have demonstrated that several AI algorithms may exhibit poor performance due to the inferior quality of data and variables. Therefore, two critical steps are required to enhance the feasibility of the variables for constructing a predictive model: variable selection modeling for reducing dimensionality and data preprocessing. This process involves data preparation and normalization to accomplish reduction or classification tasks. The Table 1 below shows the variables used in this study.

Table 1: Dataset description

Variables	Description	Value
ID	Employee's id	integer
Age	Employee's Age	Integer
Gender	Employee's gender	M or F
Marital status	Employee's status	S or M
Diploma	Employee Education Degree	Bachelor High Diploma Master, Phd
Experience years	Employee year of experience	Integer
Salary	Employee salary	Integer
Communication Level	Employee level in communication	1 to 5
Motivation enthusiasm	Employee motivation for work	Yes or No
Language score	Employee language level	1 to 5
Specialisation	Employee general Specialisation	IT, Economics HR, Network business
Effectiveness in a remote environment	Employee ability in remote	Yes or no
Seniority	Employee seniority in the company	Junior Senior Manager
Physical abilities	Employee ability to work	Yes or no
Additional Certificate	Employee additional certificate	Yes or no
Employee performance	Employee performance	BA Good

3.3 Variables selection models

In classification studies, it is crucial to determine which variables hold the most importance in distinguishing between different categories. Moreover, it is often challenging to obtain trustworthy and meaningful data. Therefore, it is essential to identify the most significant variables that can offer insights to forecast candidate performance to reduce the effort required to gather and verify data.

When creating a prediction model, it can be helpful to reduce the number of variables in order to improve computational efficiency and increase the accuracy of classification algorithms, like neural networks. To achieve this, we'll use two types of classification techniques statistical methods and artificial intelligence in order to identify the most important variables that distinguish between candidates' performance. Then, we'll choose the best variable selection model to optimize the performance of the neural network.

In this study, we are more focused on the statistical method Statistical method used in this study is chosen for its popularity in variable selection is discriminant analysis (DA) and logistic regression. Discriminant analysis is commonly utilized to identify a linear combination of features that can effectively distinguish between two or more groups, in order to reduce the number of dimensions prior to classification.

3.4 Artificial Neural Network Architecture

This study utilizes an ANN model for predicting the job performance of candidates chosen at random. As previously stated, the architecture of the ANN is critical to its functionality and effectiveness. Therefore, this section focuses on determining the optimal topology that can differentiate between a good candidate and a poor one based on the selected variables.

To define the architecture of an ANN, certain parameters must be determined such as the number of input neurons, hidden layers, and hidden neurons. According to the literature, ANNs with one hidden layer are considered the optimal structure for classification problems[13].

4 Cross validation

Any bias or bad quality due to dataset could potentially have a huge impact on determining the artificial neural network and its parameters. In this sense, the cross validation technique is made to minimize this genre of problem.

In our experiment we will use a 3 fold-cross validation technique to train and test our model to avoid over fitting.

To be precise, we divided our dataset into three equal subsets, which implies that our model will undergo training and testing procedures three times. The mean value of the accuracy measures obtained from each of the three iterations is used to evaluate the overall accuracy of the model.

5 Performance evaluation

To evaluate the performance of our model, we use this list of evaluation metrics:

Overall accuracy: In general, accuracy refers to the percentage of correctly classified records by the model. The formula for calculating accuracy can be derived from the confusion matrix presented in Table 2.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

Precision: It can be described as the proportion of the correctly predicted cases (True Positive) to the combined number of True Positive and False Positive.

Recall: It can be expressed as the proportion of the correctly predicted cases (True Positive) to the combined number of True Positive and False Negative.

Specificity: The True Negative Rate is calculated as the number of True Negatives divided by the sum of True Negatives and False Positives.

Predicted		
Actual	**BA**	**Good**
BA	True Negative	False Positive
Good	False Negative	True Positive

F-Measure: F-measures take the harmonic mean of the Precision and Recall Performance measures [17].

$$F\_\text{Measure} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

6 Empirical study

As mentioned earlier, the main aim of this study is to employ a combination of neural network and Particle Swarm Optimization to forecast the job performance of a random applicant. The initial stage in this approach, as described in the research methodology, involves choosing the suitable variables that can be utilized to create a the optimal method.

6.1 Data and variables

The dataset used in this study issued from a Moroccan firm contains the most variables used on the manual recruitment process based on the survey made inside each department of this firm. Data collected contains more than 1000 individuals. A variable class is created with two values (BA if the candidate is below the average, Good is the candidate have a good qualification). The individuals was selected randomly from a different department: IT department, finance department, HR department, Data department...

Before using our Data as input for our model, a normalizing function Eq.6 was applied to bound data values to -1 and +1 with X is the input matrix, Y is the normalized matrix, x_{min} and x_{max} are respectively the maximal and the minimum values of a variable [18].

$$Y = \frac{0.9 - 0.1}{\text{xmax} - \text{xmin}} X + \left(0.9 - \frac{0.9 - 0.1}{\text{xmax} - \text{xmin}} \text{xmax}\right) \quad (6)$$

6.2 Results

The initial step of processing and managing data involves dealing with missing values, decreasing the number of variables, and examining the most significant ones, which is crucial. So, we begin our process of building a performance job prediction model by handling the missing values. Our dataset contains many missing values so to fix this problem, we refer to KNN imputation. In fact, a new observation is imputed by finding the samples in the training set closest to it and averages these nearby points to fill in the value.[4]

Secondly, we need to determine the influence of each variables on a candidate's job performance by using variable selection techniques. We will compare common models such as Discriminant Analysis and Logistic Regression, and summarize the variables selected by each model in a table below.

Table 2: Variables selection results

Variables selection Techniques	**Number variables**	**Selected Variables**
DA	8	Gender, Marital status, Seniority , Salary Communication level Employee ethics Specialisation, Physical abilities
Logistic Regression	12	Age, Gender, Marital status, Seniority , Salary, Diploma Experience years Language score Communication score Specialisation, Additional Certificate, Effectiveness in a remote environment

The presented table displays how each model has selected a distinct set of variables based on their discriminatory power. The feature sets have been divided into two categories: the first group contains eight variables chosen by DA, and the second group includes twelve variables chosen by LR.

The ANN model's input layer will rely on the set of variables selected by the variable selection models. As a result, two hybrid neural network models, MDA-ANN and LR-ANN, are constructed accordingly.

After defining the best variables that will have a big impact on our target variable, it's time now to define the best architecture for our model, for this reason we compare the following learning algorithm based on PSO, and the hybrid artificial neural network trained separately.

Now, we have reached the step of designing the topology of our hybrid neural network. for this step, we used the following parameters:

The architecture that produced the highest performance accuracy applied to our model was determined to be 12-18-1 (12 input neurons, 18 hidden neurons, and one output neuron). The 12 input neurons in this case correspond to the number of variables selected by the logistic regression algorithm, indicating that these variables

have strong discriminatory power when it comes to predicting candidate performance.

Table 3: PSO parameters

Parameters	Architecture optimization	Weights optimization
Swarm Size	20	20
Stop criteria & iteration	100	100
Search area range	[3, 20]	[-2.0, 2.0]
Inertia factors	$(w_n = 0.9 * w_{n-1})$ $w_0 = 0.8$	$(w_n = 0.9 * w_{n-1})$ $w_0 = 0.8$

We can see also that the application of our Hybrid artificial neural network separately decrease the performance of the two models DA-ANN and LR-ANN compared to its application with the PSO. The results will be presented and analyzed in the table 5.

Note that the evaluation of the evaluation metrics alone does not give a good judgment on the quality of the prediction and the classification. In this performance comparison, we will also focus on the performance attribute to each class which gives important information about a model especially to select the variables which discriminate the performance of the candidates.

This appears clearly in the application of the hybrid algorithm: DA-ANN and DA-ANN_PSO. In fact, even with its big accuracy, they present the less rate of good classification of good candidates (47.5%, 48.3%) contrary to below average candidates (between 83.3% and 83.4%). The LR-ANN_PSO model gives the best classification rate. These findings suggest that the variables identified by the LR statistical models provide more insights into a candidate's job performance.

Table 4: Results

Model	LR-ANN	LR-ANN _PSO	DA-ANN	DA-ANN
Accuracy	72.5%	75.0%	65.0%	65.4%
Precision	70.1%	72.9%	47.5%	48.3%
Sensitivity	73.1%	75.6%	74.9%	75.1%
Specificity	72.0%	74.5%	60.3%	60.6%
F-measure	71.6%	74.2%	58.1%	58.8%
BA	74.9%	77.0%	83.4%	83,3%
Good	70.1%	72.9%	47.5%	48.3%

7 Conclusion

In this research, we have implemented a hybrid discriminant neural network relying on particle swarm optimisation and statistical variables selection techniques. The models developed takes into account the variables mostly used in the manuel performance job prediction, otherwise, the constraints of missing values was fixed by the K-nearest neighbor algorithm.

The proposed methodology of variables selection evaluated the impact of different variables selection models by comparing Multivariate Discriminant Analysis and Logistic Regression. The findings demonstrate that logistic regression perform exceptionally well as a variables selection model for Artificial Neural Networks (ANN) to distinguish between candidates job performance. Moreover, the application of the variables chosen by this technique gives the best performance for the task of prediction the candidate job performance prediction.

The hybrid neural network applied with the learning algorithm PSO gives the best results in term of optimisation and finding the local minima and then in the prediction of the job performance. This model will be very useful for recruiter to assess and predict the performance of future candidates.

References

[1] S. S. A. Mohan, Support Vector Machines for Job Performance Prediction: A Comparative Study, Ph.D. thesis, 2021.

[2] J. Zhang, Y. Liu, The Use of Big Data Analytics in Job Performance Prediction: A Literature Review, Ph.D. thesis, 2020.

[3] S. Kaur, M. Singh., "A Review of Machine Learning Algorithms for Job Performance Prediction," 2019.

[4] J. Delaney, The rise of predictive employee analytics, Ph.D. thesis, 2019.

[5] S. Krishnan, Predictive employee analytics: A new frontier in HR. Forbes., Ph.D. thesis, 2020.

[6] D. L. . C. J.Russell, Employee analytics: How to improve business performance by measuring and managing your workforce., Ph.D. thesis, 2015.

[7] H. . Y.Zhang, A review of predictive analytics in human resources management., Ph.D. thesis, 2018.

[8] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Ph.D. thesis, 2006.

[9] K. K. Y. Geoffrey K.F. Tso, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," Energy, **32**, 1761–1768, 2007, doi:doi.org/10.1016/j.energy.2006.11.010.

[10] E. D. et P. Naïm, Des réseaux de Neurones, EYROLLES, 1992.

[11] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." Journal of Applied Mathematics and Physics, **5**, 1958, doi:doi.org/10.1037/h0042519.

[12] J. L. M. David E. Rumelhart, "Parallel distributed processing: explorations in the microstructure of cognition, vol. 1 : foundations," The MIT Press, **9**, 386–408, 1987.

[13] Y.-C. L. Jae H. Min a, "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters," Expert Systems with Applications, **28**, 603–614, 2005, doi:doi.org/10.1016/j.eswa.2004.12.008.

[14] D. T. L. et C. D. Laros, Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition, WILEY, 2014.

[15] J. K. R. Eberhart, "A new optimizer using particle swarm theory," MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, doi:10.1109/MHS.1995.494215.

[16] S. A. Fatima Zahra Azayite, "Topology design of bankruptcy prediction neural networks using Particle swarm optimization and backpropagation," 1–6, 2018, doi:https://doi.org/10.1145/3230905.3230951.

[17] "Archives ourvertes," 2018.

[18] "cyberleninka," 2018.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 1, 116-124 (2023)

www.astesj.com

Special Issue on Innovation in Computing, Engineering Science & Technology

ASTESJ

ISSN: 2415-6698

Metaheuristic Optimization Algorithm Performance Comparison for Optimal Allocation of Static Synchronous Compensator

Abdulrasaq Jimoh[1], Samson Oladayo Ayanlade[*,2], Emmanuel Idowu Ogunwole[3], Dolapo Eniola Owolabi[4], Abdulsamad Bolakale Jimoh[5], Fatina Mosunmola Aremu[6]

[1]*Obafemi Awolowo University, Department of Electronic and Electrical Engineering, Ile-Ife, Nigeria*

[2]*Lead City University, Department of Electrical and Electronic Engineering, Ibadan, Nigeria*

[3]*Cape Peninsula University of Technology, Department of Electrical, Electronic and Computer Engineering, Cape Town, South Africa*

[4]*Ladoke Akintola University of Technology, Department of Electronic and Electrical Engineering, Ogbomoso, Nigeria*

[5]*University of Ilorin, Department of Electrical and Electronic Engineering, Ilorin, Nigeria*

[6]*Kwara State University, Department of Electrical Electronic Engineering, Malete, Nigeria*

ARTICLE INFO

Article history:
Received: 25 December, 2022
Accepted: 25 January, 2023
Online: 07 February, 2023

Keywords:
FACTS devices
STATCOM
Power loss
Voltage profile
Particle swarm optimization
Firefly algorithm

ABSTRACT

The relevance of static synchronous compensator (STATCOM) controllers in controlling power network parameters is causing them to be included in contemporary networks. But for the intended objectives to be attained, the best device positioning and parameter settings are essential. This work compares the performance of the particle swarm optimization (PSO) and firefly algorithm (FA) in sizing and placing a STATCOM device for the dual objectives of loss reduction and voltage deviation abasement. The effective mitigation of network loss and voltage fluctuations in the network will be achieved by the deployment of the efficient method during device allocation. While PSO and FA were taken into consideration due to their computational efficiency among other metaheuristic algorithms, STATCOM was chosen from among the Flexible Alternating Current Transmission System (FACTS) controllers as a consequence of its reactive power compensation capability. The MATLAB software was used to implement the simulations on an IEEE 14-bus system. When STATCOM was optimized with PSO and FA, it resulted in active power loss reductions of 432 and 733 kW, respectively, and reactive power loss reductions of 1622 and 2100 kVAr, respectively. As a result, the reductions in voltage variation and power losses in this instance show some benefits of FA over PSO. Additionally, this work has shown that metaheuristic algorithms are beneficial for allocating FACTS devices.

1. Introduction

In current use, a power system is a system made up of a large number of power plants, transmission lines, loads, and transformers [1–2]. Increased power consumption causes transmission lines to become overloaded, which makes the power systems unstable. The system must thus operate very near its stability limit. This typically leads to a poor voltage profile and considerable network power loss [3–4].

The deployment of Flexible Alternating Current Transmission System (FACTS) devices and the building of new transmission lines are two options for addressing the problem of the power system overloading [5]. The construction of new power generation and the upgrading of transmission lines to reduce line congestion are both fraught with challenges. Increased load demands, constraints on the economy and the environment, and power networks operating nearer to their stability limits are all implications of the reorganization of the electrical sector [6]. For the aforementioned reasons, the power networks frequently encounter losses and voltage instability,

[*]Corresonding Author: Samson Oladayo Ayanlade, Lead City University, +2348062786683, samson.ayanlade@lcu.edu.ng

which can result in voltage collapse. Sustaining the system's stability and safety is therefore a crucial and challenging problem.

To improve system stability and security, several strategies, including reactive power compensation (RPC) and phase shifting, are used [7]. The RPC is the strategy that is most frequently employed and well-liked among them since power networks are mostly reactive. Reactive power is required to maintain voltage magnitudes for transmitting active power across transmission lines. The primary source of power losses is the use of reactive power above the threshold set by the generators. By utilizing compensators, power losses may be reduced to a minimum. Different types of RPCs are employed in power networks to compensate for reactive power [8].

Power electronics and FACTS device advancements have made it possible to manage line flows, reduce overall system loss, and keep the voltage profile within permissible bounds in a power system [9]. FACTS are regulators that may alter several features of a transmission network. Through system parameter management, they also possess the capacity to swiftly and seamlessly consume or provide reactive power to the networks. These allow for voltage control on a specific bus.

Different categories have been established for FACTS devices. The work by [10] demonstrated the modeling of the FACTS device and its integration into power flow investigations. The position of the STATCOM, a shunt-type FACTS regulator, in the grid significantly affects losses and voltages and is primarily employed by power engineers for reactive power adjustment. The objective of STATCOM placement, an optimization issue, is to minimize power loss while respecting system constraints [11]. Power flow equations are utilized to demonstrate equality limitations, while upper and lower voltage limits are employed to represent inequality constraints. Swarm intelligence and population-based optimization techniques are frequently used to determine the ideal sizes for the devices, while load flow approaches continue to be a viable tool for determining the precise position for placement of these regulators.

FACTS allocation problems have been addressed using a variety of metaheuristic techniques, including Tabu Search (TS), Bat Algorithms (BAT), Whale Optimization Algorithm (WOA), Ant Lion Optimization (ALO) Algorithm, Simulated Annealing (SA), Artificial Bee Colony (ABC), etc. To boost network transfer performance, ABC was utilized by [12] to deploy FACTS regulators in the best possible way. To enhance the loadability of a power system, GA was utilized to efficiently deploy FACTS regulators in a power network. To minimize voltage magnitude changes and losses, BFOA was used by [13] , [14] to determine the best location for UPFC devices. However, there has not been much research done to date to compare the effectiveness of these techniques in FACTS regulator optimization for transmission network capability improvement. Among all the metaheuristic optimization methods, the FA and PSO are two of the most efficacious. The FA was developed based on the distinctive ways that fireflies attract one another. On the other hand, the PSO took inspiration from how insects behave while searching for food. Both the FA and PSO have been demonstrated to be reliable methods for resolving optimization problems, particularly in power systems. Thus, the efficacies of the FA and PSO in solving optimization problems in power systems cannot be overemphasized.

In this study, the STATCOM controller's allocation to enhance network voltage and diminish active and reactive losses is discussed. The implementations of PSO and FA for locating this regulator were described and applied to the IEEE 14-bus system because of their quick convergence and precision compared to other techniques. Two stages of the research were carried out: To begin with, a load flow study was done to find the buses that were over the typical range of permissible voltages. Second, PSO and FA methods were used for sizing the device needed for loss minimization. This study makes a contribution by contrasting the effectiveness of PSO and FA for deploying STATCOM controllers to enhance network functionality. Also, this study is novel in that it implements two separate metaheuristic optimization approaches to allocate STATCOM in the best way possible and determines which methodology is more effective; as a result, it assists power system engineers in society in adopting the quickest and most effective technique for resolving power system issues encountered in society to boost the general standard of living.

2. Model of STATCOM Controller

This controller is a regulator used for reducing transmission losses and alleviating voltage magnitude violation problems. It is made up of a parallel-connected controller and a static VAR generator, which uses different switching patterns within its converter to generate or absorb reactive power. To provide a sufficient supply of electricity, STATCOM corrects for reactive power in the electricity grids. When deployed, it moves more quickly between supplying and consuming reactive power, minimizing power losses and voltage fluctuations.

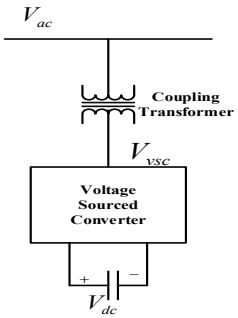


Figure 1: STATCOM controller configuration [15]

2.1. Mode of Operation

A simple STATCOM arrangement is shown in Fig. 1. It comprises a connecting transformer, a capacitor, and a voltage source converter (VSC). A series of three-phase voltages are created from the DC voltage by the VSC. The coupling transformer's functions include connecting the VSC to the high voltage side and preventing short circuits in the DC capacitor

[14]. A change in 3-phase converter voltage V_{vsc} varies the reactive supply to the network. If the STATCOM output voltage V_{vsc} more than the network's voltage V_{ac} (i.e., $V_{vsc} > V_{ac}$), the controller injects reactive power to the grid. Furthermore, if V_{vsc} does not exceed V_{ac} (i.e., $V_{vsc} < V_{ac}$), the STATCOM consumes reactive power from the grid. However, when V_{vsc} and V_{ac} the same (i.e., $V_{vsc} = V_{ac}$), the STATCOM is in standby mode.

2.2. STATCOM Power Flow Model

To control voltage, STATCOM either absorbs or provides reactive power to the network. The STATCOM connection at bus *m* is shown in Fig. 2.

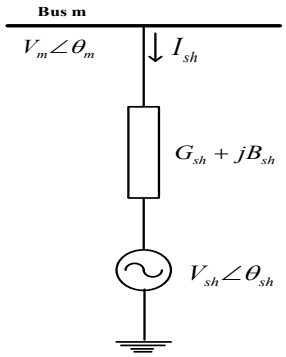


Figure 2: STATCOM Controller Equivalent

Bus *m* load flow equations following STATCOM deployment are stated as (1)–(4).

$$P_m = P_{sh} + \sum_{j=1}^{N} |V_m||V_j||Y_{mj}|\cos\left(\theta_{mj} - \delta_{mj}\right) \tag{1}$$

$$Q_m = Q_{sh} + \sum_{j=1}^{N} |V_m||V_j||Y_{mj}|\sin\left(\theta_{mj} - \delta_{mj}\right) \tag{2}$$

$$P_{sh} = G_{sh}|V_m|^2 - |V_m||V_{sh}||Y_{sh}|\cos\left(\theta_{msh} - \delta_{sh}\right) \tag{3}$$

$$Q_{sh} = B_{sh}|V_m|^2 - |V_m||V_{sh}||Y_{sh}|\sin\left(\theta_{msh} - \delta_{sh}\right) \tag{4}$$

where, $V_m\angle\theta_m$, $V_{sh}\angle\theta_{sh}$ = voltage at bus *m* and at STATCOM, respectively, P_m, Q_m and P_{sh}, Q_{sh} = bus *m* active and reactive power, and STATCOM, Y_{sh}, G_{sh} and B_{sh} = STATCOM's admittance, conductance, and susceptance, $Y_{mj}\angle\delta_{mj}$ = admittance of the line, N = number of buses.

3. Formulation of Problem

The optimum location of FACTS controllers to reduce losses is written as [16]:

Minimize $f(x, \sigma)$

subject to

$$\begin{aligned} &g(x,\sigma) = 0 \\ &h(x) < 0 \\ &x_l < x < x_u \end{aligned} \tag{5}$$

where, $g(x)$, $h(x)$ = equality and inequality constraints, $f(x)$ = total branch loss, σ = system load data, x_l and x_u = the minimum and maximum range.

The solution approach entails optimizing the objective function while satisfying the network restrictions, which include the load flow equations, voltage restrictions, and control parameter bounds [17].

3.1. Objective Function

This is done primarily to reduce overall active loss while remaining within the constraints [18].

$$\min \sum_{k\in N_g} P_{kloss} = \sum_{k\in N_g} g_k(V_i^2 + V_j^2 - 2V_iV_j\cos\theta_{ij}) \tag{6}$$

where, g_k = conductance in p.u., $k = (i, j)$, $i \in N_B$ is the bus number, V_i and V_j = voltage magnitudes in p.u., $j \in N_i$ = bus number adjusted to bus *i*.

3.2. Equality Constraints

Each particle power flow equation is represented by (7)–(8). The load flow solution employs the Newton-Raphson approach.

$$P_{gi} - P_{Li} - V_i\sum_{j\in N} V_j(g_{ij}\cos\theta_{ij} + B_{ij}\sin\theta_{ij}) = 0 \tag{7}$$

$$Q_{gi} - Q_{Li} - V_i\sum_{j\in N} V_j(g_{ij}\sin\theta_{ij} + B_{ij}\cos\theta_{ij}) = 0 \tag{8}$$

where, B_{ij}= susceptance of the branch.

3.3. Inequality Constraints

The load and generator voltages, capacitive reactive power and transformer-tap settings, active and reactive line flow restriction, and power injection are all written as

$$V_i^{\min} \le V_i \le V_i^{\max}, \quad i \in N_B \tag{9}$$

$$Q_{gi}^{\min} \le Q_{gi} \le Q_{gi}^{\max}, \quad i \in N_g \tag{10}$$

$$Q_{ci}^{\min} \le Q_{ci} \le Q_{ci}^{\max} \tag{11}$$

$$T_k^{\min} \le T_k \le T_k^{\max} \tag{12}$$

$$S_l \le S_l^{\max} \tag{13}$$

3.4. Fitness Function Formulation

It is written as

$$F_P = \sum_{q\in N} P_{qloss} + \text{PF} \tag{14}$$

The PF, which is the penalty function, is written as in (15).

$$q_1 \times \sum_{i=1}^{N_G} f(Q_{gi}) + q_2 \times \sum_{i=1}^{N} f(V_i) + q_3 \times \sum_{m=1}^{N_L} f(S_{lm}) \quad (15)$$

And q_1, q_2, q_3 are penalty factors.

$$f(x) = \begin{cases} 0, & if\ x^{\min} \le x \le x^{\max} \\ (x - x^{\max})^2, & if\ x > x^{\max} \\ (x^{\min} - x)^2, & if\ x < x^{\min} \end{cases} \quad (16)$$

where, x^{min} and x^{max} = control parameters.

4. Particle Swarm Optimization

PSO, an algorithm influenced by nature, was created in 1995 [19]. This algorithm uses particle populations to identify the optimum solution. Each particle is taken into account as a potential solution throughout the search process.

The phrases "particle," "swarm," "position," "swarm fitness," "P_{best}," "g_{best}," and the maximum and minimum permitted velocity values are all related to PSO.

Particles are generated at random by the method inside the scope of the function domain. The optimum position that individual particle i has found in the search space is shown by its current velocity (v), personal best position (y_i), and current position (x). Every particle in a d-dimensional area tracks them according to: $x_i = (x_{i1}, x_{i2},...,x_{id})$, $v_i = (v_{i1}, v_{i2},...,v_{id})$, and $P_{best} = (P_{besti,1}, P_{besti,2},...,P_{besti,d})$.

If there are s particles in the swarm.

Then, $i \in 1, ..., s$.

$$y_i(t+1) = \begin{cases} y_i(t) & if \quad f(y_i(t) \le f(x_i(t+1))) \\ x_i(t+1) & if \quad f(y_i(t) > f(x_i(t+1))) \end{cases} \quad (17)$$

$$\hat{y}(t) = \min\{f(y),\ f(\hat{y}(t))\}$$
$$y \in \{y_0(t),\ y_1(t),....,\ y_s(t)\} \quad (18)$$

At each iteration, (17) and (18) update each particle. For each dimension $j \in 1 \ldots n$, if x_{ij}, y_{ij}, and v_{ij} be the j^{th} dimension present position, personal best position and velocity of the i^{th} particle. The new velocity is given by (19).

$$v_{i,j}(t+1) = wv_{i,j}(t) + c_1 r_{1,j}(t)[y_{i,j}(t) - x_{i,j}(t)] + c_2 r_{2,j}(t)[\hat{y}_{i,j}(t) - x_{i,j}(t)] \quad (19)$$

To determine the particle's new position, the new velocity is added to its present position.

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (20)$$

To reduce the likelihood of the particle exiting the search space, all dimensional values of v_i are restricted to [$-v_{max}$, v_{max}]. The v_{max} is determined by (21).

$$v_{\max} = k \times x_{\max}, \quad \text{where } 0.1 \le k \le 1.0 \quad (21)$$

where, x_{max} = domain space search, c_1 and c_2 = coefficients of acceleration.

The PSO convergence behavior is controlled by the inertial weight, which is obtained using (22).

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{itera_{\max}}.\ itera \quad (22)$$

where, $itera_{max}$ = maximum number of iteration, $itera$ = number of iteration, w_{max} and w_{min} = maximum and minimum weighting factor.

4.1. PSO Implementation Algorithm for STATCOM Allocation

The IEEE 14-bus system was utilized to implement PSO. The particle placements were influenced by the initial control variable limitations. Computing the fitness value represented by (14), with the intention of reaching the reduced global best, yielded evaluations of the control variables. The steps for implementing the technique are as follows:

- The population size, total number of iterations, and all control parameters are specified.
- Set iteration number = 0.
- Create the populations and velocities of the particles.
- For loss calculations, run the Newton-Raphson power flow for each individual particle.
- Determine the fitness value for each particle by (14).
- Determine the P_{best} and g_{best} for each particle.
- Let $iteration$=$iteration$ + 1.
- If there is a voltage restriction breach, the velocity and displacement of each individual particle are calculated using (19).
- Find the new location of each particle by (20).
- To calculate the power, run the Newton-Raphson power flow for each particle.
- Using (14), find the fitness value for each particle.
- If the particle's current fitness P is higher than P_{best}, set P_{best} to equal P.
- Set g_{best} to P_{best}.
- Up until the allotted iteration's number is reached, continue from step 7.

The smallest loss values from the relevant fitness value are used to calculate the parameters of g_{best} and the optimum values for the control parameters.

5. Firefly Algorithm

Yang created this algorithm, which is a method for tackling challenging optimization issues quickly [20, 21].

5.1. Firefly Behavior

According to the inverse-square law, the relationship between the intensity of light, I, and distance, r, is inverse. Due to this, the majority of fireflies may be seen at night for a brief period of time, such as a few hundred meters, which is sufficient for flies to converse. A potentially optimizable objective function is used to simulate the flashing light.

5.2. Implementation of Firefly Algorithm

There are three fundamental presumptions that should be taken into account and are stated below [22] for simplicity in the FA description:

- Fireflies have no gender.
- As the distance between fireflies grows, both attractiveness and brightness decrease.
- The objective function's terrain influences the firefly brightness.

The objective functions used in FA for the optimization problem are brightness and light intensity. Finding the optimal solution is similar to being drawn to and moving toward the firefly that is brighter [23].

5.3. Light Intensity and Attractiveness

FA is influenced by two variables: light intensity fluctuation and attractiveness formation.

The brightness of the firefly i and the distance between the two fireflies are both factors in the attractiveness, I, of the firefly i to the firefly j [24]. The expression for light's intensity, which changes with distance, is stated as (23).

$$I_{(r)} = \frac{I_s}{r^2} \tag{23}$$

where, $I_{(r)}$ = light intensity, I_s = intensity of source.

The intensity is expressed as (24).

$$I_{(r)} = I_o e^{-\gamma r} \tag{24}$$

To prevent singularity at $r = 0$, (23) is estimated in gaussian notation as in (25)

$$I_{(r)} = I_o e^{-\gamma r^2} \tag{25}$$

The firefly's brightness I shows its objective function's most recent position, as given by (26).

$$I_i = f(x_i) \tag{26}$$

Each firefly has an attractiveness value represented by β, and the less-bright firefly is drawn to the more-bright firefly. The formula for the variation of β with the distance, r, is stated in (27).

$$\beta_{(r)} = \beta_o e^{-\gamma r^2} \tag{27}$$

where, γ = absorption coefficient of the media light, β_0 = attractiveness value of firefly at $r = 0$.

5.4. Distance and Movement

The formula for the distance r_{ij} between the i^{th} and j^{th} fireflies, respectively located at x_i and x_j, is given by (28).

$$r_{ij} = \left\| x_i - x_j \right\| = \sqrt{\sum_{k=1}^{d} \left(x_i, k - x_j, k \right)^2} \tag{28}$$

Where, x_i, $k = k^{th}$ component of the spatial coordinate x_i of i^{th} firefly, d = distance.

If d equals 2, then (28) changes to (29).

$$r_{ij} = \sqrt{\left(x_i - x_j \right)^2 + \left(y_i - y_j \right)^2} \tag{29}$$

The firefly i^{th} moves towards a more attractive firefly j^{th} as expressed by (29).

$$x_i^{t+1} = x_i^t + \beta_o e^{-\gamma r_{ij}^2} \left(x_j^t - x_i^t \right) + \alpha \left(rand - 0.5 \right) \tag{30}$$

Where, rand = random number within [0, 1], t = current iteration number, α = value randomly selected often inside [0, 1], x_j = brighter firefly location, x_i = less bright firefly, γ = absorption coefficient and it lies in the range of 0.01 and 10.

The algorithm compares the new firefly attractiveness position value to the previous value. If the new site has a higher attraction rating than the old one, the firefly moves there; otherwise, it stays put. A predetermined fitness value determines the FA termination criterion. The brightest firefly will travel at random, according to (31).

$$x_i^{t+1} = x_i^{\ t} + \alpha \varepsilon_i \tag{31}$$

The firefly will move randomly if there are no other fireflies around that are brighter. Up until the stopping condition is satisfied, the aforementioned procedures are repeated. The largest and best-predicted position and capacity are represented by the brightest firefly [25].

5.5. FA Implementation Algorithm for STATCOM Allocation

The following are the procedures in the Firefly algorithm for power flow incorporating a STATCOM controller.

- Enter the network data (independent parameters such as active power of all generators except the swing bus, generators' voltages, regulating transformer-tap setting, reactive power injection) while meeting different equality and inequality constraints.
- Initiate the firefly algorithm's parameters and constants, such as α, β_0 and γ.
- Set the iteration count to 1 and generate 'n' fireflies at random.
- Execute the base case load flow.
- Use the mathematical formulation of the objective function in (14), to calculate the fitness function of each firefly for loss minimization.
- The fitness values are used to generate P_{best} values for all of the fireflies, with g_{best} being the best of the P_{best} values.
- Calculate each firefly's attraction distance by utilizing (29).
- For each firefly, new values are computed.
- Firefly's position is updated using (30).

- For each of the fireflies' new places, new fitness values are calculated. If a firefly's new fitness value is higher than its old P_{best} value, it is set to its current fitness value. G_{best} is calculated using the most recent P_{best} data.
- The iteration number is increased, and if it has not attained its maximum, the process proceeds to step 3 unless convergence is obtained.
- Sort the fireflies into categories based on the current global best. The optimal STATCOM capacities in 'n' candidates are determined by G_{best} firefly, with the position denoting the location and the results presented.

6. Results and Discussion

The load flow study findings, in addition to the applicability of the suggested PSO and FA for STATCOM controller optimum allocation to minimize losses and voltage violations of the IEEE 14-bus network, are shown. The control variables that were tuned include the voltage magnitude, tap parameters of the transformer, and STATCOM output. Table 1 shows these data for these control variables.

Accounting for the STATCOM power injection concept, MATLAB codes for a load flow study were written. These were employed for the load flow study in both cases—without and with the STATCOM controller. During the implementation and evaluation of both approaches on the IEEE 14-bus system, voltage profile augmentation and real as well as reactive power losses were employed as performance metrics.

Table 1: Restrictions on Control Parameters

S/N	Parameters	Limits
1	Voltage Magnitude	0.95 – 1.05 p.u
2	Tap Settings of the Transformer	0.90 – 1.10 p.u
3	Static Compensator MVAr	0.00 – 100 MVAr

6.1. Voltage Profile

The magnitudes of the network voltages are shown in Fig. 3. This implies that the bus voltage magnitudes were greatly enhanced following STATCOM regulator optimization using FA as opposed to when the PSO approach was used for the identical device setup. Bus voltage magnitudes across the entire test network are all within the permissible limits of 0.95 to 1.05 p.u., culminating in dependable network operation. The discrepancy was lessened by these two methods. FA did, however, provide the greatest voltage deviation minimization results in this circumstance.

Buses 2 and 3 offer a compelling justification for this performance. When optimizing with PSO, the bus 2 voltage was 1.048 p.u., and it was 1.046 p.u. after the controller was deployed with FA. The voltage magnitude at bus 3 increased from 0.96 to 0.98 p.u. and then to 0.99 p.u. as a result of optimization utilizing PSO and FA. Given that the anticipated voltage is 1.00 p.u., the best suitable method is one in which network influences attempt to return the voltage to that value.

6.2. Minimization of Active Power Loss

Utilizing optimization techniques for placement strategies, the FACTS controller decreased the active power loss. Fig. 4 depicts the active power loss data for both the PSO and FA-placed STATCOM controllers, as well as the base case. The overall active loss for the base case was recorded at 6.251 MW. Applying PSO and FA to integrate the device reduced the loss to 5.819 and 5.518 MW. The overall loss was minimized by 0.733 MW following the device's incorporation using FA as opposed to 0.432 MW when PSO was employed.

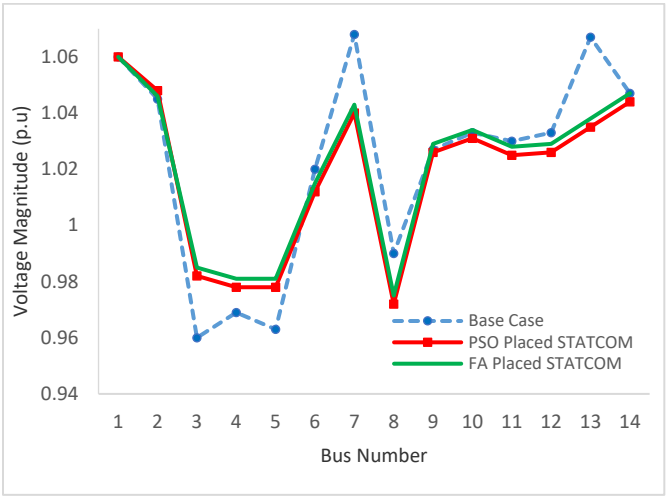


Figure 3: Voltage profile comparison

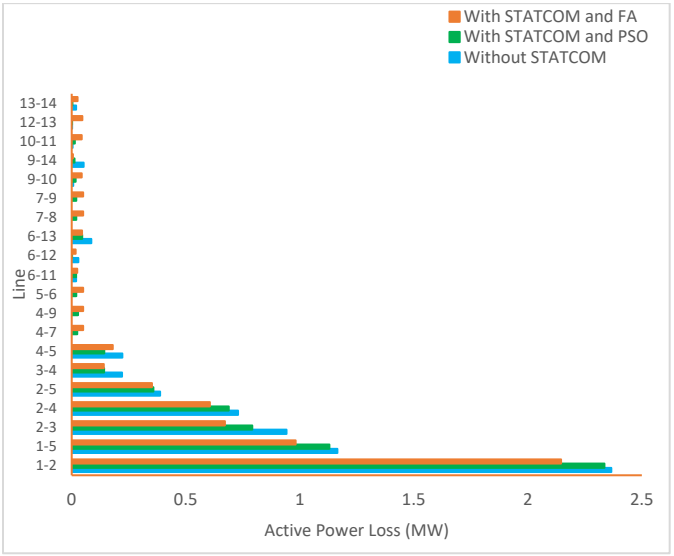


Figure 4: Active power loss minimization for all the cases

By rerouting the system load flow, the loss was reduced. PSO had a loss reduction of 6.9%, whereas FA had a loss reduction of 11.73%. This indicates that FA fared better than PSO in the active loss reduction of the system under study. A more thorough evaluation of the effectiveness of the two techniques in terms of loss reduction is also illustrated in Fig. 4. All of the lines displaying loss decreased following the controller installation utilizing the FA and PSO techniques. The degree of loss reduction does, however, differ between the two strategies. The green bars (loss with the PSO technique) have a substantially higher magnitude than the red bars (FA-placed STATCOM). As seen in the red illustration, these reductions with FA-placed STATCOM substantially outweigh those with

PSO placement. The overall loss minimization is shown in Fig. 5 to help understand how well PSO and FA may be used to optimize STATCOM controllers. It is impossible to exaggerate the advantages of FA over the PSO algorithm. Cost reductions were achieved as a consequence of FA's better minimization findings for active loss and voltage profile augmentation. The STATCOM controller's presence led to a redistribution of network power that improved network operation.

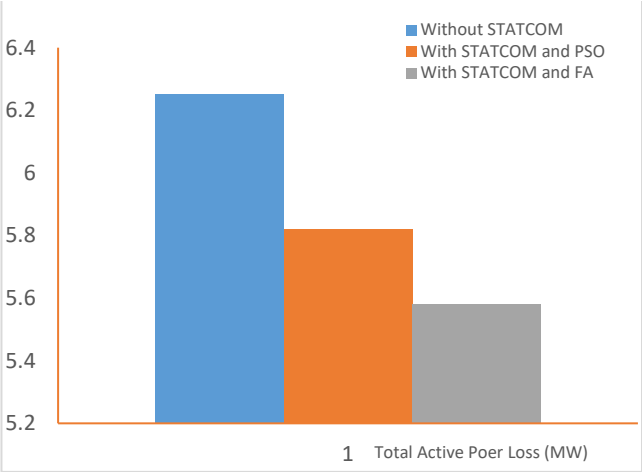


Figure 5: Overall active power losses

The controller offered a different flow path, allowing electricity to flow through less-loaded lines and reducing loss on the original lines as a consequence. The active power flows for the base case, which were 69.9246, 68.7359, 51.8444, 38.2318, 7.5796, 17.2293, 3.6629, and 5.4713 MW for transmission lines 1–5, 2-3, 4, 2, 5, 6, 12, 6, 13, and 14, were modified to 69.8589, 68.7203, 52.1509, 38.5039, 7.5424, 17.1875, and 3.5072 MW. On the other hand, the line flow was improved by 1.40, 1.04, 0.43, 0.08, 0.02, 0.09, 0.20, and 0.03 MW compared to the flow recorded with the PSO technique application.

The system's overall active power flow is therefore increased by utilizing these algorithms, from 621.5 to 623.4 MW with the PSO-placed STATCOM and to 626.64 MW with the FA-placed STATCOM.

6.3. Reduction of Reactive Power Loss

The findings of the network branch losses for the system under study, before and following STATCOM installation, are shown in Fig. 6. The overall power loss without the device was 14.256 MVAr; nevertheless, when STATCOM's optimum configuration was attained with PSO, this decreased to 12.59 MVAr. Following appropriate STATCOM integration using FA, this loss was further reduced to 12.16 MVAr. The STATCOM device's integration with PSO and FA resulted in achievements of 1.62 and 2.10 MVAr, or 11.37 and 14.73%, respectively, in overall reduction. When the two optimization techniques are compared for effectiveness, FA outperforms PSO in minimizing reactive power loss.

As illustrated in Fig. 6, all transmission lines—aside from lines 3–4—were loss-minimized utilizing FA-placed STATCOM. The disparities in reactive loss magnitude for appropriately located STATCOM with PSO and FA show that FA has a loss reduction boost over PSO. The reduction of the system's overall reactive loss with and without correctly positioned STATCOM is illustrated in Fig. 7.

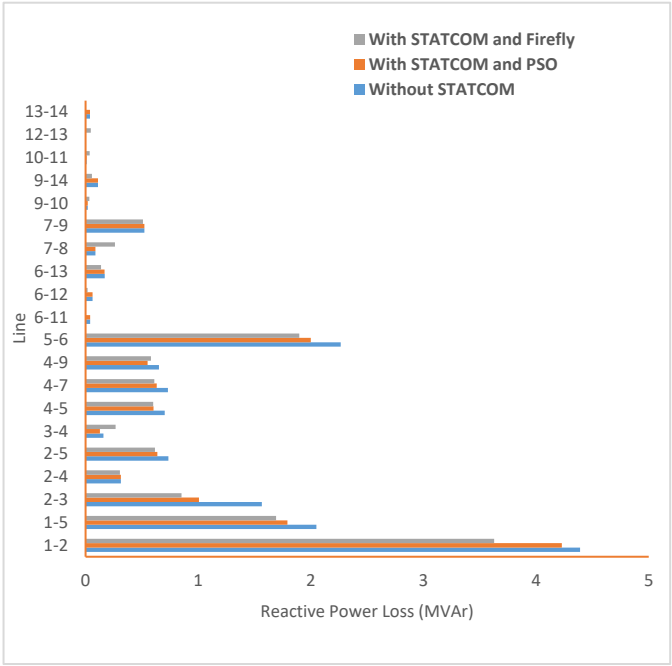


Figure 6: Reduction of reactive loss for all the cases

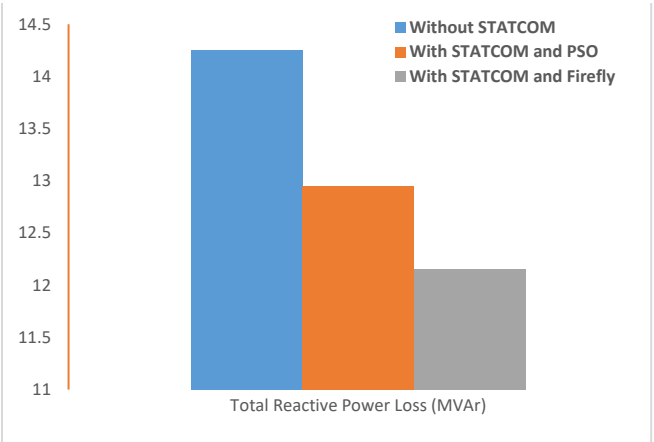


Figure 7: Overall reactive losses

This considerably decreased reactive loss on the network and outperformed the PSO technique. This reduction greatly aided in reducing the bus voltage magnitude deviation, which increased network stability and security. The test system's active and reactive loss information is shown in Table 2. Columns three and four, respectively, reflect the active and reactive power losses experienced by the network during the base case study. With PSO-placed STATCOM, the active power loss is recorded in column 5, while with FA-placed SATCOM, it is recorded in column 7. Columns six and eight of the table contain their related reactive power losses. As a consequence of the device using these two techniques, there is an overall line-by-line decrease for the active and reactive power.

Table 2: The IEEE 14-Bus Network Line Losses

Bus Number		Steady State (Base case)		STATCOM (PSO-placed)		STATCOM (Firefly-placed)	
From	***To***	***(MW)***	***(MVAr)***	***(MW)***	***(MVAr)***	***(MW)***	***(MVAr)***
1	2	2.366	4.390	2.346	4.370	2.146	3.628
1	5	1.165	2.049	1.129	1.787	0.982	1.692
2	3	0.942	1.565	0.819	0.947	0.672	0.952
2	4	0.729	0.313	0.726	0.415	0.706	0.395
2	5	0.388	0.736	0.372	0.676	0.352	0.696
3	4	0.221	0.158	0.161	0.247	0.141	0.267
4	5	0.222	0.703	0.200	0.698	0.180	0.678
4	7	0.000	0.731	0.030	0.671	0.050	0.651
4	9	0.000	0.651	0.030	0.601	0.050	0.581
5	6	0.000	1.898	0.030	2.265	0.050	2.245
6	11	0.019	0.041	0.004	0.023	0.024	0.003
6	12	0.029	0.062	0.002	0.038	0.017	0.018
6	13	0.086	0.170	0.065	0.158	0.045	0.138
7	8	0.000	0.087	0.030	0.281	0.050	0.261
7	9	0.000	0.522	0.030	0.530	0.050	0.510
9	10	0.007	0.019	0.024	0.014	0.044	0.034
9	14	0.052	0.111	0.020	0.077	0.005	0.057
10	11	0.004	0.009	0.024	0.016	0.044	0.036
12	13	0.002	0.001	0.027	0.027	0.047	0.047
13	14	0.019	0.039	0.006	0.017	0.026	0.002
Total		6.251	14.256	5.819	12.954	5.681	12.891

Comparing with the PSO technique, FA's efficacy cannot be highlighted enough. With this performance, FA was able to minimize active and reactive power losses as well as voltage fluctuations more effectively, which reduced costs.

For better comprehension, Table 3 shows the entire system power flows and the corresponding total loss projections. Without a STATCOM device, Table 3 shows that the network is capable of handling an apparent power of 653.38 MVA. But with the PSO and FA installed STATCOM controllers, the apparent power increased to 671.9 and 676.1 MVA, respectively. The system loss decreased from 15.57 to 14.20 and 13.81 MVA, respectively, as a result of this rise in total network power, as depicted in Table 3, when the device was strategically placed with PSO and FA, respectively.

Following the utilization of PSO and FA algorithms, the STATCOM allocation resulted in an improvement in overall flow of 2.8 and 3.5%, respectively. Deploying this device and employing the PSO and FA algorithms led to a reduction of the overall network loss of 8.78 and 11.26%, respectively. The stated FA performance in loss reduction and total network flow clearly demonstrates that FA is superior to PSO in the deployment of STATCOM device controllers. Table 3 indicates the differences in STATCOM device capacities.

Reactive power injection is represented by column 5, while STATCOM controller voltages and angles are shown in columns 3 and 4, respectively. Column 2 displays the device location that was selected. Table 4 shows the comparison of the total parameter settings and STATCOM controller location that led to the network performance for FA and PSO that was previously described. The table makes it evident that both algorithms' shunt reactances fall within the same range. As a consequence, the final device rating—which depends on the controller capacity and potential costs for the two techniques—is rather similar. Due to this, FA outperforms PSO in terms of cost.

Table 3: The IEEE 14-Bus Network Line Losses Network Overall Power Flows and Losses

	Active and Reactive Power Flows			**Active and Reactive Power Losses**		
	Base Case	***PSO-placed STATCOM***	***FA-placed STATCOM***	***Base Case***	***PSO-placed STATCOM***	***FA-placed STATCOM***
Active (MW)	621.5	623.4	626.6	6.3	5.8	5.7
Reactive (MVAr)	201.7	250.8	253.9	14.3	12.9	12.9
Apparent (MVA)	653.4	671.9	676.1	15.6	14.2	14.1

Table 4: STATCOM Parameters Settings and Location

Technique	**Location**	**Voltage Value (p.u)**	**Angle (deg.)**	**STATCOM Size (MVAr)**
FA	9	1.029	0.926	9.54
PSO	11	1.025	3.769	8.96

7. Conclusion

This study looked into and proved the efficacy of the FA algorithm over the PSO method for placing STATCOM devices optimally. In this research study, the ideal STATCOM controller placement and parameter settings were made with the goals of reducing voltage magnitude variations and active and reactive power losses. The outcomes produced utilizing the IEEE 14-bus network show how appropriate these optimization strategies are. The capacity of the STATCOM controller to produce the best results for the specified objectives served as evidence of the applicability of PSO and FA for the best STATCOM controller position. According to the research, the STATCOM controller's performance with FA placement is superior to the PSO's. This means that FA performance in the optimum STATCOM controller configuration outperforms PSO in voltage profile augmentation and loss mitigation situations. Future research may be carried out to compare the effectiveness of the FA with other recently developed metaheuristic optimization algorithms that deliver superior performance at a reduced cost of STATCOM allocations on the power transmission and sub-transmission networks.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] E. I. Ogunwole, S. O. Ayanlade, D. E. Owolabi, A. Jimoh, A. B. Jimoh and F. M. Aremu, "Performance Comparative Evaluation of Metaheuristic Optimization Algorithms for Optimal Placement of Flexible Alternating Current Transmission System Device," in 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET), 2022, 1-8, doi: 10.1109/ICECET55527.2022.9872866.

[2] B. Behera and K. C. Rout, "Comparative performance analysis of SVC, STATCOM UPFC during three-phase symmetrical fault," Proc. Int. Conf.

Inven. Commun. Comput. Technol. ICICCT 2018, **2**, 1695–1700, 2018, doi: 10.1109/ICICCT.2018.8473226.

[3] K. G. Damor, D. M. Patel, V. Agrawal, and H. G. Patel, "Comparison of different FACT devices," Int. J. Sci. Eng., **1**(1), 372–375, 2014, doi: 10.1109/APCC.2016.7581484.

[4] Jimoh, A., Ayanlade, S. O., Ariyo, F. K. and Jimoh, A. B., "Variations in phase conductor size and spacing on power losses on the Nigerian distribution network. Bulletin of Electrical Engineering and Informatics," **11**(3), 1222 - 1233, 2022, doi: https://doi.org/10.11591/eei.v11i3.3753.

[5] E. I. Ogunwole, "Optimal placement of statcom controllers with metaheuristic algorithms for network power loss reduction and voltage profile deviation minimization," M.Tech Dissatation, Kwa Zulu Natal University, 2020.

[6] B. O. Adewolu and A. K. Saha, "FACTS devices loss consideration in placement approach for available transfer capability enhancement," Int. J. Eng. Res. Africa, **49**, 104–129, 2020, doi: 10.4028/www.scientific.net/JERA.49.104.

[7] S. O. Ayanlade and O. A. Komolafe, "Distribution system voltage profile improvement based on network structural characteristics," in OAU Faculty of Technology Conference (OAUTEKConF2019), 2019, 75–80.

[8] A. K. Rawat et al., "Design of microcontroller based static VAR compensator," 2015 17th Eur. Conf. Power Electron. Appl. EPE-ECCE Eur. 2015, **4**(1), 1–6, 2017, doi: 10.1515/ijeeps-2017-0145.

[9] A. Gupta and P. R. Sharma, "Optimal placement of FACTS devices for voltage stability using line indicators," in 2012 IEEE 5th Power India Conf. PICONF 2012, 4–6, 2012, doi: 10.1109/PowerI.2012.6479518.

[10] S. T. Fadhil and A. M. Vural, "Comparison of dynamic performances of TCSC, STATCOM, SSSC on inter-area oscillations," in 2018 5th Int. Conf. Electr. Electron. Eng. ICEEE, 138–142, 2018, doi: 10.1109/ICEEE2.2018.8391317.

[11] Y. Zhang, Y. Zhang, B. Wu, and J. Zhou, "Power injection model of STATCOM with control and operating limit for power flow and voltage stability analysis," Electr. Power Syst. Res., **76**(12), 1003–1010, 2006, doi: 10.1016/j.epsr.2005.12.005.

[12] D. Karaboga and B. Akay, "A comparative study of artificial bee colony algorithm," Appl. Math. Comput., **214**(1), 108–132, 2009, doi: 10.1016/j.amc.2009.03.090.

[13] M. Sankaramoorthy and M. Veluchamy, "A hybrid MACO and BFOA algorithm for power loss minimization and total cost reduction in distribution systems," Turkish J. Electr. Eng. Comput. Sci., **25**(1), 337–351, 2017, doi: 10.3906/elk-1410-191.

[14] A. Elansari, J. Burr, S. Finney, and M. Edrah, "Optimal location for shunt connected reactive power compensation," in Proc. Univ. Power Eng. Conf., 1–6, 2014, doi: 10.1109/UPEC.2014.6934743.

[15] M. O. Okelola, S. A. Salimon, O. A. Adegbola, E. I. Ogunwole, S. O. Ayanlade, and B. A. Aderemi, "Optimal siting and sizing of D-STATCOM in distribution system using new voltage stability index and bat algorithm," **2**(2), 2–6, 2021.

[16] S. Majumdar, A. K. Chakraborty, and P. K. Chattopadhyay, "Active power loss minimization with FACTS devices using SA/PSO techniques," in 2009 Int. Conf. Power Syst. ICPS '09, 1–5, 2009, doi: 10.1109/ICPWS.2009.5442726.

[17] A. A. A. Esmin and G. Lambert-Torres, "Loss power minimization using particle swarm optimization," in IEEE Int. Conf. Neural Networks - Conf. Proc., 1988–1992, 2006, doi: 10.1109/ijcnn.2006.246945.

[18] S. O. Ayanlade, E. I. Ogunwole, S. A. Salimon, and S. O. Ezekiel, "Effect of optimal placement of shunt facts devices on transmission network using firefly algorithm for voltage profile improvement and loss minimization," Advances on Intelligent Informatics and Computing: Health Informatics, Intelligent Systems, Data Science and Smart Computing, **127,** 385-396, 2022.

[19] M. O. Okelola, S. O. Ayanlade, and E. I. Ogunwole, "Particle swarm optimisation for optimal allocation of STATCOM on transmission network," in Journal of Physics: Conference Series, 2021.

[20] X. S. Yang and X. He, "Firefly algorithm: recent advances and applications," Int. J. Swarm Intell., **1**(1), 36, 2013, doi: 10.1504/ijsi.2013.055801.

[21] X. S. Yang, "Firefly algorithms for multimodal optimization," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5792 LNCS, 169–178, 2009, doi: 10.1007/978-3-642-04944-6_14.

[22] A. Ritthipakdee, A. Thammano, N. Premasathian, and D. Jitkongchuen, "Firefly mating algorithm for continuous optimization problems," Comput. Intell. Neurosci., **2017**, 2017, doi: 10.1155/2017/8034573.

[23] F. S. Moustafa, N. M. Badra, and A. Y. Abdelaziz, "Evaluation of the performance of different firefly algorithms to the economic load dispatch problem in electrical power systems," Int. J. Eng. Sci. Technol., **9**(2), 1, 2017, doi: 10.4314/ijest.v9i2.1.

[24] N. F. Johari, A. M. Zain, N. H. Mustaffa, and A. Udin, "Firefly algorithm for optimization problem," Appl. Mech. Mater., **421**, 512–517, 2013, doi: 10.4028/www.scientific.net/AMM.421.512.

[25] S. O. Ayanlade, E. I. Ogunwole, A. Jimoh, S. O. Ezekiel, D. E. Owolabi, and A. B. Jimoh, "STATCOM Allocation Using Firefly Algorithm for Loss Minimization and Voltage Profile Enhancement," in 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), 2022, 1-6, doi: 10.1109/ICECCME55909.2022.9988475.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 118-135 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Social Financial Technologies for the Development of Enterprises and the Russian Economy

Evgeniy Kostyrin*, Evgeniy Sokolov

Bauman Moscow State Technical University, Engineering Business and Management, Finance, Moscow, 105005, Russia

ARTICLE INFO

Article history:
Received: 27 February, 2023
Accepted: 05 May, 2023
Online: 21 May, 2023

Keywords:
Economic and mathematical model
Working citizen
Social state

ABSTRACT

*The main **contradiction** identified in the study is that the existing scientific and methodological solves of economy development management processes does not create prerequisites for improving the efficiency of their work, the introduction of progressive technologies for material and moral stimulation of the work of performers and administrative and managerial personnel, advanced social mechanisms for country's economy development and social security employees. Economic and mathematical modeling of the complex system of social financing of enterprises and the economy of the country, scientifically sound personnel policy and the system of motivation of performers and administrative and management personnel **is an important and urgent problem**. **The purpose of the study is** to develop and implement an economic and mathematical model of a comprehensive system of social financing of enterprises and the economy of the country, optimizing the wages of the workforce, consistent with revenue growth, deductions for the development of the enterprise (relevant for the employer and the entire workforce), taxation and social contributions (important for the state). **The results** of the studies conducted and presented in this article allow us to conclude that the proposed social financial technologies for the development of enterprises and the economy of Russia, make it possible, at quite achievable rates of growth of gross domestic product (revenue of enterprises) by 3% per year, to ensure an increase in the wages of working citizens for 5 years by 34 %, which will make it possible to practically end poverty, and to increase contributions to the development fund over 5 years by 16%. Starting from 2026, increase receipts from income tax, tax on profit rate and value added tax and bring this growth to 30% by 2041, which will allow the state to solve many social problems.*

1. Introduction

Social financial technologies for the development of enterprises and the economy of any country in the world are connected with the life of a person and his relationships with other people in society. The realization of the most diverse needs of people in goods, works, services in society is provided only in the process of work, thus, the social financial technologies considered in this study should be perceived as a reflection of labor relations aimed at self-realization of citizens of working age.

For the first time, social protection of the population, including citizens working at enterprises, was issued through collective insurance.

According to the existing legislation of the Russian Federation, all enterprises, companies, organizations must make contributions: to the Pension Fund of Russia (PFR) – 22% of wages; to the Federal Compulsory Medical Insurance Fund (FCMIF) – 5.1% of wages; to the Social Insurance Fund (SIF) – 2.9% of wages (in case of temporary disability and maternity). Also, according to Federal Law No. 517-FZ of December 19, 2022, depending on the class of occupational risk, contributions in the amount of 0.2% to 8.5% of wages are paid.

All citizens of the Russian Federation are associated with the aforementioned funds. So, working citizens make contributions to the PFR and the FCMIF, and this is 82,678 thousand people, more than 36 million people receive a penis in our country, which means that in total it is already more than 118 million people, i.e. more than 80% of the population of the country, which indicates the importance and urgent need to study issues related to the effective management of funds of social funds and optimization of contributions of enterprises to these funds, so that the relationship with the funds does not lead to a decrease in the efficiency of enterprises and organizations, but on the contrary, is

*Corresponding Author: Evgeniy Kostyrin, mauntain76@mail.ru

an incentive to the development of the enterprise, the economy of the country, motivation of personnel to work.

On the topic of this study, we analyzed 176 articles by different researchers and specialists. The analysis of articles devoted to the practical implementation of financial and social mechanisms of personnel management and enterprise development, which take into account the interests of the workforce, managers, shareholders and owners of enterprises, as well as the state as one of the stakeholders in the growth of wages of enterprise personnel, its development and increase in tax deductions to the federal and territorial budgets, can be divided into the following groups: Cash Flow Management (30 articles), Investment Efficiency (34 works), Socio-economic Security (23 articles), Welfare, Wage Growth (27 articles), Enterprise Development (12 scientific studies), Economic and Mathematical Modeling (23 scientific articles) and Business Process Management (27 articles). The relative weight of each group in the review is shown in Figure 1.

Figure 1: Distribution of Articles by Research Topics

Figure 1 shows that the largest share of the studied articles on the research topic falls on the Investment Efficiency section (34 articles, or a fifth of all works), followed by a problem related to cash Flow Management (Cash Flow Management), with the number of articles 30, which is 17% of the literature review. The group of articles devoted to Business Process Management closes the top three most popular research topics (28 papers, 16% of articles). The smallest group of works belongs to the Enterprise Development section with a total number of articles equal to 12, which is 7% of the list of analyzed articles.

The degree of elaboration of the research problem and its analysis are presented in detail in Table 1.

Table 1: Analysis of works devoted to the problems of this study

Articles directions	Scientists and specialists who have contributed to the subject of the study
Welfare, Wage Growth	[1-27]
Enterprise Development	[28-39]
Investment Efficiency	[40-73]
Socio-economic Security	[74-96]
Cash Flow Management	[97-126]
Economic and Mathematical Modeling	[127-149]
Business Process Management	[150-176]

A comparative analysis of the scientific results obtained by the authors and the results of other scientists and specialists dealing with certain aspects of the problems raised in the article is presented in Table 2.

Table 2: Comparison with Other Studies and Scientific Increment of Knowledge

Number	Research directions	Scientific result	Scientific novelty
1	Welfare, Wage Growth	Social technologies have been developed for the financing of enterprises and the development of the country's economy, maximizing the wages of employees, taxation and social contributions of enterprises, which can significantly increase the welfare of working citizens and ensure the supply of products in accordance with the increased demand for it and practically end poverty.	In contrast to the models of social financing of enterprises used in practice [14; 22] the basis of this social technology of enterprise financing is a breakthrough mechanism for increasing wage growth, material and moral incentives for workers, based on taking into account the needs of working citizens, managers, owners and shareholders of enterprises and the state, which allows the entire workforce to participate in the process of enterprise management, be a full participant in socio-economic relations in society, the development of the country and enterprises by directing part of the company's revenue to its development and the growth of the welfare of citizens of the country, equipping the workplaces of employees, improving their qualifications.

2	Enterprise Development	A unique social technology for financing the economy of the country and its enterprises has been developed, which allows linking profits from the sale of products and services with additional remuneration of personnel, labor, taxation and social contributions of enterprises to the FCMIF, PFR and SIF, as well as contributions to the development of the enterprise.	In contrast to the well-known macroeconomic models of managing the country's economy and microeconomic models of managing the development of enterprises [28-30] the developed method allows us to simultaneously take into account the interests of the labor collective, managers and shareholders of enterprises and the state as a whole through the use of progressive mathematical models aimed at solving the main tasks, the purpose of which is the effective management of social contributions to the PFR, FCMIF, SIF and effective, interrelated relationships with the needs of working citizens in wage growth, by the state and owners on the basis of making complex scientifically based management decisions with a synergistic effect. This approach gives a significant increase in the economic, social and professional efficiency of working citizens and the economy of the country as a whole in comparison with their work, provided that the interests of only one of the parties are taken into account: working citizens, business owners or the state as an institution for data collection. taxes and social contributions of citizens and enterprises. Thus, the difference between the methodology of this scientific research compared to analogues lies in the complexity and comprehensiveness of the management of the enterprise and working citizens, based on the interests of all participating groups of beneficiaries.
3	Investment Efficiency	An approach to the economic assessment of the effectiveness of the development of the country's economy and enterprises, as well as investments in the development of enterprises, equipping the workplace of an employee and improving his qualifications, which consists in making a management decision to redistribute the effect of reducing the cost of products, services, works, goods associated with the growth of income of enterprises per employee, and the direction of this effect on the increase in wages and the development of the enterprise.	This approach differs from other methods of assessing enterprises and the country's economy, as well as the effectiveness of investments [45; 54; 66] in that the criterion for the effectiveness of the management decision is the growth of wages of the workforce and the growth of the welfare of the population, which is ensured by the optimal redistribution of the effect of reducing the cost of goods between wages, enterprise development, social contributions, taxation of enterprises, while taking into account the interests of all beneficiaries. In addition, in the author's approach, additional profit from increased sales is directed to investments in the entrepreneurship development fund, payroll, taxation and social contributions, the amount of which significantly depends on the average monthly revenue of enterprises per working citizen, which motivates him to high effective, high-performance work makes him not in words, but on a participant in labor relations interested in the development of the enterprise where he works and the country in which he lives. Thus, the author's methodology makes such an approach to social technologies for financing enterprises and the country's economy closed, complex and dynamic, increasing the well-being of citizens working at the enterprise.
4	Socio-economic Security	An approach to ensuring the social and financial security of the state and the population is proposed, the essence of which is that the authors propose technologies that ensure wage growth, bringing the pension level to 40% of wages in four years; in 8 years – up to 60% of wages; in 10 years – up to 80% of wages fees, an increase in total FCMIF and PFR receipts, as well as tax on profit, income tax and VAT by 30% of the base level, while reducing FCMIF and PFR receipts by 14.08% over five years.	In contrast to the works [74; 80], the developed technology is aimed at increasing the investment attractiveness of the country's economy and domestic enterprises by applying in practice the progressive system of personnel labor incentives developed by the authors and the optimal distribution of funds from reducing the cost of production and investments between the payroll, development fund and contributions to the PFR, FCMIF, income tax, VAT and tax on profit are based on solving the problem of economic and mathematical modeling and using social technologies of financing to manage enterprises and the economy of the country as a whole.

5	Cash Flow Management	A comprehensive model for managing the cash flows of an enterprise, its income and expenses, cost, unit fixed costs and unit variable costs, the effect of cost reduction, tax and social deductions from the sale of products has been developed.	The proposed model differs from the models used for managing financial flows of enterprises [99; 105; 109; 126], by enabling management decision makers to coordinate investment programs and plans depending on the prices of final products, goods, works, services, their volumes and production costs, which contributes to the growth of profitability of investments in the economy of the country and enterprises, labor productivity of personnel (citizens working at the enterprise) and finance for the development of enterprises.
6	Economic and Mathematical Modeling	A new formulation and approach to solving the problem of nonlinear programming about the optimal combination of labor wages compatible with an increase in company revenue, tax revenues, which is important for the development of the state and the growth of the welfare of its citizens, and to the enterprise development fund, which is important for its management and shareholders, owners, are proposed.	In contrast to the famous researchers devoted to solving the problem of investment allocation [130; 133; 137; 138], in the author's approach, the company's budget is used as a source of financing, which is formed by contributions to the development fund, depending on the effectiveness of each employee's activities. This approach allows the most efficient redistribution of financial flows between the payroll fund, the enterprise development fund, taxation and social contributions and provides sources of financing for equipping the workplace of personnel, improving the skills of employees, involves all personnel in the enterprise management.
7	Business Process Management	An economic and mathematical model of wage maximization consistent with the growth of the revenue, process management of cash flows of enterprises, characterized by a systematic combination of methods of nonlinear programming, economic and mathematical modeling, social financial technologies, has been developed, this makes it possible to create tools for managing the cash flow and development of enterprises and to develop standard projects of management decision support systems with the prospect of their integration into existing and promising information and analytical systems at enterprises that ensure a combination of the interests of the workforce, owners, managers of enterprises and the state.	The developed complex model of nonlinear programming, which maximizes wages consistent with the growth of the company's revenue and the funds that are released from the company by reducing the cost, allows us to find an optimal scientifically based management solution to the complex problem of combining and taking into account the development goals of the state, the enterprise and the needs of the workforce by dividing into smaller subtasks: increasing the wages of the labor collective through the use in practice of material and moral factors to increase productivity and labor efficiency, revenue growth at the enterprise, reduction in the unit cost of production, increase in contributions to the development fund and wages, tax deductions and social contributions to the FCMIF, PFR and SIF, each of which represents a well-known task of economic andmathematical modeling, and then the integration of the results into a comprehensive management decision-making system at the stage of practical implementation by applying additional criteria: market capacity, profit share, aimed at increasing material incentives for personnel, for the development of the economies of countries and enterprises, for investment, etc.

Thus, the literature review presented in Tables 1 and 2 allowed us to conclude that in scientific research there are practically no economic and mathematical models and technologies of social financing of enterprises using methods of nonlinear programming and mathematical optimization, which allow the manager to ensure the relationship of financial indicators of enterprises (revenue, profit, unit cost of production) and a progressive system of material incentives for the labor of citizens working at enterprises, as well as tax and social contributions to the FCMIF, PFR and SIF, which allows for a synergistic effect as a result of taking into account the needs of workers (citizens working at enterprises), owners, managers of enterprises, budgets of all levels (federal and territorial) and the state as a whole. The article [25] shows that social financial technologies are based on the construction of models of economic and mathematical optimization, flowcharts, tools, algorithms, software products and environments and allow solving two problems:

1) Efficient allocation of enterprises' contributions to off-budget funds and their subsequent use (PFR, FCMIF, SIF).

2) Effective distribution and accounting at enterprises of the needs of the state, owners, shareholders and the workforce.

The solution to the first problem should be sought in the transition of healthcare and pension insurance to personalized medical and pension accounts of citizens [25-27; 85-95].

The article [25] solves the second problem of optimizing social financial technologies, taking into account various interests

both in the country as a whole and within labor relations at enterprises.

In the same article, using a nonlinear model, it is also proved that in the case of an increase in wages at an enterprise above the average in Russia, it is advantageous for the state to reduce deductions to FCMIF by increasing tax deductions due to wage growth.

In the article [93], using a nonlinear model, it is also proved that when wages increase above the average in Russia (consistent with revenue growth), it is advantageous for state and employer to reduce deductions to the PFR.

In the article [95], using a nonlinear model, it is proved that when switching, starting from 2021, to personalized accumulative pension accounts (PAPA) for 19 years (by 2039), 11,833,027 rubles will accumulate on PAPA of working citizens (with an average salary of 54,175 rubles), which is enough to provide a pension in the amount of 80% of wages for 20 years, which corresponds to the expected period of survival.

In other words, after 2038, every working citizen with a salary of 52,355 rubles (the average salary in Russia in August 2021) will have an amount of 11,099,133.99 rubles on his account, which is enough to receive a monthly pension of 80% of the national average salary for the rest of his life (the survival period). Accordingly, enterprises after 2038 can stop deducting 22% of wages in favor of citizens, which will reduce the unit cost of goods produced in the country, and hence the price of their sale to the population, and this in turn is already an effective method of combating inflation in the country.

This article shows that the post-retirement transition to a reduction in the percentage of deductions from wages for pension and medical support of citizens from the first (2021), and not from 2038, is much more profitable for the labor collective, employers and the state.

In this article, the main ***contradiction*** is noted, which consists in the fact that the existing organizational, economic and scientific support for the development of enterprises does not create prerequisites for improving the efficiency of their work, the introduction of progressive technologies for material and moral stimulation of the labor of performers and administrative and managerial personnel, advanced social technologies for financing the economy of the country and the development of enterprises, as well as social security of employees.

The existing processes of socio-financial development of the country's economy and individual enterprises can be characterized as extremely inefficient, since there is no holistic scientific and methodological approach integrated into everyday practice to develop and apply effective management solutions based on optimization tools, models and breakthrough financial management algorithms. The low efficiency of managing the country's economy and the socio-financial activities of enterprises is associated with the fragmentation and imperfection of the mathematical apparatus and tools used in practice, as well as methods of stimulating labor. Attempts to create a full-fledged integrated enterprise development management system, including a progressive incentive system for employees, receipts to the development fund, as well as effective financial resource management mechanisms based on the methodology of nonlinear programming of enterprise development management processes and their individual assets, organizational and structural units built in this article, are presented in [25-27; 85-95], but their main drawback is that the issues of social financing of the Russian economy and domestic enterprises have not been fully worked out, mutually linking the interests of the state, the labor collective, owners, shareholders, heads of enterprises, problems of enterprise taxation management, social contributions to the PFR, FCMIF and SIF are not fully reflected depending on from the growth of revenue of enterprises, the growth of employee productivity, criteria for making informed management decisions have not been developed, aimed at organizing effective labor relations in the company.

Ambitious goals require advanced social financial technologies and optimization models for the development of enterprises and promising systems for financing their activities, taking into account the optimization of taxation and social contributions to FCMIF, PFR and SIF. Nonlinear processes and computational methods in the management of such systems are becoming increasingly relevant as the most effective tools for making managerial decisions and providing scientifically sound algorithms and models for the development of management objects. When managing complex systems with many interrelated parameters, it becomes necessary to use algorithms and methods of nonlinear programming to achieve the optimal result from the set of possible values of the dependent variable with limited ranges of changes in influencing factors. As a rule, both the objective function and each of the inequalities of the system of constraints of the optimization problem in most modern control models of real processes are nonlinear functions, which imposes additional restrictions on control objects and requires special mathematical models and instrumental methods for solving such problems. Thus, optimization methods, in particular nonlinear programming models, have proven themselves well in macroeconomic problems and problems of managing the development of enterprises using social financing technologies [85; 88; 91].

Thus, economic and mathematical modeling of a complex system of social financing of the country's economy and enterprises, optimization of employees' wages tied to the company's revenue, optimization of the size of the enterprise development fund, which is important for the owners of enterprises and all employees, as well as taxes, which ensures the performance of state functions, a well-thought-out policy towards existing and potential consumers and categories of goods, an instrumental basis for managing the development of enterprises, taking into account promising and effective technologies for their financing and investment in development, structural analysis, system analysis, factors of the internal and external environment, a scientifically sound personnel policy and a system of motivation of performers and administrative and managerial personnel ***is an important problem for the national economy***.

The purpose of the study is to build and implement an optimization model of an integrated system of social financing of the country's economy and enterprises, maximizing the wages of the workforce associated with income for the development of the enterprise, which is important and relevant for owners and

shareholders, the management of the enterprise, increasing revenue, taxation and contributions to off-budget funds, which is relevant for the state.

The object of the study is the average Russian enterprise, its financial flows, taxation and social contributions to extra-budgetary funds (FCMIF, PFR and SIF). The article discusses social financial technologies with the average monthly salary of employees according to the Federal State Statistics Service, the share of wages in the structure of gross domestic product (GDP) of Russia in the amount of 44.9% and the profitability of products in the amount of 9.9%.

The subject of the study is modeling the optimal distribution of investments in wages of labor, consistent with revenue growth, in the enterprise development fund, taxation and deductions to off-budget funds, which is relevant for the state, using methods of nonlinear programming and process management.

For the purposes of this article, we will define the concepts of "products", "goods", "work", "service". According to Article 38 of the Tax Code of the Russian Federation, by ***goods*** we will mean any property sold or intended for sale. Without loss of meaning in this article, the terms "products" and "goods" are identified.

Work is an activity whose results have a material expression and can be implemented to meet the needs of an organization and (or) individuals.

A ***service*** is an activity whose results have no material expression, are realized and consumed in the process of carrying out this activity.

Materials and Methods

The financial system of Russia consists of three subsystems: public finance; finance of enterprises (organizations) and household finance. As shown in [92], the basis of the Russian financial system is working citizens, on whose well-organized and motivated work its condition depends. The personnel of enterprises providing services to the population, performing work, producing products and essential goods also need to create and implement in daily practice an effective motivation system that can become a source of scientifically sound management decisions, involve each employee of the enterprise (organization) in the management system of all divisions of the enterprise. enterprises create the material and moral foundations of responsibility for the quality and efficiency of their work. The foundation of the motivation system can be an economic and mathematical model of enterprise profit management, which is considered in the works [25; 88; 90; 95]. The introduction of this model into the daily activities of enterprises for many years has shown their high efficiency, allows you to manage profits due to the optimal ratio of such economic indicators of enterprises as prices for products, goods, works, services, production volumes and sales of enterprise products, cost. When managing enterprises, optimization models of profit management make it possible to link the reduction in the cost of enterprises' products, which provides additional demand for products and goods, with an increase in their production and sales volumes. Moreover, the growth of volumes allows you to get a double effect. The first is directly related to the growth of volumes. The second is associated with a reduction in the cost of production, due to a reduction in unit fixed costs with the growth of the company's production.

One of the key tasks of this work is to build a comprehensive system of stimulating the work of personnel of enterprises, analysis and justification of ways to improve the working conditions and workplace of employees of enterprises, without motivation and initiative work of which it is impossible to achieve high financial results.

The work [94] shows that the most important factor in improving the Russian finance system are working citizens, on whose material and moral incentives depend not only household incomes (family budgets), but also the fullness of budgets at all levels of the financial system, namely: the finances of the state and the finances of economic entities (enterprises, organizations). Thus, there is a need to create such an economic and mathematical model that would allow calculating the amount of material incentives for employees of enterprises and all restrictions on the range of changes affecting its size factors. We will build an optimization model that ensures the maximization of material incentives for the work of employees of enterprises and the level of financial contributions for its development is not lower than indicated.

A common practice of managing the development of enterprises is the material stimulation of workers' labor as a percentage of total revenue. Therefore, as a target function of the developed model, we will take the amount of deductions from the company's revenue directed to financial incentives for employees: $Sal_j = \theta_j \cdot Rev_j / 12 \rightarrow max$, where Sal_j stands for the monthly financial remuneration of employees of the j-th division of the enterprise in rubles; θ_j stands for the percentage of revenue from the revenue of the j-th division of the enterprise, directed to stimulating the work of performers (employees of the j-th division) in fractions of units; Rev_j stands for the revenue of the j-th division of the enterprise in rubles.

To stimulate the company's employees to increase revenue and productivity of their labor, it is necessary to create a progressive scale of material remuneration. In other words, the percentage of income allocated to the remuneration of employees of the enterprise should depend on the amount of this income, i.e. $\theta_j(Rev_j)$. We will assume that all income earned by employees of the enterprise in excess of the base amount is distributed between employees and the enterprise in pre-established ratios. Let's call this ratio the coefficient of redistribution of the financial result between the labor collective and the management, owners, shareholders of the enterprise and denote it ξ. Then the financial result received by the enterprise will be redistributed between employees and the enterprise in the following proportions:

1) $Sal_j = (Rev_j / 12) \bullet \theta_{bj} + \xi \bullet (FR_j - FR_{bj}) / 12$ – this part of the income will be used to stimulate the work of employees.

2) Therefore, the amount of income that will be directed to the development of the enterprise is equal to $Rev_{devj} = (FR_{bj} / 12) + \xi \bullet [(FR_j - FR_{bj}) / 12] \bullet (1 - Tax_{prof})$.

From the formulas presented above, it is possible to deduce the dependence of the parameter θ_j on the revenue growth of the enterprise, i.e. to create a progressive system of material and

moral incentives for employees, in which the amount of remuneration directly depends on the revenue of the enterprise and at the same time the value of the percentage θ_j does not remain constant, but grows with the growth of revenue. This means that the increase in material incentives for employees occurs under the influence of a double effect: a) depending on the growth of revenue; b) on the change in the amount of interest directed to stimulating the work of employees. Thus, the amount of interest allocated to stimulate the work of employees has the following form: $\theta_j(Rev_j) = \frac{(Rev_j/12)\cdot\theta_{bj}+\xi\cdot(FR_j-FR_{bj})/12}{Rev_{bj}/12}$, where Rev_{bj} stands for the base revenue of the j-th division of the enterprise in rubles; θ_{bj} stands for the share of the basic income of the enterprise, which goes to stimulate the work of employees in fractions of units.

Combining all of the above, we will build an optimization model that belongs to the class of nonlinear programming models that maximizes the wages of employees with the growth of the company's income, integrated with income for its development (relevant for the employer and the entire workforce as a whole), taxation and contributions to off-budget funds):

Target Function

$$Sal = Rev \cdot \theta_b + \xi \cdot (FR - FR_b) \rightarrow max, \quad (1)$$

$$Rev_{dev} = FR_b + (1-\xi)\cdot(FR - FR_b)\cdot\left(1 - Tax_{prof}\right), \quad (2)$$

$$\theta = (Rev \cdot \theta_b + \xi \cdot FR)/Rev_b, \quad (3)$$

$$\Delta C = V \cdot \left(C_{var} + \frac{C_{fix}}{\sum_{i=1}^{n} V_i}\right) - V_b \cdot \left(C_{var} + \frac{C_{fix}}{\sum_{i=1}^{n} V_i}\right), \quad (4)$$

$$D_{PFR} = Sal \cdot \varphi_{PFR} + (Rev - C_{var}) \cdot VAT, \quad (5)$$

$$D_{FCMIF} = Sal \cdot \varphi_{FCMIF} + Sal \cdot Tax_{inc} + FR \cdot Tax_{prof}, \quad (6)$$

$$D = D_{PFR} + D_{FCMIF}, \quad (7)$$

$$\varphi_{FCMIF} = \varphi_{FCMIF5.1\%} - \Delta\varphi_{FCMIFstim} - \Delta\varphi_{FCMIFcost}, \quad (8)$$

$$\Delta\varphi_{FCMIFstim} = [(Sal - Sal_b)/Sal_b] \cdot \varphi_{FCMIF5.1\%}, \quad (9)$$

$$\Delta\varphi_{FCMIFcost} = [\Delta C/FR_b] \cdot \varphi_{FCMIF5,1\%}, \quad (10)$$

$$\varphi_{PFR} = \varphi_{PFR22.0\%} - \Delta\varphi_{PFRstim} - \Delta\varphi_{PFRTax}, \quad (11)$$

$$\Delta\varphi_{PFRstim} = [(Sal - Sal_b)/Sal_b] \cdot \varphi_{PFR22.0\%}, \quad (12)$$

$$\Delta\varphi_{PFRTax} = [(FR - FR_b)/Sal_b] \cdot \varphi_{PFR22,0\%}, \quad (13)$$

$$FR = Rev - V \cdot \left(C_{var} + \frac{C_{fix}}{\sum_{i=1}^{n} V_i}\right), \quad (14)$$

$$\omega_{fix} = \frac{\frac{C_{fix}}{\sum_{i=1}^{n} V_i}}{C_{var} + \frac{C_{fix}}{\sum_{i=1}^{n} V_i}}, \quad (15)$$

$$\omega_{var} = \frac{C_{var}}{C_{var} + \frac{C_{fix}}{\sum_{i=1}^{n} V_i}}, \quad (16)$$

when *Sal* stands for the salary of employees in rubles; Sal_b stands for the base salary of employees in the first case simulation in rubles; *Rev* stands for the revenue in rubles; θ stands for the share of the income of the enterprise, which goes to stimulate the work of employees in fractions of units; ξ stands for the coefficient of redistribution of the financial result between the labor collective and the management, owners, shareholders of the enterprise; ΔC stands for the unit cost reduction in rubles; Rev_{dev} stands for the value of the development fund in rubles; Rev_b stands for the base revenue in the first case simulation in rubles; θ_b stands for the share of the basic income of the enterprise, which goes to stimulate the work of employees in fractions of units; *V* stands for the number of products sold by the enterprise in units; V_b stands for the base number of products sold by the enterprise in the first case simulation in units; C_{var} stands for the unit variable costs in rubles; C_{fix} stands for the total fixed costs in rubles; $\sum_{i=1}^{n} V_i$ stands for the total number of products sold by the enterprise in units; *n* stands for the total number of divisions of the enterprise in units; D_{PFR} stands for the amount of value added tax and receipts to the PFR in rubles; D_{FCMIF} stands for the amount of tax on profit, income tax and receipts in FCMIF in rubles; *D* stands for the total deductions of the enterprise in rubles; φ_{PFR} stands for the rate contributions to the PFR in %; $\varphi_{PFR22.0\%}$ stands for the base rate of deductions to the PFR in the first case simulation, equal to 22.0% of the wage fund (WF) in %; $\Delta\varphi_{PFRstim}$ stands for the change in the rate of deductions to the PFR related to wage growth in %; $\Delta\varphi_{PFRTax}$ stands for the change in the rate of deductions to the PFR due to the increase in value added tax in %; φ_{FCMIF} stands for the rate of deductions to the FCMIF in %; $\varphi_{FCMIF5.1\%}$ stands for the base rate of deductions to the FCMIF in the first case simulation, equal to 5.1% of the wage fund (WF) in %; $\Delta\varphi_{FCMIFstim}$ stands for the change in the rate of deductions in FCMIF related to wage growth in %; $\Delta\varphi_{FCMIFcost}$ stands for the change in the rate of deductions to the FCMIF due to the effect of cost reduction in %; *FR* stands for the financial result or profit in rubles; FR_b stands for the base financial result or base profit in the first case simulation in rubles; *VAT* stands for the rate of value added tax (VAT), 20%; Tax_{prof} stands for the tax on profit rate, 20%; Tax_{inc} stands for the income tax rate, 13%; ω_{var} stands for the share of variable costs in fractions of units; ω_{fix} stands for the share of fixed costs in fractions of units.

The main limitations, prerequisites and assumptions used in the development and practical implementation of the optimization model (1)-(16):

1) personalized pension accounts and medical savings accounts are used as a source of financing for pension provision and healthcare of citizens working at the enterprise, which means that part of the social contributions that the enterprise makes to the FCMIF and the PFR go to medical savings accounts and personalized accounts of citizens;
2) information about the cost structure, operating and investment activities, as well as financial results provided by the management and employees of the analyzed enterprise is correct, complete and reliable;

3) any hidden (not explicitly stated) factors will not have a significant impact on the company and the results of the practical application of the model (1)-(16);
4) the company is active and will continue its business activities in the foreseeable future;
5) in the future, the responsible attitude of the owners of the enterprise and the competent management of its operational activities will remain;
6) the company will comply with all applicable provisions of laws and regulations, especially in terms of taxation and social contributions;
7) the enterprise has, will receive or will extend all necessary permits and licenses on which the functioning of the enterprise and the applicability of the economic and mathematical model developed by the authors are based (1)-(16);
8) all cash flows received from operating activities occur during the same year to which the corresponding income received and expenses incurred relate.

Figure 2 shows a flowchart of the integrated system of social financing of enterprises, which reflects in detail the main aspects and criteria for making key management decisions and shows the optimal strategy for the practical implementation of the optimization model (1)-(16).

A more detailed algorithm and tools for the practical application of social technologies for financing the Russian economy and enterprises are given in the section "Results".

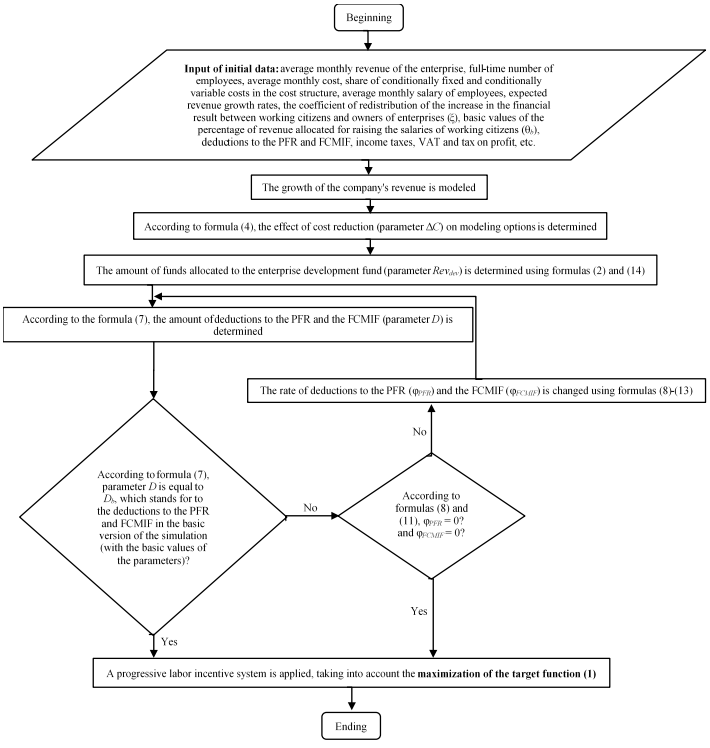


Figure 2: Block Diagram of the Complex System of Social Financing of Enterprises

Results

Table 3 shows the practical implementation of the optimization model (1)-(16). Column 1 shows the number of the simulation variant, which corresponds to the year shown in columns 2 and 15 of Table 3. The first simulation variant is the basic one and corresponds to the ***2021st year, in which the basic values of the parameters being changed are indicated***. The salary according to the Federal State Statistics Service for November 2023 is 63,060 rubles (see row 1, columns 10 and 13 of Table 3). The share of wages in the structure of Russia's gross domestic product (GDP) is 44.9%, and the profitability of products sold is 9.9%. Formula (1) establishes the basic parameters and the relationship between them, namely: wages (*Sal*), enterprise income (*Rev*) and the share of the income of the enterprise, which goes to stimulate the work of employees. Since $FR - FR_b = 0$ in the first case simulation, it is possible to determine the revenue of enterprises by dividing wages by the share of wages in the GDP structure ($Rev = Sal : \theta_b$), i.e. 63,060 rubles : 0.449 (44.9%) = 140,445.43 rubles (see column 3 of Table 3).

The cost of production is equal to the revenue of enterprises minus the profit from the sale of products, which is 140,445.43 rubles · 0.099 = 13,904.10 rubles. This means that the cost price will be: 140,445.43 rubles – 13,904.10 rubles = 126,541.34 rubles (see the first row, column 4 of Table 3). Next, we increase the company's revenue by 3% per year, which corresponds to the growth rate of average wages in [95]. As shown above, the revenue of enterprises is equal to the amount of wages divided by its share in the structure of GDP. So, for the second row of Table 1, the income is 63,060 rubles • 1.03 (average wage growth rate) : 0.449 = 144,658.80 rubles, for the tenth modeling option 63,060 rubles • 1.039 : 0.449 = 183,249.44 rubles, etc. In the latest version of the simulation, the revenue of enterprises per employee exceeds the basic version of the simulation by 4.38 times.

In the production of products, the total costs are divided into fixed (63,035.89 rubles, see column 7), which do not depend on the volume of products, and variable (42,023.92 rubles in the basic version of modeling, see column 8 of Table 3). At the same time, the share of fixed costs is equal to 60% of total costs in the basic version of modeling (ω_{fix} = 60%, see line 1 of Table 3), and the share of variables is 40% of total costs (ω_{var} = 40%). With an increase in production volumes, unit costs decrease, and the share of fixed costs in the cost structure also decreases, which makes it possible to direct the released financial resources to additional material incentives for employees, i.e. to introduce a progressive system of material incentives for labor (see column 13 of Table 3). The meaning of this system is that with an increase in labor productivity increases the share of the income of the enterprise, which goes to stimulate the work of employees (θ), as shown in column 11 of Table 3. The share of fixed costs (ω_{fix}) is determined by the formula (15), and the share of variable costs (ω_{var}) is determined by the formula (16). Variable costs depend on the quantity of products produced, while fixed costs do not.

The progressive system of material remuneration (objective function (1)) makes it possible to increase wages faster (see column 13) compared to the data in column 10 of Table 3, namely: more than eight times compared to the basic modeling option at annual growth rates of the volumes of products, services and works produced and sold, equal to 3.0% (see column 14 of Table 3).

Table 3: Progressive Labor Incentive System

Option number	Year	The revenue of enterprises per employee with a share of wages in its structure of 44.9% in the first (basic) version of the simulation, rubles per month	The cost of production per employee with a profitability of 9.9% in the first (basic) version of the simulation, rubles per month	The share of fixed costs	The share of variable costs	Total fixed costs, rubles per month	Total variable costs, rubles per month	The effect of cost reduction, rubles per month	The average monthly nominal accrued salary, rubles per month	The share of the company's income allocated to wages	Wage growth, rubles per month	Innovative technology of material remuneration of employees, rubles per month	Wage index
1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	2023	140,445.43	126,541.34	60.00%	40.00%	75,924.80	50,616.53	0.00	63,060	44.90%	0.00	63,060	1.000
2	2024	144,658.80	128,059.83	59.29%	40.71%	75,924.80	52,135.03	2,277.74	64,952	45.40%	2,155.89	67,108	1.064
3	2025	148,998,56	129,623.88	58.57%	41.43%	75,924.80	53,699.08	4,623.82	66,900	45.90%	4,376.46	71,277	1.130
4	2026	153,468.52	131,234.86	57.85%	42.15%	75,924.80	55,310.05	7,040.28	68,907	46.40%	6,663.65	75,571	1.198
5	2027	158,072.57	132,894.16	57.13%	42.87%	75,924.80	56,969.36	9,529.23	70,975	46.90%	9,019.45	79,994	1.269
6	2028	162,814.75	134,603.24	56.41%	43.59%	75,924.80	58,678.44	12,092.85	73,104	47.40%	11,445.93	84,550	1.341
7	2029	167,699.19	136,363.59	55.68%	44.32%	75,924.80	60,438.79	14,733.38	75,297	47.90%	13,945.20	89,242	1.415
8	2030	172,730.17	138,176.75	54.95%	45.05%	75,924.80	62,251.95	17,453.13	77,556	48.40%	16,519.45	94,075	1.492
9	2031	177,912.07	140,044.31	54.21%	45.79%	75,924.80	64,119.51	20,254.47	79,883	48.90%	19,170.93	99,053	1.571
10	2032	183,249.44	141,967.90	53.48%	46.52%	75,924.80	66,043.10	23,139.84	82,279	49.40%	21,901.95	104,181	1.652
11	2033	188,746.92	143,949.19	52.74%	47.26%	75,924.80	68,024.39	26,111.78	84,747	49.90%	24,714.90	109,462	1.736
12	2034	194,409.33	145,989.92	52.01%	47.99%	75,924.80	70,065.12	29,172.88	87,290	50.40%	27,612.24	114,902	1.822
13	2035	200,241.61	148,091.88	51.27%	48.73%	75,924.80	72,167.08	32,325.81	89,908	50.90%	30,596.51	120,505	1.911
14	2036	206,248.86	150,256.89	50.53%	49.47%	75,924.80	74,332.09	35,573.33	92,606	51.40%	33,670.29	126,276	2.002
15	2037	212,436.32	152,486.85	49.79%	50.21%	75,924.80	76,562.05	38,918.27	95,384	51.90%	36,836.30	132,220	2.097
16	2038	218,809.41	154,783.71	49.05%	50.95%	75,924.80	78,858.91	42,363.57	98,245	52.40%	40,097.28	138,343	2.194
17	2039	225,373.69	157,149.48	48.31%	51.69%	75,924.80	81,224.68	45,912.22	101,193	52.90%	43,456.09	144,649	2.294
18	2040	232,134.90	159,586.22	47.58%	52.42%	75,924.80	83,661.42	49,567.33	104,229	53.40%	46,915.67	151,144	2.397
19	2041	239,098.95	162,096.06	46.84%	53.16%	75,924.80	86,171.26	53,332.09	107,355	53.90%	50,479.03	157,834	2.503
20	2042	246,271.92	164,681.20	46.10%	53.90%	75,924.80	88,756.40	57,209.80	110,576	54.40%	54,149.30	164,725	2.612
21	2043	253,660.08	167,343.89	45.37%	54.63%	75,924.80	91,419.09	61,203.84	113,893	54.90%	57,929.67	171,823	2.725
22	2044	261,269.88	170,086.47	44.64%	55.36%	75,924.80	94,161.66	65,317.69	117,310	55.40%	61,823.45	179,134	2.841
.....													
50	2072	597,766.59	291,359.88	26.06%	73.94%	75,924.80	215,435.08	247,227.82	268,397	69.40%	234,002.09	502,399	7.967
51	2073	615,699.58	297,822.93	25.49%	74.51%	75,924.80	221,898.13	256,922.39	276,449	69.90%	243,178.04	519,627	8.240

Table 4: Progressive system of reducing the tax burden on enterprises

Option number	Year	Growth of the development fund, rubles per month	Receipts to the development fund, rubles per month	The rate contributions to the PFR	Index of deduc-tions to the PFR	The rate of deductions to the FCMIF	Index of deduc-tions to the FCMIF	Income tax rate	Tax on profit rate	VAT rate	Financial result, rubles per month	Increase in VAT deductions, rubles	Deductions to the PFR + VAT, rubles per month	Deductions to the FCMIF + income tax + tax on profit rate, rubles per month	Deductions to the PFR + deductions to the FCMIF + income tax +tax on profit rate + VAT, rubles per month
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	2023	0.00	13,904.10	22.00%	1.00	5.10%	1.00	13.00%	20.00%	20.00%	13,904.10	0.00	31,838.98	14,194.68	46,033.66
2	2024	431.18	14,335.28	19.87%	0.90	3.88%	0.76	13.00%	20.00%	20.00%	14,335.28	538.97	31,838.98	14,194.68	46,033.66
3	2025	875.29	14,779.39	17.93%	0.81	2.77%	0.54	13.00%	20.00%	20.00%	14,779.39	1,094.12	31,838.98	14,194.68	46,033.66
4	2026	1,332.73	15,236.83	16.15%	0.73	1.75%	0.34	13.00%	20.00%	20.00%	15,236.83	1,665.91	31,838.98	14,194.68	46,033.66
5	2027	1,803.89	15,707.99	14.52%	0.66	0.82%	0.16	13.00%	20.00%	20.00%	15,707.99	2,254.86	31,838.98	14,194.68	46,033.66
6	2028	2,289.19	16,193.28	13.02%	0.59	0.00%	0.00	13.00%	20.00%	20.00%	16,193.28	2,861.48	31,838.98	14,230.13	46,069.10
7	2029	2,789.04	16,693.14	11.64%	0.53	0.00%	0.00	13.00%	20.00%	20.00%	16,693.14	3,486.30	31,838.98	14,940.11	46,779.09
8	2030	3,303.89	17,207.99	10.36%	0.47	0.00%	0.00	13.00%	20.00%	20.00%	17,207.99	4,129.86	31,838.98	15,671.39	47,510.37
9	2031	3,834.19	17,738.28	9.17%	0.42	0.00%	0.00	13.00%	20.00%	20.00%	17,738.28	4,792.73	31,838.98	16,424.61	48,263.59
10	2032	4,380.39	18,284.49	8.06%	0.37	0.00%	0.00	13.00%	20.00%	20.00%	18,284.49	5,475.49	31,838.98	17,200.42	49,039.40
11	2033	4,942.98	18,847.08	7.03%	0.32	0.00%	0.00	13.00%	20.00%	20.00%	18,847.08	6,178.73	31,838.98	17,999.51	49,838.49
12	2034	5,522.45	19,426.55	6.07%	0.28	0.00%	0.00	13.00%	20.00%	20.00%	19,426.55	6,903.06	31,838.98	18,822.57	50,661.55
13	2035	6,119.30	20,023.40	5.16%	0.23	0.00%	0.00	13.00%	20.00%	20.00%	20,023.40	7,649.13	31,838.98	19,670.33	51,509.31
14	2036	6,734.06	20,638.16	4.32%	0.20	0.00%	0.00	13.00%	20.00%	20.00%	20,638.16	8,417.57	31,838.98	20,543.52	52,382.50
15	2037	7,367.26	21,271.36	3.53%	0.16	0.00%	0.00	13.00%	20.00%	20.00%	21,271.36	9,209.07	31,838.98	21,442.90	53,281.88
16	2038	8,019.46	21,923.55	2.78%	0.13	0.00%	0.00	13.00%	20.00%	20.00%	21,923.55	10,024.32	31,838.98	22,369.26	54,208.24
17	2039	8,691.22	22,595.32	2.08%	0.09	0.00%	0.00	13.00%	20.00%	20.00%	22,595.32	10,864.02	31,838.98	23,323.42	55,162.40
18	2040	9,383.13	23,287.23	1.42%	0.06	0.00%	0.00	13.00%	20.00%	20.00%	23,287.23	11,728.92	31,838.98	24,306.20	56,145.18
19	2041	10,095.81	23,999.90	0.79%	0.04	0.00%	0.00	13.00%	20.00%	20.00%	23,999.90	12,619.76	31,838.98	25,318.46	57,157.44
20	2042	10,829.86	24,733.96	0.20%	0.01	0.00%	0.00	13.00%	20.00%	20.00%	24,733.96	13,537.32	31,838.98	26,361.09	58,200.07
21	2043	11,585.93	25,490.03	0.00%	0.00	0.00%	0.00	13.00%	20.00%	20.00%	25,490.03	14,482.42	32,448.20	27,435.00	59,883.20
22	2044	12,364.69	26,268.79	0.00%	0.00	0.00%	0.00	13.00%	20.00%	20.00%	26,268.79	15,455.86	33,421.64	28,541.13	61,962.77
.....															
50	2072	46,800.42	60,704.52	0.00%	0.00	0.00%	0.00	13.00%	20.00%	20.00%	60,704.52	58,500.52	76,466.30	77,452.81	153,919.11
51	2073	48,635.61	62,539.71	0.00%	0.00	0.00%	0.00	13.00%	20.00%	20.00%	62,539.71	60,794.51	78,760.29	80,059.47	158,819.76

The developed optimization model makes it possible to increase the salaries of employees, which is beneficial to both the state and shareholders, managers and owners of enterprises. The algorithm of practical implementation of the optimization model is as follows.

Formula (14) is used to calculate the financial result, which is given in column 12 of Table 4, and formula (4) is used to calculate the effect of reducing unit costs (see column 9 of Table 3).

Formula (5) is used to calculate receipts to the VAT and PFR (see column 14 of Table 4), and formula (6) is used to determine the values of column 15 of Table 4. To calculate the data presented in column 16 of Table 4, formula (7) is used. The data in column 16 of Table 4 are equal to the sum of the values in the columns 14 and 15 of Table 4.

Formula (3) is used to calculate the share of the company's income, which is used to increase the salaries of employees. Formulas (1) and (2) are used to calculate the increase in wages and the development fund (column 12 of Table 3 and column 3 of Table 4, respectively). The coefficient ξ is equal to 0.8, which means that 80% of the increase in the financial result of enterprises from the sale of products is directed to the salaries of employees, and 20% is directed to the development fund. This distribution of the financial result is taken from the ratio of wages and financial result in the basic version (63,060 rubles : 13,904.10 rubles = 4.5). We will show an algorithm for calculating deductions for salary increases and to the entrepreneurship development fund using the example of the second row of column 12 of Table 3 and column 3 of Table 4. According to the formula (14), the financial result is 144 658.80 rubles (the company's revenue per employee, see row 2, column 3 of Table 3) – 128 059.83 rubles (the cost of production per employee) = 16,598.96 rubles. The increase compared to the basic version of the simulation is 2,694.86 rubles = 16,598.96 rubles – 13,904.10 rubles (the financial result in the basic version of the simulation, see row 1, column 12 of Table 4). Then for the second row of column 12 of Table 3, the wage increase according to the formula (1) is equal to 2,155.89 rubles. = 2,694.86 rubles (increase in financial result relative to the base value) • 0.8 (coefficient of redistribution of financial result between employees and owners, managers, shareholders of enterprises), and an increase in contributions to the development fund (see the second row, column 3 of Table 4) is 431.18 rubles = 2,694.86 rubles (increase in financial result relative to the base value) • (1 – 0.8 (coefficient of redistribution of the increase in financial result between employees and managers of enterprises) • 0.8 (adjustment of the income tax rate, 20%, see formula (2)). Similarly, for the other rows of column 12 of Table 3 and column 3 of Table 4. For the 51st variant of the simulation, the growth of the development fund amounted to 48,635.61 rubles per employee (see the last row of column 3 of Table 4), which is more than the basic option development fund [(48,635.61 rubles : 13,904.10 rubles) • 100% = 350%] by 350%.

In column 4 of Table 4 the average monthly contributions to the development fund are given, determined by adding an increase in contributions to the development fund to the base amount of the financial result (see formula (2)). So, for the second row of column 4 of Table 4 the value of 14,335.28 rubles = 13,904.10 rubles (see the first row, column 4 of Table 4) + 431.18 rubles (see the second row, column 3 of Table 4). Similarly, for the remaining rows of column 4 of Table 4. In column 5 of Table 4 shows the size of the deduction rate in the PFR, which are determined by the formula (11) of the economic and mathematical model (1)-(16) and include the effect of wage growth and VAT. The VAT rate is 20%, the value is shown in column 11 of Table 4. The article [95] shows that pension funding is carried out from two sources: deductions from the wages of employees (22%) and the federal budget (for 2023 – 4,822.23 rubles on average per pensioner). It was emphasized above that wage growth increases VAT and pension contributions and is not beneficial to the labor collective and the owner.

This article proposes a mechanism of state regulation that encourages enterprises to increase wages, which consists in reducing pension contributions and deductions to the FCMIF depending on wage growth, but at the same time does not allow a reduction in the basic amount of deductions for VAT, pension provision, income tax, income tax and deductions to the FCMIF.

According to the second term of the formula (5) of the economic and mathematical model (1)-(16) for the basic (first variant, Table 4) the value of VAT is equal 17,965.78 rubles = (140,445.43 rubles (the average monthly revenue of enterprises, see the first row, column 3 of Table 3) – 50,616.53 rubles (conditionally variable costs, see the first row, column 8 of Table 3)) · 0.20 (VAT rate). Further, according to the first term of formula (5), deductions to the PFR are calculated as the product of wages at the rate of deductions to the PFR (63,060 rubles · 0.22 = 13,873.20 rubles). Total deductions, according to formula (5), will be 17,965.78 + 13,873.20 rubles = 31,838.98 rubles (first row, column 14 of Table 4). On the example of the second row of column 14 of Table 4 we will show a ***detailed methodology for calculating*** the rate of monthly deductions to the PFR and VAT, taking into account the reduction in the rate of deductions to the PFR (column 5 of Table 4):

Step 1. We determine the amount of monthly deductions to the PFR and VAT (formula (5) of the economic and mathematical model (1)-(16)). To do this, we will calculate deductions to the PFR and VAT at the base rate of deductions to the PFR (22.0%). And in the future, we will reduce the deduction rate in proportion to the increased amount of deductions so that in row 2, column 14 of Table 4 get an irreducible amount 31,838.98 rubles, i.e. as in the basic version of the simulation.

Step 2. We calculate the amount of deductions to the PFR at a rate of 22.0% (see the first term of formula (5)): 67,108 rubles (the average monthly nominal accrued salary, taking into account the progressive labor incentive system, see row 2, column 13 of Table 3) • 0.22 = 14,763.76 rubles.

Step 3. Calculate the amount of VAT deductions. According to the second term formula (5) the amount of VAT deductions is equal to (144,658.80 rubles (the average monthly revenue of enterprises, see the second row, column 3 of Table 3) – 52,135.03 rubles (conditionally variable costs of enterprises in the sale of products, services and works, see the second row, column 8 of Table 3)) • 0.2 (VAT deduction rate) = 18,504.75 rubles.

Step 4. According to formula (5), to the value obtained in step 3 of the algorithm, we add the amount of deductions to the PFR from wages at a rate of 22.0%, we have 18,504.75 rubles + 14,763.76 rubles = 33,268.51 rubles.

Step 5. To calculate the rates of deductions to the PFR, the condition is accepted that the amount of monthly deductions to the PFR and VAT must be at least the amount of these deductions in the basic version of the simulation. Thus, the rate of deductions to the PFR can be reduced with an increased salary level so that deductions to the PFR and VAT amount to 31,838.98 rubles (see row 1, column 14 of Table 4). Then, according to the formula (5) of the economic and mathematical model (1)-(16) the amount of deductions to the PFR it should be 31,838.98 rubles – 18,504.75 rubles (total VAT deductions according to step 3 of the algorithm) = 13,334.23 rubles. In other words, the product of increased wages (67,108 rubles for the second row of column 13 of Table 3) the rate of deductions to the PFR should bring at least 13,334.23 rubles, hence we get the rate of deductions to the PFR in the amount of 13,334.23 rubles : 67,108 rubles • 100% = 19.87%, which is indicated in the second row of column 5 of Table 4. Similarly, for all other rows of column 19 of Table 3.

In other words, with this approach, PFR and the federal budget receive an irreducible amount of 31,838.98 rubles per month from each employee, and the company's deductions to the PFR for modeling options are reduced in accordance with formulas (11)-(13) from 22.00% in the basic version to 0.00% in the 21st version of modeling, which corresponds to the 2043st year (see line 21, column 5 of Table 4). Further, the rate of deductions to the PFR remains unchanged and equal to 0.00% (see lines 21-51, column 5 of Table 4). To calculate the rates of deductions to the PFR, the condition of not reducing the amount of monthly deductions due to VAT and deductions to the PFR (formula (5) of the model) was also taken into account, at least the amounts necessary to accumulate funds for retirement for the period of survival (see Table 3-6 [95]), as can be seen from the analysis of the data presented in column 14 of Table 4. Thus, for column 5 of Table 4 the reduction in the rates of deductions to the PFR occurs due to wage growth and the redistribution of VAT funds from the increased amount of wages of working citizens and VAT from the increased volume of manufactured and sold products, goods, works and services to personalized pension accounts of citizens in banks (see formulas (11)-(13)).

Despite the reduction in the rate of deductions to the PFR, the financing of pension provision is not reduced, since it is fully compensated by the increase in VAT deductions received by the federal budget and directed to personalized pension accounts of employees in banks.

Let's look at this process in more detail. In column 13 of Table 4 shows an increase in deductions from VAT in the PFR, compensating for a decrease in the rate of deductions from wages in the PFR (column 5 of Table 4). In the second version of the simulation (2024), deductions from VAT in the PFR will amount to 538.97 rubles (row 2, column 13), in the tenth version of the simulation 5,475.49 rubles, in 20 option (2042) 13,537.32 rubles (row 20, column 13 of Table 4). In all these variants (from 1 to 20), deductions to the PFR and VAT are constant (31,838.98 rubles, column 14) and correspond to the basic variant. But already in 21 variants (2043) with the rate of deductions to the PFR equal to 0 (row 21, column 5 of Table 4) deductions to the PFR and VAT begin to increase 32,448.20 (line 21, column 14 of Table 4) and to 51 variants in accordance with formula 5 of the economic and mathematical model (1)-(16) amount to 78,760.29 rubles, exceeding the basic modeling variant in (78,760.29 : 31,838.98 = 2.47) 2.47 times.

It should be emphasized that Tables 3 and 4 is suitable only for citizens who started working in 2023. For those who have already worked, the table is shifted a year ago. That is, for those who have already worked for 1 year, the calculation is carried out by 2023 and zero deductions to the PFR will begin from 2044. For those who have been working for 5 years before the implementation of social financing technologies, zero deductions will begin from 2047 and so on.

Thus, the proposed mechanism contributes to the growth of wages of the labor collective, the growth of deductions for development and the growth of revenues to PFR and the federal budget.

Contributions to the FCMIF are formed as follows. According to the formula (6) for the basic (first variant, Table 4) the amount of income tax is determined (63,060 rubles • 0.13 (income tax rate, see column 9 of Table 4) = 8,197.80 rubles, income tax (13,904.10 (financial result, see column 12 of Table 4) · 0.20 (income tax rate, see column 10 of Table 4) = 2,780.82 rubles) and deductions to the FCMIF (63,060 rubles · 0.051 = 3,216.06 rubles), which in total will be: 8,197.80 rubles + 2,780.82 rubles + 3,216.06 rubles = 14,194.68 rubles (first row, column 15 of Table 4).

In other words, with this approach, the territorial budget and the FCMIF receive an irreducible amount of 14,194.68 rubles per month from each employee, and the company's deductions to the FCMIF for modeling options are reduced in accordance with formulas (8)-(10) from 5.10% in the basic version to 0.00% in the 6th version of modeling, which corresponds to the year 2028 (see row 6, column 7 of Table 4). Further, the rate of deductions to the FCMIF remains unchanged and equal to 0.00% (see rows 6-51, column 7 of Table 4). To calculate the rates of deductions to the FCMIF, the condition of not reducing the amount of monthly deductions due to income tax, income tax and deductions to the FCMIF (formula (6) of the model) at least the amount of monthly deductions in the basic version of the simulation was also taken into account, as can be seen from the analysis of the data presented in column 15 of Table 4. Thus, for column 7 of Table 4, the reduction in the rates of deductions to the FCMIF occurs due to wage growth and the redistribution of income tax funds from the increased amount of wages of working citizens and income tax from the increased volume of manufactured and sold products, goods, works and services to personalized medical savings accounts of citizens in banks.

It should be emphasized that despite the reduction in the rate of deductions to the FCMIF, the financing of medical care is not reduced, since it is fully compensated by the increase in income tax and income tax deductions received by the territorial budget and directed to personalized medical savings accounts of working citizens in banks.

Table 5: The Economic Effect

Line number	Year	The growth index of the revenue of the enterprise	Revenue of enterprises per employee, rubles	Wage growth index	Salary + deductions from wages, rubles	Annual wage growth, thousand rubles	Average annual contributions to the development fund, thousand rubles	The rate of deductions to the PFR,	The rate of deductions to the FCMIF
1	2	3	4	5	6	7	8	9	10
0	2023	1.00	140,445.43	1.000	63,060	62,564,096,160	13,794,756,169	22.00%	5.10%
1	2024	1.03	144,658.80	1.064	67,108	66,579,958,692	14,222,544,098	19.87%	3.88%
2	2025	1.06	148,998,56	1.130	71,277	70,716,297,100	14,663,165,666	17.93%	2.77%
3	2026	1.09	153,468.52	1.198	75,571	74,976,725,661	15,117,005,880	16.15%	1.75%
4	2027	1.12	158,072.57	1.269	79,994	79,364,967,078	15,584,461,301	14.52%	0.82%
5	2028	1.15	162,814.75	1.341	84,550	83,884,855,738	16,065,940,384	13.02%	0.00%

And the total deductions due to income tax, income tax and deductions to the FCMIF (column 15 of Table 4, formula (6) of the economic and mathematical model (1)-(16)), despite the reduction in the rate of deductions to the FCMIF, increase from 14,194.68 rubles (in the basic version and variants 2 to 5), starting from option 6 (14,230.13 rubles) to 80,059.47 rubles in the 51st version of the simulation (***5.64 times***).

Thus, the proposed mechanism contributes to the growth of wages of the labor collective, the growth of deductions for development and the growth of revenues to FCMIF.

It is worth noting that the total deductions due to income tax, income tax, VAT, deductions to the FCMIF and the PFR (see formula (7)) increase from 46,033.66 rubles in the basic version of the simulation to 158,819.76 rubles in the 51st version of the simulation, i.e. 158,819.76 rubles : 46,033.66 rubles = 3.45 times (see column 16 of Table 4).

The rate of deductions to the FCMIF becomes zero already for the salary of a working citizen of Russia in the amount of 84,550 rubles, which is 34.08% higher than the average salary in the country as of November 2022 (84,550 rubles : 63,060 rubles = 1,341), as can be seen by comparing the data presented in line 6, columns 13 of Table 3 and column 7 of Table 4.

Discussion

The use of a progressive labor incentive system and the reduction of deductions to the PFR and the FCMIF with a corresponding increase in revenue from VAT, income tax and profit tax is beneficial to working citizens, business owners and the state. The economic effect for working citizens, owners of enterprises and the state from an annual increase in GDP (revenue of enterprises) by 3.0% over the first five years of the introduction of social financing technology is presented in Table 5.

Column 1 of Table 5 shows the year number, and column 2 of Table 5 shows the year. The first row (year zero) of Table 5 corresponds to the basic modeling variant (2023), i.e. row 1 of Tables 3 and 4. It is assumed that the annual revenue of enterprises will increase annually for 5 years by 3.0% from the baseline (row 1 of Tables 3 and 4). The data presented in columns 2-6, 9 and 10 are taken from similar columns of Tables 3 and 4. Annual wage growth (62,564,096,160 thousand rubles, see the first row, column 7 of Table 3) calculated as the product of wages (63,060 rubles, see row 1, column 13 of Table 3), by 12 (the number of months in a year) and by the population of working age according to Rosstat (82,678 thousand people according to Rosstat). With an increase in the revenue of enterprises by 3.0%, wage growth is 6.4% (see the first year, second row, column 5 of Table 5). With an increase in revenue by 3.0% per year for five years, wage growth per employee is 34.1% (see the last row, column 5 of Table 5), and the cumulative annual wage growth for all workers in Russia is 83,884,855,738 thousand rubles (see the last row, column 7 of Table 5). Thus, the wage increase is almost 20 trillion rubles.

Average annual contributions to the development fund (column 8 of Table 5) are determined by multiplying the amount of contributions to the development fund (column 4 of Table 4) by 12 (the number of months in a year) and by 82,678 thousand people (the population of working age). So, for the basic variant (year zero, the first row of column 8 of Table 5) the value of 13,794,756,169 thousand rubles = 13,904.10 rubles • 12 • 82,678 thousand people. For the second row, column 8 of Table 5 value 14,222,544,098 thousand rubles = 14,335.28 rubles • 12 • 82,678 thousand people. For the owners of enterprises, the use of social technologies developed in this article for financing enterprises leads to an increase in the size of average monthly contributions to the development fund to 16,193.28 rubles (see the sixth row, column 4 of Table 4) and this is only from one employee at the enterprise. There are 82,678 thousand citizens of working age in Russia. Thus, with an increase in the revenue of enterprises by 15%, the average annual contributions to the development fund using social technologies for financing enterprises are equal to 16,193.28 rubles • 12 • 82,678 thousand people = 16,065,940,384 thousand rubles (see the last row, column 8 of Table 5), the growth is 16,065,940,384 thousand rubles : 13,794,756,169 thousand rubles = 1.16 times, i.e. average annual contributions to the development fund increased by 16% compared to the base variant, the zero year.

An important result of the use of social financial technologies is the reduction of the deduction rate in the PFR for 5 years to 13.02% and in the FCMIF to 0 for workers with average wages

who started working in 2023 (columns 9 and 10 of Table 5), which allows owners to significantly reduce social contributions and, accordingly, reduce the cost of all manufactured in Russia, goods, works and services.

Conclusion

The proposed technology of financing enterprises and the Russian economy, harmoniously combining the interests of working citizens, owners and the state, makes it possible:

1. At quite achievable rates of GDP growth (enterprise revenue) by 3% per year, ensure a 34% increase in wages of working citizens over 5 years (see the last row, column 5 of Table 5), ***which will practically end poverty***. Under the current funding system, this has not been done in 30 years.

2. To ensure in four years the level of pension provision for current and future pensioners in the amount of 40% of wages; in 8 years – 60% of wages; in 10 years – 80% of wages. Under the current funding system, this has also not been done in 30 years.

3. To increase contributions to the development fund for 5 years by 16% while reducing social contributions (PFR and FCMIF by 14.08%), which, first of all, the owners of enterprises are interested in, since this ensures the growth of their incomes, a significant reduction in cost and the possibility of constant modernization and updating of technological equipment and the release of new competitive products. In other words, if the owner motivates employees by increasing wages to increase sales volumes, then funds for development will grow at a higher rate than revenue. If the owner takes all the profits for himself, as is currently being done at many enterprises, then he will not be able to increase the volume of sales without motivation due to wage growth of working citizens, and, consequently, there will be much less funds for development. Thus, the motivation of employees by increasing wages, and owners by reducing social contributions is extremely beneficial for the owner. It is also important that wage growth, rigidly linked to an increase in product sales, stimulates the entire workforce to develop the enterprise. In other words, not only the owner and senior management, but the entire workforce becomes interested in the development of their enterprise.

4. First, to stabilize, and starting from 2028, to increase deductions to the PFR and the FCMIF and income tax, income tax and VAT receipts and bring this growth to 30% by 2043, which will allow the state to solve many social problems. Reduce deductions to the PFR and FCMIF by 14.08% by 2028, which will significantly reduce the cost of products, services and works of all enterprises in Russia.

5. The wage growth provided by the proposed social financial technologies contributes to the growth of the purchasing power of citizens of the relevant region (stimulates demand), and this, in turn, allows enterprises to increase products sales, which together ensures their development and subsequent wage growth of workers.

6. At the beginning of this article, it was shown that in social relations, the main thing is not to distribute, but to create, and that all goods, works and services are produced in the process of labor at enterprises. Therefore, state (public) funds should, first of all, perform the function of enterprise development, and only then will the incomes of working citizens grow and there will be funds for social support. The reduction in the rate of deductions to the PFR proposed in this article at wages above a certain level is, as shown in this article, an effective tool to support the development of enterprises. Why drive money up in the form of taxes and social contributions, and then bring it to enterprises and citizens, creating the ground for corruption along the way, when they can immediately be sent to PAPA and medical bills and to enterprises in the form of a reduction in the rate of deductions to the PFR and FCMIF.

The developed optimization model (1)-(16) allows:

1. To determine the optimal ratios of the main economic indicators of the company's division, such as revenue, the price of products, goods, works, services, discounts on them, which are provided to increase demand for the company's products, cost and profit, as well as to provide a progressive system of material incentives for employees involved in the business processes of the enterprise, in depending on the growth or decline of the specified economic indicators of the unit.

2. To create the necessary economic mechanisms for moral and material encouragement of employees of the enterprise to increase the quality and effectiveness of their work, to involve them in the process of managing the activities of the enterprise, to create incentives for the formation of responsibility for the results of work in the workplace.

3. To form long-term work plans of divisions and the enterprise as a whole, to identify deviations in the work of divisions in order to be able to correct them in a timely manner and provide employees with a clear and understandable information base for self-improvement, self-development, formation of independence and responsibility for decisions made.

4. Calculate the size of consumption funds and accumulation funds. Determine, depending on the results of the work, the amount of material remuneration and the amount of deductions for the development of the enterprise, from which sources can be formed in the future for updating the material and technical base of the division, improving the skills of personnel, increasing the capacity of the division of the enterprise and (or) the equipment of the enterprise, the purchase of new technological and auxiliary equipment and necessary consumables for it, which will further contribute to the growth of the volume of manufactured and sold products of the enterprise, goods, works, services, reducing their cost and, ultimately, increasing the profit of the enterprise.

5. To create additional sources of formation and strengthening of the financial system of Russia, since, as mentioned above, working citizens are the basis of the financial system of Russia. This means that the stability of the entire financial system depends on the organized and well-motivated work of working citizens, and the developed economic and mathematical model is a tool for managing the work of employees, increasing their material and moral remuneration depending on the results of work. At the same time, a distinctive feature of the developed staff motivation system is the dependence of the percentage of deductions from income for

stimulating the work of employees on the amount of this income, which makes it possible to use a progressive scale of material incentives.

The developed economic and mathematical model and the system of personnel motivation, which contributes to the growth of the company's revenue, as well as the system of progressive material and moral incentives for employees is a very important source of their development for enterprises.

An essential factor in the proposed system of material and moral incentives for personnel is their direct participation in the decision-making process, together with managers and owners of enterprises, on the acquisition of technological and auxiliary equipment that improves the quality of products, goods, works, services and the throughput capacity of the enterprise division and (or) equipment from the funds accumulated in each division for its development. Such participation morally stimulates each employee of the enterprise to continuously improve their professional qualifications, increases the prestige and demand for their work and at the same time contributes to the development of this division and the enterprise as a whole.

The developed economic and mathematical model of material incentives for employees, consistent with the growth of the company's revenue, allows you to increase deductions for material incentives for employees, as well as deductions for the development of the enterprise. The indisputable practical significance of the model is the functional scientifically based relationship between the structure of the cost of manufactured and sold products, goods, works, services, financial incentives for employees involved directly in the business processes of the enterprise, the amount of deductions to the development fund and revenue, which allows the model to be used for the analysis and development of optimal long-term work plans of divisions and the enterprise as a whole.

Progressive financial incentives lead to an increase in deductions for the development of the enterprise. These funds, allocated for the purchase of advanced technological and auxiliary equipment and staff training, significantly improve the quality of the company's products and contribute to the influx of new customers.

The economic and mathematical model presented in the scientific study is a development plan for the division and the enterprise as a whole. It models how, depending on revenue growth, the profitability and profit of the division's work increases, the cost and prices of the company's products decrease, the material remuneration of personnel and deductions for the development of the division and the enterprise as a whole increase. Social financial technologies for the development of enterprises allows employees to participate in management, coordinating with management their needs for the purchase of technological and auxiliary equipment, based on the amount of financial resources transferred by each division to the development fund. The proposed mechanism of material and moral incentives contributes to the fact that not only the management and owners of the enterprise and their deputies think about the development of the enterprise, but the entire workforce is interested in increasing professional growth, prestige and demand for their work, which contributes to a significant increase in the quality and availability of the enterprise's products for citizens of the state.

Conclusion, recommendations and direction of future research. The economic and mathematical model of the integrated system of social financing of enterprises and the country's economy developed by the authors can be used to improve the accuracy, efficiency and validity of management decisions in the interests of enterprise development, increase the profitability of its activities, increase employee salaries and contributions to the development fund.

The results of the development of the scientific and methodological apparatus and the implementation of practical tools of this study allow us to conclude that the goal of the study has been achieved. The completed scientific work provides management decision makers with effective tools for social financing of enterprises and the country's economy. ***The directions of further research on the problems of the article are:*** the introduction of a progressive system of stimulating the work of employees in other fields of activity, for example, the provision of educational services to motivate highly effective work of scientific and pedagogical workers, improving their qualifications and professional level; expansion of the technology of setting and solving the problem of nonlinear programming for prospective investors and the search for optimal sources of financial resources from the point of view of the weighted average price for the implementation of investments in the development of enterprises and the economy of the country; adaptation of the economic and mathematical model developed in the article at enterprises of all sectors of the economy; the inclusion of the developed economic and mathematical tools into a unified information and analytical system for managing the financial flows of the enterprise, its interaction with widely used application software products and others.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. Prskawetz, G. Feichtinger, M. Luptacik, A Milik, “Endogenous growth of population and income depending on resource and knowledge,” European Journal of Population, **14**(4), 305-331, 1999, doi: 10.1007/BF02863319.

[2] A. Bellou, B. Kaymak, “The cyclical of job quality and real wage grow,” American Economic Rewiew Insigths, **3**(1), 83-96, 2021, doi: 10.1257/aeri.20190553.

[3] Y. Avraamova, “The risks of reduced welfare of the population: stratification analysis,” Social Sciences, **48**(3), 37-50, 2017, doi: 10.21557/SSC.49373409.

[4] A.R. Cardoso, “Long-term impact of minimum wages on workers’ careers: evidence from two decades of longitudinal linked employer–employee data,” The Scandinavian Journal of Economics, **121**(4), 1337-1380, 2019, doi: 10.1111/sjoe.12327.

[5] C. Krolage, A. Peichl, “Long-run trends in top income shares: the role of income and population growth,” The Journal of Economic Inequality, **20**, 97-118, 2022, doi: 10.1007/s10888-021-09520-8.

[6] C.C. Chao, M.S. Ee, X. Nguyen, E.S.H. Yu, “Minimum wage, firm dynamics, and wage inequality: theory and evidence,” International Journal of Economic Theory, **18**(3), 2021, doi: 10.1111/ijet.12307.

[7] Y.Y. Finogenova, “Analysis of new financing opportunities for Russian pension insurance system,” European Journal of Economics and Management Sciences, **1**, 47-53, 2017, doi: 10.20534/EJEMS-17-1-47-52.

[8] N. Gavkalova, I. Kolupaieva, M. Barka Zine, “Analysis of the efficiency of levers in the context of implementation of the state regulatory policy,”

Економічний часопис-хxі, **165**(5-6), 41-46, 2017, doi: 10.21003/ea.V165-09.
[9] J. Barry, "Real wage growth in the U.S. health workforce and the narrowing of the gender pay gap," Human Resources for Health, **19**(1), 2021, doi:10.1186/s12960-021-00647-3.
[10] J. Adda, C. Dustmann, "Sources of wage growth," Journal of Political Economy, **131**(2), 2023, doi: 10.1086/721657.
[11] E.V. Kabashova, "Statistical methods in the analysis factors of welfare of the population," Modern European Researches, **1**, 27-34, 2017.
[12] J. Končar, R.M. Marić, S. Yučenović, G. Yukmirović, "Employee welfare in the western balkanretail sector: how to improve it through socio-organizational variables," Revija za Socijalnu Politiku, **27**(2), 151-170, 2020, doi: 10.15838/esc.2022.1.79.6.
[13] L. Malcoci, V. Mocanu, "The middle class in the republic of Moldova: determinant factors of formation," Economie si Sociologie: Revista Teoretico-stiintifica, **1-2**, 103-111, 2017.
[14] E. Pelinescu, M. Iordan, M.N. Chilian, "Digitization and population welfare in the new EU member states," Romanian Journal of Economic Forecasting, **24**(4), 59-75, 2021.
[15] J. Pi, Y. Ge, J. Yin, "The impacts of rural property rights on urban unemployment, wage inequality, and welfare in developing countries," Journal of Economic Analysis & Policy, **17**(1), 2015-2025, 2017, doi: 10.1515/bejeap-2015-0225.
[16] P. Litynski, "Living in sprawling areas: a cost-benefit analysis in Poland," Journal of Housing and the Built Environment, 2022, doi:10.1007/s10901-022-09986-6.
[17] S.N. Rastvortseva, I.V. Manaeva, "Modern development of small and medium-sized cities: trends and drivers," Economic and Social Changes: Facts, Trends, Forecast, **15**(1), 110-127, 2022, doi: 10.15838/esc.2022.1.79.6.
[18] G.N. Semenova, E.I. Larionova, O.G. Karpovich, S.V. Shkodinsky, F.M. Ouroumova, "Social integration as a factor of economic growth: experience and perspectives of developing countries," International Journal of Sociology and Social Policy, 2020, doi: 10.1108/IJSSP-03-2020-0083.
[19] R.S.P. Singh, "The 4 quadrants of monthly household budgets," Research Trends in Multidisciplinary Subjects, **1 264**, 264-271, 2022.
[20] E. Sofilda, R. Hamzah, M. Zilal Hamzah, "The effect analysis of minimum regional wages and macroeconomic on poverty level in Indonesia period 2010-2015," Oida International Journal of Sustainable Development, **11**(10), 11-24, 2018.
[21] S. Brunow, S. Losch, O. Okhrin, "Labor market tightness and individual wage growth: evidence from Germany," Journal for Labour Market Research, **56**(1), 2022.
[22] L. Synyayeva, A. Yarchuk, D. Verba, I. Verkhohod, D. Aleksandrov, "Resourses of educational and healthcare industries and population welfare: comparative analysis in post-socialist and OECD countries," WSEAS Transactions on Business and Economics, **18**, 531-542, 2021, doi: 10.37394/23207.2021.18.54.
[23] U. Ulugkhan, "Employment of the population on the basis of development and regulation of the labor market, demographic factors and welfare economic development of the region," Internauka, **182**(6-2), 66-67, 2021.
[24] X. Lu, X. Wu, R. Xu, "Why low unemployment rate in the United States has not delivered meaningful wage growth," Highlights in Business, Economics and Management, **2**, 2022, doi: 10.54097/hbem.v2i.2386.
[25] E.V. Sokolov, E.V. Kostyrin, A.B. Balantsev, "Social technologies of enterprise financing," Economics and Management: Problems, Solutions, **3**(4), 13-27, 2021, doi: 10.36871 / ek.up.pr2021.04.03.002.
[26] E.V. Sokolov, E.V. Kostyrin, "Justification of expediency of transition of financing of domestic healthcare to medical savings accounts," Economics and Management: Problems, Solutions, **4**(8), 194-212, 2018.
[27] E.V. Sokolov, E.V. Kostyrin, "Justification of the necessity and effectiveness of the introduction of medical savings accounts for all subjects of the Russian Federation and Russia as a whole," Economics and Management: Problems, Solutions, **11**(1), 52-65, 2018.
[28] A.P.B. Etges, V. Grenon, M. Lu et al., "Development of an enterprise risk inventory for healthcare," BMC Health Serv Res, **18**, 578, 2018, doi: 10.1186/s12913-018-3400-7.
[29] J. Horkoff, M.A. Jeusfeld, J. Ralyté et al., "Enterprise modeling for business agility," Bus Inf Syst Eng, **60**, 1-2, 2018, doi: 10.1007/s12599-017-0515-z.
[30] J.E. Vahlne, W.A. Bhatti, "Relationship development: a micro-foundation for the internationalization process of the multinational business enterprise," Manag Int Rev, **59**, 203–228, 2019, doi: 10.1007/s11575-018-0373-z.
[31] A.A. Adamenko, D.V. Petrov, V.V. Markelov, "Effective management of cash flows of an economic entity," Bulletin of the Academy of Knowledge, **6**(35), 14-18, 2019.
[32] K.B. Akhmedjanov, I.S. Musakhonzoda, "Financial performance management system as a factor of efficiency of a balanced financial management system," Journal of Advanced Research in Dynamical and Control Systems, **5**(12), 301-310, 2020, doi: 10.5373/JARDCS/V12I5/20201718
[33] M. Alnaimat, N. Rudyk, A. Al-Naimi, A. Panchenko, I. Turski, "The impact of international economic sanctions on the use of financial technologies," WSEAS Transactions on Business and Economics, **20**, 682-693, 2023, doi: 10.37394/23207.2023.20.63.
[34] L.N. Altunina, I.E. Khorolskaya, A.I. Smirnova, "Evaluation of the effectiveness of cash flow management of a commercial organization," Bulletin of the Academy of Knowledge, **4**(39), 41-46, 2020, doi: 10.24411/2304-6139-2020-10435.
[35] S. K. Azimov, "The development of financial markets and financial theory," Theoretical & Applied Science, **8**(112), 2409, 2022, doi: 10.15863/TAS.2022.08.112.3
[36] L.E. Basovsky, Financial management, Moscow, INFRA-M, 2019.
[37] M.U. Boboev, D.A. Gaibulloeva, "Baҳisobgirii moliyavӣ ҳamchun manbai ittilooti nizomi menecҷmenti moliyavӣ," Bulletin of the Technological University of Tajikistan, **1**(44), 125, 2021.
[38] T. Bolgar, V. Varenyk, Z. Pestovska, I. Miro, "Innovative information technologies in financial management," Academic Review, **2**, 98-110, 2022, doi: 10.32342/2074-5354-2022-2-57-8.
[39] J.L. Combes, P. Plane, T. Kinda, R. Ouedraogo, "Financial flows and economic growth in developing countries," Economic Modelling, **83**(c), 195-209, 2019, doi: 10.1016/j.econmod.2019.02.010.
[40] A. Aurora Ndelo, Y. Permatasari, I. Harymawan, N. Anridho, "Corporate tax avoidance and investment efficiency: evidence from the enforcement of tax amnesty in Indonesia," Economies, **10**(10):25, 1-23, 2022, doi:10.3390/economies10100251.
[41] A. Ali, "Investment projects portfolio analyses using fuzzy evaluation methods," Advances in Intelligent Systems and Computing, **1306**, 685-693, 2021, doi: 10.1007/978-3-030-64058-3_85.
[42] K. Amel, "Corporate sustainability disclosure and investment efficiency: the Saudi Arabian context," Sustainability, **14**(21):13984, 1-13, 2022, doi: 10.3390/su142113984.
[43] A. Costăngioară, "Introducing a new requirement in the assessment of the impact of companies on the environment, the multidimensional approach," in 19th International Scientific Geoconference, 271-278, 2019, doi: 10.5593/sgem2019/4.1/S17.035.
[44] E. Vasilyeva, T. Kudryavtseva, D.V. Ovsyanko, "Evaluation of the effectiveness of investment in innovation in industry," Innovation Project Management, **9**, 13-18, 2019, doi:10.17513/vaael.693.
[45] M.A. Filina, Z.M. Umarova, "Investment risk: types and evaluation methods," Era of Science, **20**, 477-480, 2019, doi: 10.24411/2409-3203-2019-12099.
[46] I. Golaydo, I. Parshutina, G. Gudimenko, A. Lazarenko, N. Shelepina, "Evaluation, forecasting and management of the investment potential of the territory," Journal of Applied Economic Sciences, **2**(48):12, 618-635, 2017.
[47] M.V. Grenaderova, "Methodological approaches to the evaluation of investment efficiency taking into account environmental factors," in IOP Conference Series Materials Science and Engineering, **753**(8):082036, 1-5, 2020, doi: 10.1088/1757-899X/753/8/082036.
[48] J.J. Heung, M.O. Hyun, "Debt origin and investment efficiency from Korea," International Journal of Financial Studies, **8**(3):47, 1-27, 2020, doi: 10.3390/ijfs8030047.
[49] D.D. Ierkovska, M.Yu. Bugayko, Hryhorak, Yu. Zaloznova, N. Trushkina, S.I. Gritsenko, T.I. Dovgan, V.Z. Ninich, V.A. Kulik, V.Ye. Marchuk, O.M. Harmash, O. Karpun, N.M. Perederii, "Intellectualization of logistics and supply chain management," Electronic Scientific and Practical Collection, 9, 2021, doi:10.46783/smart-scm/2021-9.
[50] J. Bian, Ya. Shan, G. Zhao, "Evaluation and analysis of environmental protection investment efficiency in China based on DEA model," Research Square, 1-16, 2021, doi: 10.21203/rs.3.rs-277471/v1.
[51] K. Yang, S. Fahad, F. Yuan, "Evaluating the influence of financial investment in compulsory education on the health of Chinese adolescents: a novel approach," BMC Public Health, **22**(1):1725, 1-15, 2022, doi: 10.1186/s12889-022-14125-5.
[52] K. Kim, J. Koo, J. Kim, "Development of an investment efficiency evaluation model for waterworks maintenance through data envelopment

analysis," Desalination and Water Treatment, 73-72, 2018, doi: 10.5004/dwt.2018.21640.
[53] M.V. Korolkova, T.S. Novikova, "Approaches to the efficiency evaluation for the complex of interrelated investment projects," World of Economics and Management, **18**(3):66-80, 2018, doi: 10.25205/2542-0429-2018-18-3-66-80.
[54] P. Kovalenko, A. Rokochinskiy, P. Volk, V. Turcheniuk, N. Frolenkova, R. Tykhenko, "Evaluation of ecological and economic efficiency of investment in water management and land reclamation projects," Journal of Water and Land Development, **1-3**(48), 81-87, 2021, doi: 10.24425/jwld.2021.136149.
[55] L. Lijia, X. Guanglong, L. Keyao, H. Juhua, M. Wanzhen, W. Xuejie, Zh. Huiru, "Investment efficiency assessment of distribution network for the high proportion of renewable energy: a hybrid multiattribute decision-making method," Mathematical Problems in Engineering, 2022, doi: 10.1155/2022/2373363.
[56] L. Wang, J. Liang, J. Liu, "Research on investment efficiency evaluation of wind power projects under supply-side reform," in IOP Conference Series Earth and Environmental Science, **508**(1):012089, 1-5, 2022, doi: 10.1088/1755-1315/508/1/012089.
[57] Y. Liu, H. Zhang, Y. Wu, Y. Dong, "Ranking range based approach to MADM under incomplete context and its application in venture investment evaluation," Technological and Economic Development of Economy, **5**(25), 877-899, 2019, doi: 10.3846/tede.2019.10296.
[58] M. Luchko, R. Ruska, G. Lew, I. Vovk, "Modelling the optimal size of investment portfolio in a non-state pension fund," Journal of International Studies, **1**(12), 239-252, 2019, doi: 10.14254/2071-8330.2019/12-1/16.
[59] S.K. Malhotra, A. Saran, D. John, H. White, N.D. Cruz, J. Eyers, E. Beveridge, N. Blöndal, "Studies of the effectiveness of transport sector interventions in low- and middle-income countries: an evidence and gap map," Campbell Systematic Reviews, **4**(17), e1203, 2021, doi: 10.1002/cl2.1203.
[60] M. Nourani, Q.L. Kweh, W.M. Lu, I. Gurrib, "Operational and investment efficiency of investment trust companies: do foreign firms outperform domestic firms?" Financial Innovation, **8**(79), 2022, doi: 10.1186/s40854-022-00382-1.
[61] V. Obinna, C. Amarachi, L. Anthony, D. Ajibare, Oladayo, D. Oluleye, "Small and medium enterprises assessment of investment decisions and financial performance of small and medium enterprises in federal capital territory," Nigeria, 40-49, 2022, doi: 10.46281/ijsmes.v5i1.1813.
[62] Y. Qixiong, L. Zhenqiu, C. Yu, Z. Ying, "An investment efficiency evaluation model for distribution network with distributed renewable energy resources," Frontiers in Energy Research, **10**, 1-4, 2022, doi: 10.3389/fenrg.2022.931486.
[63] Q. Trung Tran, Q. Dat Tran, "How does national culture affect corporate investment efficiency?" Global Business Review, 2022, doi: 10.1177/09721509221088898.
[64] O. Sumets, "Evaluation of the investments efficiency in the development of the key component of the supply chain," Electronic Scientific and Practical Publication in Economic Sciences, **5**, 43-61, 2021, doi: 10.46783/smart-scm/2021-5-4.
[65] S.T. Do, N.N.N. Tran, "An investment willingness assessment model for private sector in PPP transportation," Public-Private Partnership Transportation Projects in Vietnam, 2019.
[66] P. Wang, "Application of cloud computing and information fusion technology in green investment evaluation system," Journal of Sensors, 1-13, 2021, doi: 10.1155/2021/2292267.
[67] X. Wang, W. Chen, W.J. Lekse, "Investment selection and evaluation for china express delivery market," International Journal of Industrial Engineering: Theory Applications and Practice, **2**(28), 190-208, 2021.
[68] W. Hao, H. Gao, Z. Liu, "An evaluation study on investment efficiency: a predictive machine learning approach," Complexity, 1-9, 2021, doi: 10.1155/2021/6658516.
[69] X. Tian, Y. Zhang, G. Qu, "The impact of digital economy on the efficiency of green financial investment in China's provinces," International Journal of Environmental Research and Public Health, **19**(14):8884, 1-18, 2022, doi:10.3390/ijerph19148884.
[70] G. Xiong, L. Wang, "Factors and economic evaluation of transnational investment risks," Discrete Dynamics in Nature and Society, 1030183, 2021, doi: 10.1155/2021/1030183.
[71] G. Kou, X. Chao, Y. Peng, F. Wang, "Network resilience in the financial sectors: advances, key elements, applications, and challenges for financial stability regulation," Technological and Economic Development of Economy, **28**(2), 531-558, 2022, doi:10.3846/tede.2022.16500.
[72] L. Lehoux, T.V. Morozova, E.G. Safonova, A.D. Balashova, M.V. Protasov, "Practical aspects in calculating of impairment of financial assets according to IFRS 9 Financial instruments," in Proceedings of the 33RD International Business Information Management Association Conference, IBIMA 2019: Education Excellence and Innovation Management Through Vision 2020, 6624, 2019.
[73] P. Liu, A. Hendalianpour, "A branch cut/metaheuristic optimization of financial supply chain based on input-output network flows: investigating the Iranian orthopedic footwear," Journal of Intelligent and Fuzzy Systems, **41**(2), 2561, 2021, doi: 10.3233/JIFS-201068.
[74] U.S. Aliyu, H.L. Ozdeser, B.L. Çavuşoğlu, M.A.M. Usman, "Food security sustainability: a synthesis of the current concepts and empirical approaches for meeting SDGS," Sustainability, **21**(13), 2021, doi: 10.3390/su132111728.
[75] H. El Bilali, "Research on agro-food sustainability transitions: where are food security and nutrition?" Food Security, **11**(3), 559-577, 2019, doi: 10.1007/s12571-019-00922-1.
[76] D. Enahoro, K.M. Rich, S.S. Staal, D. Mason-D'croz, M. Mul, T.P. Robinson, P. Thornton, "Supporting sustainable expansion of livestock production in south Asia and sub-saharan Africa: scenario analysis of investment options," Global Food Security, **20**, 114-121, 2019, doi: 10.1016/j.gfs.2019.01.001.
[77] D. Hall, "National food security through corporate globalization: Japanese strategies in the global grain trade since the 2007–8 food crisis," Journal of Peasant Studies, **47**(5), 993-1029, 2020, doi: 10.1080/03066150.2019.1615459.
[78] S. Li, X. Li, L. Sun, G. Cao, G. Fischer, S. Tramberend, "An estimation of the extent of cropland abandonment in mountainous regions of China," Land Degradation and Development, **29**(5), 1327-1342, 2018. doi: 10.1002/ldr.2924.
[79] A. Semin, A. Kibirov, U. Rassukhanov, "Problems and main mechanisms to increase investment attractiveness of agricultural production," European Research Studies Journal, **21**(2), 378-400, 2018, doi: 10.35808/ersj/1009.
[80] G.T. Shakulikova, A.S. Baidalinova, A.M. Uakhitzhanova, G.B. Baimuldina, E.B. Ikmatova, "Agriculture financing – a basic premise for ensuring food security in Kazakhstan," Journal of Applied Economic Sciences, **13**(1), 216-226, 2018.
[81] M. Svanidze, L. Götz, I. Djuric, T. Glauben, "Food security and the functioning of wheat markets in Eurasia: a comparative price transmission analysis for the countries of central Asia and the south Caucasus," Food Security, **11**(3), 733-752, 2019, doi: 10.1007/s12571-019-00933-y.
[82] O.V. Vaganova, N.E. Solovjeva, O.N. Polukhin, V.M. Zakharov, G.G. Zabnina, R.V. Lesovik, S.L. Lesovaya, M.E. Ageykina, "Analysis of supply chain in investment activity in the Russian agricultural complex," International Journal of Supply Chain Management, **9**(5), 1615-1622, 2020.
[83] V. Verma, B. Vishal, A. Kohli, P.P. Kumar, "Systems-based rice improvement approaches for sustainable food and nutritional security," Plant Cell Reports, **40**(11): 2021-2036, 2021, doi: 10.1007/s00299-021-02790-6.
[84] K. Wakjira, T. Negera, A. Zacepins, A. Kviesis, V. Komasilovs, S. Fiedler, S. Kirchner, O. Hensel, D. Purnomo, M. Nawawi, A. Paramita, O.F. Rachman, A. Pratama, N.A. Faizah, M. Lemma, S. Schaedlich, A. Zur, M. Sper, K. Proschek, K. Gratzer, R. Brodschneider, "Smart apiculture management services for developing countries-the case of SAMS project in Ethiopia and Indonesia," Peerj. Computer Science, **7**, e484, 2021, doi: 10.7717/PEERJ-CS.484.
[85] E.V. Sokolov, E.V. Kostyrin, "Medical savings accounts as a tool for increasing doctors' salaries and motivating Russian citizens to high-performance work and a healthy lifestyle," Economics and Management: Problems, Solutions, **7**(2), 24-31, 2020, doi: 10.34684/ek.up.p.r.2020.07.02.004.
[86] E.V. Sokolov, E.V. Kostyrin, "Organization of the transition of citizens of the Sverdlovsk region to medical savings accounts," Economics and Management: Problems, Solutions, **12**(108):1, 39-60, 2020, doi: 10.36871/ek.up.p.r.2020.12.01.007.
[87] E.V. Sokolov, E.V. Kostyrin, "The economic effect of using medical savings accounts instead of the existing system of healthcare financing," Economics and Management: Problems, Solutions, **2**(110):1, 16-26, 2021, doi: 10.36871/ek.up.p.r.2021.02.01.003.
[88] E.V. Sokolov, E.V. Kostyrin, "The mechanism of financing health care on the basis of medical savings accounts," Economics and Management: Problems, Solutions, **3**(5), 64-85, 2019.
[89] E.V. Sokolov, E.V. Kostyrin, S.V. Lasunova, "Financial technologies for the development of enterprises and the economy of Russia," Economics

and Management: Problems, Solutions, **10**(118):1, 91-106, 2021, doi: 10.36871/ek.up.p.r.2021.10.01.013.
[90] E.V. Sokolov, E.V. Kostyrin, P.A. Nevezhin, "Modeling of the insurance and accumulative parts of the old-age labor pension," Economics and Management: Problems, Solutions, **9**(1), 132-153, 2018.
[91] E.V. Sokolov, E.V. Kostyrin, K.V. Rudnev, "Social financial technologies for the development of enterprises and the economy of Russia," Soft Measurements and Calculations, **9**(46), 74-96, 2021, doi: 10.36871/2618-9976.2021.09.004.
[92] E.V. Sokolov, E.V. Kostyrin, "Social financial technologies for the development of large scale healthcare systems and the Russian economy," in 2022 15th International Conference Management of Large-Scale System Development (MLSD), 1-5, 2022. doi: 10.1109/MLSD55143.2022.9934748.
[93] E.V. Sokolov, P.A. Nevezhin, "Breakthrough technologies of old-age labor pension financing," Economics and Management: Problems, Solutions, **7**(3), 4-9, 2018.
[94] E.V. Sokolov, "The main source of development of the financial system of Russia," Economics and Management: Problems, Solutions, **9**(2), 158-161, 2016.
[95] E.V. Sokolov, E.V. Kostyrin, "Breakthrough technologies of old-age labor pension financing," Economics and Management: Problems, Solutions, **7**(115):1, 63-80, 2021, doi: 10.36871/ek.up.p.r.2021.07.01.009.
[96] E.V. Kostyrin, "Progressive system of stimulating the work of doctors," Economics and Entrepreneurship, **2**(103), 1122-1131, 2019.
[97] S. Abbasov, "Improving cash flow management," Economic Herald of the Donbas, 33-38, 2021, doi: 10.12958/1817-3772-2021-4(66)-33-38.
[98] M. Apsite, D. Belova, "Financial analysis as a cash flow management tool," Interactive Science, 2019, doi: 10.21661/r-509061.
[99] N. Atakul, "Exploring the cash flow management strategies of Turkish construction companies," Journal of Construction Engineering, Management & Innovation, **5**, 2022, doi:10.31462/jcemi.2022.03168180.
[100] O. Chubka, I. Skoropad, "Features of cash flow management in public finance," Odessa National University Herald. Economy, **25**, 2020, doi: 10.32782/2304-0920/2-81-25.
[101] O. Chubka, R. Zhelizniak, "Cash flow management in banking," International Humanitarian University Herald. Economics and Management, 2019, doi: 10.32841/2413-2675/2019-40-21.
[102] L. Dvořáková, J. Kronych, A. Malá, "Cash flow management as a tool for corporate processes optimization," Smart Science, **6**(1-7), 2018, doi: 10.1080/23080477.2018.1505370.
[103] E. Etim, E. Daferighe, E. Enang, M. Nyong, "Cash flow management and financial performance of selected listed companies in Nigeria," Indo-Asian Journal of Finance and Accounting, **3**, 27-46,2022, doi: 10.47509/IAJFA.v03i01.03.
[104] Z. Imanbayeva, H. Kusainov, B. Zhakupova, A. Niyazbayeva, B. Bimbetova, "Ways to improve the company's cash flow management," Reports, **4**, 177-185, 2020, doi: 10.32014/2020.2518-1483.107.
[105] K. Ketova, I. Rusyak, E. Kasatkina, E. Saburova, D. Vavilova, "Organizing the cash flow management in the construction industry in the Russian Federation," in IOP Conference Series: Materials Science and Engineering, **862**, 042035, 2020, doi: 10.1088/1757-899X/862/4/042035.
[106] K. Koopman, R. Cumberlege, "Cash flow management by contractors," in IOP Conference Series: Earth and Environmental Science, **654**, 012028, 2021, doi: 10.1088/1755-1315/654/1/012028.
[107] O. Korobova, M. Blum, "Application of digital technologies in cash flow management at a commercial enterprise," Voprosy Sovremennoj Nauki i Praktiki, 062-070, 2021, doi: 10.17277/voprosy.2021.02.pp.062-070.
[108] T. Kucherenko, H. Anishchenko, "Accounting and analytical support of cash flow management of enterprises," Efektyvna Ekonomika, 2020, doi: 10.32702/2307-2105-2022.2.12.
[109] O. Kudyrko, I. Kopchykova, "Methodological approaches to cash flow management at the enterprise," European Journal of Economics and Management, **8**, 17-22, 2022, doi: 10.46340/eujem.2022.8.3.3.
[110] A. Nanggala, "Free cash flows, management ownership, dividend policy, and debt policy," Jurnal Ekonomi Akuntansi dan Manajemen, **19**, 30, 2020, doi: 10.19184/jeam.v19i1.17544.
[111] E. Nangih, T. Ofor, O. Joshua, "Cash flow management and financial performance of quoted oil and gas firms in Nigeria," Journal of Accounting and Financial Management, **6**, 2020.
[112] M. Nashkerska, N. Patriki, "Instruments for cash flow management of construction enterprises," Economic Analysis, 223-229, 2020, doi: 10.35774/econa2020.01.02.223.
[113] O. Oladimeji, O. Aina, "Cash flow management techniques practices of local firms in Nigeria," International Journal of Construction Management, **21**, 1-9, 2018, doi: 10.1080/15623599.2018.1541705.
[114] J. Ongpeng, K. Aviso, D. Foo, R. Tan, "Graphical pinch analysis approach to cash flow management in engineering project," Chemical Engineering Transactions, **76**, 493, 2019, doi: 10.3303/CET1976083.
[115] K.V. Oriekhova, O.Hr. Golovko, "Cash flow management strategy," Economics and Law, 89-97, 2022, doi: 10.15407/econlaw.2022.01.089.
[116] X. Peng, Z. Ren, "Design and implementation of electronic commerce cash flow management system," Agro Food Industry Hi-Tech, **28**, 2535-2540, 2017.
[117] T. Phuong, N. Thuy, "Impact of cash flow management on shareholder value of listed real estate companies in Vietnam," Vnu Journal of Economics and Business, **1**, 2021, doi: 10.57110/jeb.v1i4.4584.
[118] N. Piontkevich, E. Shatkovskaya, "An organization's cash flow management in digital economy," in Proceedings of the Ecological-Socio-Economic Systems: Models of Competition and Cooperation (ESES 2019), 2020, doi: 10.2991/assehr.k.200113.099.
[119] S.T. Dinh, N.C. Phuc, "Foreign financial flows, human capital and economic growth in African developing countries," International Journal of Finance and Economics, **27**(5), 2020, doi: 10.1002/ijfe.2310.
[120] L.N. Dobryshina, "Socio-economic security: essence, evolution, factors," Transport Business of Russia, **10**, 5-7, 2011.
[121] K.V. Ekimova, I.P. Savelyeva, K.V. Kardapoltsev, Financial management, Moscow, Yurayt Publishing House, 2019.
[122] Explanatory note to the draft federal law "On compulsory medical insurance in the Russian Federation using medical accounts," 2019, URL: http: // sokolov.expert.
[123] V.A. Fedorov, "Methodology for assessing the impact of the effect of financial leverage on the total cash flow of the company," Innovations and Investments, **10**, 60-62, 2021.
[124] A.M. Galimova, A.N. Kirpikov, "Perspective economic assessment of financial stability indicators organizations with the use of economic and mathematical modeling of cash flows," Economic Bulletin of the Republic of Tatarstan, **1**, 64-72, 2019.
[125] A.A. Ilyinykh, "The economic essence of the cash flow of the organization," Young Scientist, **14**, 103-105, 2019.
[126] Z. JingJing, "Risk assessment method of agricultural management investment based on genetic neural network," Security and Communication Networks, **1-10**, 2022, doi: 10.1155/2022/2373363.
[127] A. Podgornaya, K. Romanov, "Actual issues of cash flow management in enterprises in Russia," Gênero & Direito, **8**, 2019, doi: 10.22478/ufpb.2179-7137.2019v8n6.49183.
[128] E.V. Kostyrin, "Economic and mathematical models of financial incentives for the personnel at medical organization departments," International Journal of Pharmaceutical Research, **4**(12), 1769-1780, 2020, doi: 10.31838/ijpr/2020.12.04.253.
[129] E.V. Kostyrin, "Economic and mathematical modeling of financial resource management in medical organizations," Industrial Engineering and Management Systems, **3**(19), 716-729, 2020, doi: 10.7232/iems.2020.19.3.716.
[130] A.V. Kemenov, Cash flow management, Moscow, UNITY-DANA, 2015.
[131] K.V. Ketova, D.D. Vavilova, "Optimization of financial flows in a building company using an escrow account in the Russian Federation," in Recent Research in Control Engineering and Decision Making, 427-442, 2021, doi: 10.1007/978-3-030-65283-8_35.
[132] A.N. Kirpikov, T.A. Sibgatullin, "Simulation modeling of cash flows in the system of predictive analysis and financial management of an organization," Scientific Review: Theory and Practice, **7**(63):9, 1086-1100, 2019, doi: 10.35679/2226-0226-2019-9-7-1086-1100.
[133] A.E. Kisova, V.K. Zolotareva, "Cash flows as a factor in ensuring financial stability of an organization," Notes of a Scientist, **8**, 362, 2021.
[134] M.A. Magomedov, E.D. Ozdeadzhieva, "Features of managing financial flows of the enterprise," Economics and Entrepreneurship, **3**(140), 1038-1041, 2022, doi: 10.34925/EIP.2022.140.03.196.
[135] S. Mahdi, Z. Grzegorz, A. Arash, E.G. Fratemeh, "The impact of investment efficiency on firm value and moderating role of institutional ownership and board independence," Journal of Risk and Financial Management, **15**(4):170, 1-13, 2022, doi: 10.3390/jrfm15040170.
[136] M.B. Melikhov, "Methodological foundations of systematic economic and statistical modeling of financial flows," Bulletin of the Tula branch of the Financial University, **1-1**, 227-238, 2019.
[137] S.S. Morozkina, A.V. Rykalo "Analysis of the organization's cash flows," Natural Sciences and Humanities Research, **24**(2), 55-59, 2019.
[138] E.A. Pirogova, V.S. Kirsanova, "Financial flow management," Development Trends and Actual Problems of Assessment, Management and Regulatory Support of the Financial System of Russia, **2**, 205, 2020.

[139] N.S. Plaskova, N.A. Prodanova, E.V. Prokofieva, "Methods of financial analysis of the organization's cash flows and assessment of the effectiveness of cash flow management," Financial Analysis: Theory and Practice, 106, 2021.
[140] A. Ramli, L. Yekini, "Cash flow management among micro-traders: responses to the COVID-19 pandemic," Sustainability, **14**, 10931, 2022, doi: 10.3390/su141710931.
[141] I. Rosemary, I. Abner, A. Jack, O. Fausat, E. Enoch, U. Samuel, "Cash flow management and industrial firms' performance in Nigeria," Universal Journal of Accounting and Finance, **9**, 2021, doi: 10.13189/ujaf.2021.090416.
[142] A. Shash, A. Qarra, "Cash flow management of construction projects in Saudi Arabia," Project Management Journal, **49**(2):875697281878797, 2018, doi: 10.1177/8756972818787976.
[143] M. Sofyan, U. Ludigdo, A.D. Mulawarman, (2021). "The meaning of cash flow management for the non-bank housing developers," International Journal of Research in Business and Social Science, (2147-4478), **10**, 195-203, doi: 10.20525/ijrbs.v10i4.1174.
[144] A. Sulla, D. Slepchenko, I. Kuzmicheva, E. Zaostrovskikh, "Financial logistics and its application in cash flow management", 2021, doi: 10.2991/assehr.k.210322.206.
[145] O. Vodolazska, K. Petrenko, "Enterprise cash flows: management principles and methods," Eastern Europe: Economy, Business and Management, 2019, doi: 10.32782/easterneurope.23-88.
[146] V. Voloshina-Sidey, I. Rud, O. Portnenko, "Cash flow management at the enterprise during COVID-19," Market Infrastructure, 2021, doi: 10.32843/infrastruct54-33.
[147] K.S. Zaryvakhina, "Cash flow management of the corporation in the conditions of instability," Scientific Review. Economic Sciences, **10-15**, 2022, doi: 10.17513/sres.1102.
[148] U.D. Atmond, V. Vyatkin, Z. Salcic, K.I.K. Wang, "A service-oriented programming approach for dynamic distributed manufacturing systems," IEEE Transactions on Industrial Informatics, **1**(16), 151-160, 2020, doi: 10.1109/TII.2019.2919153.
[149] M. Dehgnani, Z. Montazeri, A. Ehsanifar, A.R. Seifi, M.J. Ebadi, O.M. Grechko, "Planning of energy carriers based on final energy consumption using dynamic programming and particle swarm optimization," Електротехніка і Електромеханіка, **5**, 62-71, 2018, doi: 10.20998/2074-272X.2018.5.10.
[150] B. Di, A. Lamperski, "Newton's method, Bellman recursion and differential dynamic programming for unconstrained nonlinear dynamic," Dynamic Games and Applications, **413**, 87-102, 2021, doi: 10.1007/s13235-021-00399-8.
[151] B. Doerr, A. Eremeev, F. Neumann, M. Theile, C. Thyssen, "Evolutionary algorithms and dynamic programming," Theoretical Computer Science, **43**(412), 6020-6035, 2011, doi: 10.1016/j.tcs.2011.07.024.
[152] M.I. Gomoyunov, "Dynamic programming principle and Hamilton-Jacobi-Bellman equations for fractional-order systems," Siam Journal on Control and Optimization, **6**(58), 3185-3211, 2020, doi: 10.1137/19M1279368.
[153] M. Justiz, B. Bychko, S. Soler, V. Frolov, O. Malafeyev, A. Vasileva, "Application of dynamic programming to minimize energy consumption in industrial dryers," Bulletin de L'académie International Concorde, **2**, 3-20, 2021.
[154] T.S. Karaseva, "Genetic programming algorithm for the dynamic systems identification," Young People. Society. Modern Science, Technology and Innovation, **19**, 299-301, 2020.
[155] K. Land, B. Vogel-Heuser, S. Cha, "Applying dynamic programming to test case scheduling for automated production systems," Communication in Computer and Information Science, **1262**, 3-20, 2020, doi: 10.1007/978-3-030-58167-1_1.
[156] V. Struchenkov, D.A. Karpov, "High-speed dynamic programming algorithms in applied problems of a special kind," Mathematics and Statistics, **3**(8), 339-346, 2020, doi: 10.13189/ms.2020.080313.
[157] A. Yamaganov, A. Agafonov, V. Myasnikov, "An improved map matching algorithm based on dynamic programming approach," Lecture Notes in Business Information Processing, **413**, 87-102, 2021, doi: 10.1007/978-3-030-71846-6_5.
[158] Y. Zhu, G. Jia, "Dynamic programming and Hamilton-Jacobi-Bellman equations on time scales," Complexity, 7683082, 2020, doi: 10.1155/2020/7683082.
[159] J. Butt, "A conceptual framework to support digital transformation in manufacturing using an integrated business process management approach," Designs, **4**(3), 1-39, 2020, doi: 10.3390/designs4030017.
[160] K.A. Krylyvetz, A.A. Krylyvetz, "Process approach in the quality management system," Young People. Society. Modern Science, Technology and Innovation, **19**, 201-203, 2020.
[161] F. Li, G. Fang, "Process-aware accounting information system based on business process management," Wireless Communications and Mobile Computing, **2022**, 7266164, 2022, doi: 10.1155/2022/7266164.
[162] O. Olshanskiy, "Development of methods of improvement of business process management," Technology Audit and Production Reserves, **5**(4), 20-25, 2018, doi: 10.15587/2312-8372.2018.146862.
[163] P. Saragiotis, "Business process management in the port sector: a literature review," Maritime Bussiness Review, **4**(1), 49-70, 2019, doi: 10.1108/MABR-10-2018-0042.
[164] S.G. Sboeva, Y.A. Klyueva, N.L. Burdaev, M.A. Zaharchenko(2019). "Development of methodical bases for business process management optimization in clinical trials," Journal of Advanced Pharmacy Education and Research, **9**(2), 137-142, 2019.
[165] S. Tabassam, O. Hassan, E. AL-Qahtnae, N. AL-Ahmary, "Question metrics and its application to process management and improvement," National Journal on Engineering Applications, **7**(2), 52-58, 2019, doi: 10.15866/irea.v7i2.17013.
[166] L.A. Taskymbayeva, A.A. Shaikh, R.A. Salimbayeva, "Application of business process management methods in higher education institutions," Central Asian Economic Review, **3**(144), 45-55, 2022, doi: 10.52821/2789-4401-2022-3-45-55.
[167] A.S. Voskovskaya, T.A. Karpova, P.P. Rostovtseva, N.V. Guseva, A.V. Shelygov, "Development of the learning process management in the context of digitization," Revista Inclusiones, **7**(S4-5), 240-249, 2020.
[168] N. Yehorchenkova, O. Yehorchenkov, "Modeling of project portfolio management process by cart algorithm," Advances in Intelligent Systems and Computing, **1265**, 353-363, 2021, doi: 10.1007/978-3-030-58124-4_34.
[169] M. Saha, K.D. Dutta, "Nexus of financial inclusion, competition, concentration and financial stability: Cross-country empirical evidence," Competitiveness Review, 2020, doi: 10.1108/CR-12-2019-0136.
[170] M.Y. Shakatreh, M.M.A. Orabi, B.R.T. Shammout, "The role of financial vigilance in predicting possible financial distress among foreign banks," Journal of Management Information and Decision Science, **24**(5), 1-18, 2021.
[171] E.G. Spodareva, Ya.V. Sazhnikova, "Monitoring as a way of managing financial flows at the enterprise," Bulletin of the Ural Institute of Economics, Management and Law, **2**(59), 4, 2022.
[172] S. Suhadak, R.S. Mangesti, S.R. Handayani, "GCG, financial architecture on stock return, financial performance and corporate value," International Journal of Productivity and Performance Management, **69**(9), 1813, 2019, doi: 10.1108/IJPPM-09-2017-0224.
[173] L.B. Sungatullina, E.S. Golovchenko, "The economic essence of the company's cash flows as an object of financial management," Accounting in Budgetary and Non-profit Organizations, **1**(505), 14, 2021.
[174] X. Yujing, Z. Qinli, W. Daolin, W. Shihai, "Mining investment risk assessment for nations along the belt and road initiative," Land, **11**(8):1287, 2022, doi: 10.3390/land11081287.
[175] Q. Zhang, F. Li, "Financial resilience and financial reliability for systemic risk assessment of electricity markets with high-penetration renewables," IEEE Transactions on Power Systems, **37**(3), 2312-2321, 2022, doi: 10.1109/TPWRS.2021.3115499.
[176] N.F. Zhokabine, "Cash flows in the financial resources management system of the enterprise," Bulletin of Lugansk State University named after Vladimir Dal, **1**(43), 62-66, 2021.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 2, 123-129 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Assessment of Scattered-Bend Loss in Polymer Optical Fiber (POF) Displacement Sensor

Latifah Sarah Supian[*,1], Danial Haikal Mohd Razali[1], Chew Sue Ping[1], Nurul Sheeda Suhaimi[1], Sharifah Aishah Syed Ali[2], Nani Fadzlina Naim[3], Harry Ramza[4]

[1]*Department of Electrical and Electronics Engineering, Faculty of Engineering, National Defence University of Malaysia, Kuala Lumpur, 57000, Malaysia*

[2]*Faculty of Defense Science and Technology, National Defence University of Malaysia, Kuala Lumpur, 57000, Malaysia*

[3]*School of Engineering, College of Engineering, UiTM, Shah Alam, 40450, Malaysia*

[4]*Department of Electrical Engineering, Faculty of Industrial Technology and Informatics, Universitas Muhammadiyah Prof. Dr. HAMKA, Jakarta, 13420, Indonesia*

ARTICLE INFO

Article history:
Received: 09 January, 2023
Accepted: 30 March, 2023
Online: 28 April, 2023

Keywords:
Bending loss
Coupling effect
Displacement
Optical sensor
Polymer optical fiber
Scattering loss

ABSTRACT

This work investigated the coupling behavior of the scattered-bend loss in displacement sensor during the bending of the fiber by using a multimode polymer optical fiber (POF). To utilize the scattered-bend effect for displacement measurement, a side coupling technique can be used by twisting a pair of POF fibers and bent the structure into a loop. The working principle of the sensor is quite simple. The bent radius grows smaller as the fiber draughts which simulate a change of displacement. The scattered-bend loss increases as the illuminating fiber is bent in decreasing angle and the light being coupled to the receiving fiber. The fabricated sensor is tested based on static measurement analysis and the sensor is characterized by its sensitivity, resolution, linearity, and repeatability error. From the experiment, the fabricated sensor has a range of roughly 160 mm with a sensitivity of 0.817 nW/mm, a resolution of 1.228 mm, and a repeatability error of 1.856 %. The sensor exhibits high linearity from 0 mm to 80 mm. The sensor's design structure and analysis are simple, comprehensive, and cost-effective, with potential benefits in industrial applications.

1. Introduction

An optical fiber is a data transmission medium that uses lightwave propagation in conjunction with a fiber that is often constructed of glass or plastic. Optical fiber is mostly used in the application of high-speed and long communication. Various features of light behavior in optical fiber have been researched through time, including bend-loss studies which this work is based on and this paper is an extension of work originally presented in 2022 IEEE 9th International Conference of Photonics[1] among others [2], light propagation [3], coupled-mode theory [4] and scattering [5]. Until now, researchers have been attempting to explain many occurrences and properties in various types of optical fiber.

In sensing applications, most of the sensors can sense a variety of parameters such as temperature [6], pressure [7], displacement [8], biomedical [9], food quality [10,11] and chemical [12]. Many fibers have been used in the application of sensors such as glass optical fiber [13], polymer optical fiber, Fiber Bragg Grating [14], etc. In comparison to other fibers, POF is inexpensive, flexible [15] and is well known for its high reliability in short-distance communication and sensing applications. POF is widely recognized for its physical toughness, which can withstand the huge physical strain, as well as its low weight when compared to silica-based fiber, which is much more delicate and fragile due to the incredibly thin glass fiber it contains [16]. POFs are also immune to electromagnetic interference and have multiplexing capabilities [17].

[*]Corresponding Author: L.S. Supian, Faculty of Engineering, NDUM, +60129266933, cawa711@gmail.com/ sarah@upnm.edu.my

1.1. Common Techniques

Researchers had offered many techniques to obtain a high-performance sensor with a simple structure and low-cost manufacture in Polymer Optical Fiber (POF) sensor applications [18]. Among the techniques proposed by researchers are long-period gratings [19], nonlinear effects [20], surface plasmon resonance [21] and fiber bragg gratings [22]. An intensity-based technique has addressed a high-performance, simple, and a low-cost sensor for various detections based on the approaches indicated above since it does not require specific equipment [23]. A sensor that uses light intensity as the measurement detecting technique is known as an intensity-based sensor [24]. The common sensor that uses an intensity-based sensor is a pressure sensor, temperature sensor, turbidity sensor, and displacement sensor. In terms of displacement measurement sensors, various methods have been proposed. Most of the methods are able to detect static, dynamic, and plane-in-out measurement analysis.

Diffraction grating technique [25] has achieved a 4 mm to 14 mm range of the best linearity for displacement range, however, it has a complicated design setup and complicated analysis, and the range is very small. The technique is found to be complicated because it requires a specific angle cut or called a diffraction angle at the end of the illuminating fiber which can cause a loss if the cut not clean enough. The analysis of this method is also difficult to determine the diffraction order its need to obtain the LED light wavelength, light cone angle, the period of the diffraction grating, and diffraction angle of the illuminating fiber.

This works integrated the principle of macro-bend loss and scattering loss to realize the fabricated sensor. The works studied and investigated the right tapered depth, bending angle and turns in order to optimize the results.

1.2. Macro-bend and Scattering Loss

The displacement sensor described in this research is based on the coupling of scattered-bend loss where to determine the scattering loss in POF, two fibers were twisted together. The first fiber is acting as illuminating fiber, while the second fiber couples scattered-bend radiation loss using the side coupling approach [26]. Fiber loss increases as a result of bending, as does coupling power. Based on this technique, the power coupling structure is employed for the displacement measurement sensor. The power coupling was visualized by measuring the outputs and calculate the coupling ratio at each bend diameter.

In optical application, there are a few detected types of losses in light transmission which are bend loss, dispersion loss, scattering and attenuation loss [27]. A bending loss is a loss that occurred due to the physical pressure where bending is applied to the optical fiber strand. There are two types of bending losses which are macroscopic bending and microscopic bending [28]. When the light source propagates, the power gets transferred into other modes, so the changing of the mode due to different refraction index due to bending makes the power leaked where the power will not continue to propagate in the fiber core, which the radiated light is known as scattering loss. This loss is caused by the material compositional fluctuation, density of the material, and manufacturing defects of the fiber [29]. Due to the bending of the fiber and density fluctuation in the core of the fiber, the loss is called scattered-bend loss.

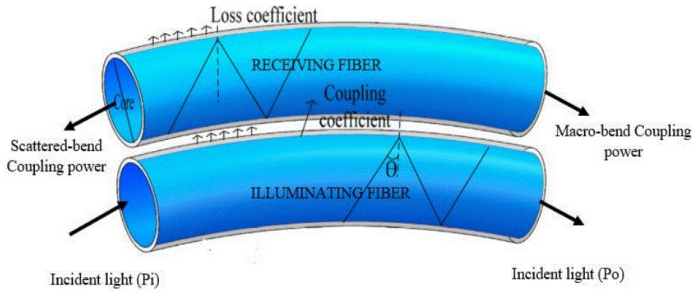


Figure 1: Coupled macro-bend and scattered-bend illustration.

In this method, the coupled scattered-bend loss has generated a polynomial to exponential-like curve while the macro-bend is producing an exponential curve [30]. This is caused by the power of the macro-bend being reflected at back-end of the receiving fiber. In general, due to the bending of fiber, consequently, there are two losses increase which are the macro-bend loss and scattered-bend loss. The macro-bend loss is propagating in the same direction as the light while a scattered-bend loss propagates in opposite direction toward the back-end of the receiving fiber. This can be illustrated in Figure 1.

To utilize the scattered-bend loss in the sensor application in this research, a polymer optical fiber, POF from ESKA Mitsubishi SK-40 bare multimode is used due to its durability and flexibility to the tightest bend and has a larger core which is 0.98 mm. To observe the relationship of the scattered-bend and macro-bend based on the changing of the bend radius, the coupling power received at both ends of the receiving fiber is measured and the ratio with respect to the LED input power can be calculated as:

$$C_s = \frac{P_2}{P_i}\% \quad (1)$$

$$C_m = \frac{P_1}{P_i}\% \quad (2)$$

Where C_s is the coupling ratio of scattered-bend, P_2 is the received power at the back-end of the fiber, C_m is the coupling ratio of macro-bend, P_1 is the received power at forward-end of the fiber and P_i is the input power of the light source.

2. Experimental Design and Setup

Based on the coupling ratio of both macro-bend and scattered-bend equations, it can be observed that with the increases of bend-diameter, the power received at both back-end and forward-end is increasing. In this research, variation of the received power and bending diameter testing parameters are applied in displacement detection. The design of the sensor structure is shown in Figure 2 setup.

In the proposed structure, only a single light source as input is required for the illuminating fiber. The LED used is from Advanced Fiber Solutions, OS417N with an operating wavelength of 650 nm and output power measured, 6.475 mW. The power meter used is also from Advanced Fiber Solution, OM110N for

detection of the received fiber at the back-end and forward-end of the receiving fiber. The power meter is set to 650 nm for calibration and the resolution of 1 mW or 0.01 dB.

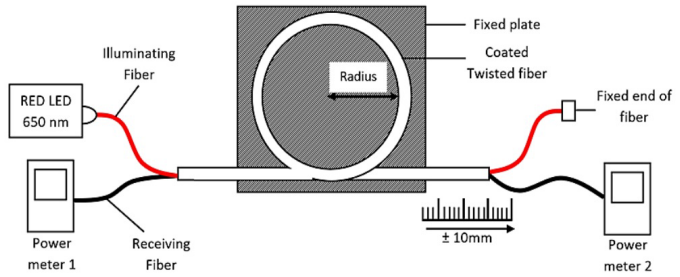


Figure 2: POF displacement sensor design structure.

For the main sensing part, a pair of twisted polymer optical fiber of SK-40 Bare Multimode POF is used and the fiber is coated with a black tube where in this experiment a black electrical shrink tube is used. The first fiber is connected to an LED source as an illuminating fiber. The second fiber is connected with a power meter to measure the received power both back-end and forward-end as receiving fiber. To test the sensor, the twisted fiber is bent to an initial 100 mm of bend radius, and adhesive tape is used to put the fiber at a fixed acrylic plate. The sensor is analyzed by using static measurement analysis with multiple variations of initial bend diameter.

During the experiment, the twisted fiber is manually pulled up to 100 mm for initial set loop with the changing of decreasing 10 mm displacement at each time. Each of the measurements is taken for both ends for scattered-bend and forward-end for macro-bend. From the result obtained, the losses of the light can be observed by the study of the graph of the received power. The tests are repeated three times to measure the repeatability of the reading for both the back-end and forward-end. The step will be repeated with setup loop of 80 mm and 60 mm loop. At the end of the experiment, the best result among the set loop of the sensor is characterized for the sensor parameter. The characterization of the sensor is based on performance parameters of resolution, sensitivity, repeatability error, and linearity of the reading and the obtained characteristic is compared with another studies.

3. Experimental Results

3.1 Coupling Power Ratio

Before the fabrication of the sensor, the coupling power response is studied and the result is being used as the reference element and for the verification of the sensor. Based on the coupling power ratio curve in Figure 3 and Figure 4, both losses in receiving fiber increase along with the decreasing bent diameter. This happened due to the increases of both losses in the illuminating fiber where with the side coupling effect, the light propagated from the illuminating fiber is radiated based on the evanescent wave theorem to receiving fiber. Most of the radiated power propagates parallel with the source but some of the power is refracted toward the back-end of the receiving fiber which is known as scattered-bend coupling power. When the fiber loop is pulled, the bend diameter is decreasing which causes the variation of coupling power.

Based on Table 1, the losses of the light in receiving fiber are decreasing at the forward end of the fiber from -44.06 dBm at 100 mm to -32.15 dBm at 20 mm. This also same goes to the power received at the back-end where the light losses also show a significant decrease from -51.50 dBm at 100 mm to -41.81 dBm at 20 mm. From observed Table 1, it is used for the displacement sensing and as guidance or reference to validate the sensor.

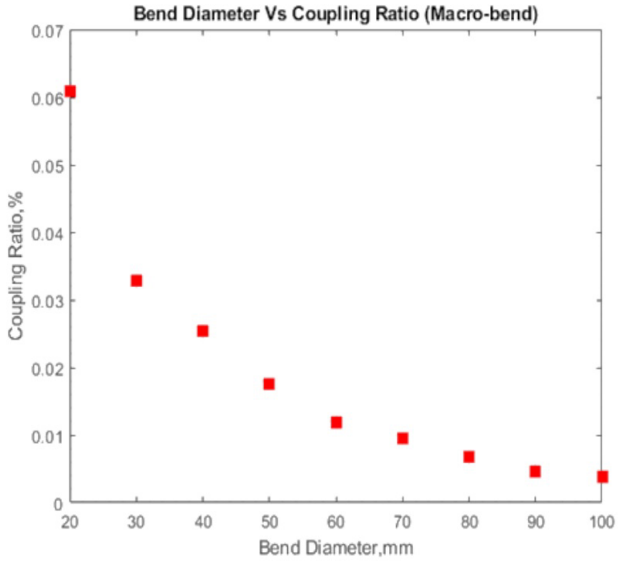


Figure 3: Macro-bend coupling power ratio.

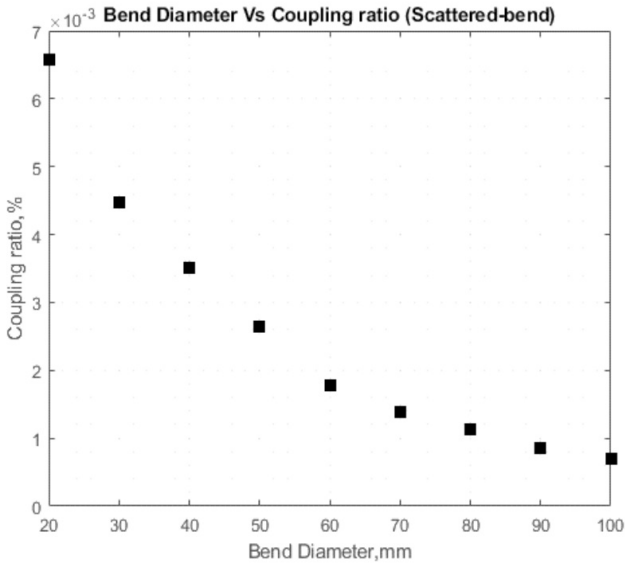


Figure 4: Scattered-bend coupling power ratio.

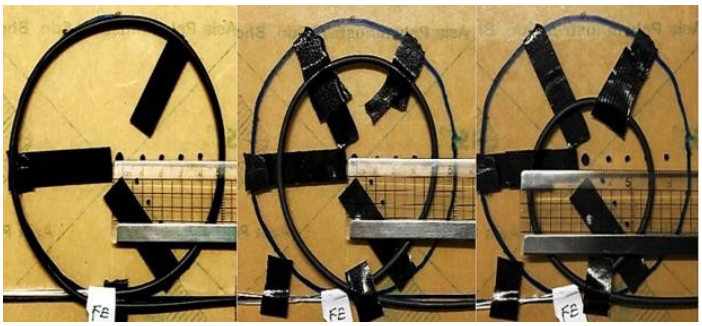
Table 1: Power losses from the reference power output.

	Power Received (dBm)		Total Power (dBm)	
Diameter (mm)	Macro-bend	Scattered-bend	Macro-bend	Scattered-bend
100	-35.95	-43.39	-44.06	-51.50
90	-35.18	-42.56	-43.29	-50.67
80	-33.60	-41.34	-41.71	-49.45

70	-32.08	-40.49	-40.19	-48.60
60	-31.16	-39.37	-39.27	-47.48
50	-29.46	-37.64	-37.57	-45.75
40	-27.85	-36.42	-35.96	-44.53
30	-26.72	-35.38	-34.83	-43.49
20	-24.04	-33.70	-32.15	-41.81

3.2 Displacement Sensing Test

Figure 5 shows the initial bending diameter of the twisted bend part at 100 mm, 80 mm, and 60 mm used in the sensor structure. The bending diameter of the structure decreased in 10 mm at each test which increases the coupled power received at receiving fiber.



(a) (b) (c)

Figure 5: Initial bending diameter of sensor structure for (a) 100 mm, (b) 80 mm and (c) 60 mm

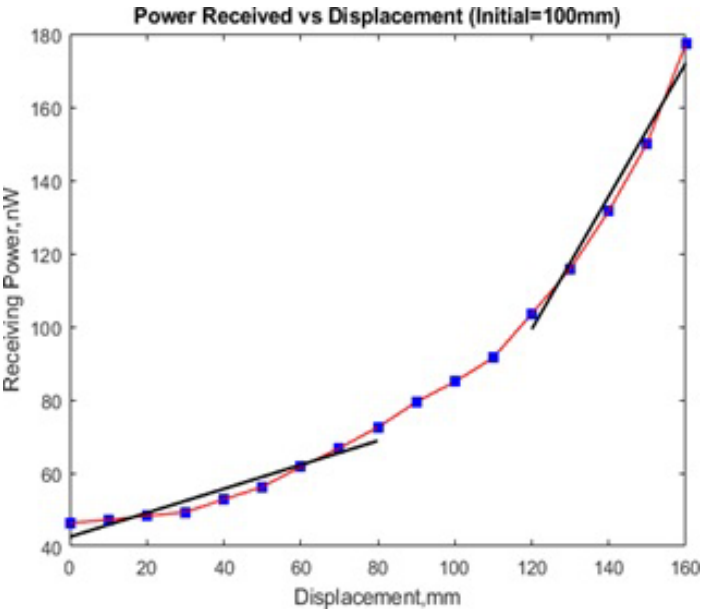


Figure 6: Received power at back-end of receiving fiber (Scattered-Bend coupling power).

Figure 6 showing the power received for the initial 100 mm bend diameter. It clearly shows that when the fiber is dragged 10 mm in each reading, the power received also increases. For macro-bend coupling power, it is producing an exponential curve relation as shown in Figure 7. From the experiment, both losses have producing good repeatability.

The step of the experiment is repeated by changing the loop bend diameter to 80 mm then 60 mm. As for a structure with an initial loop bend diameter of 100 mm, the smallest bend diameter at 160 mm displacement length is 50 mm. While the smallest diameter for initial loop bend of 80 mm is 30 mm at 160 mm displacement length, the smallest diameter for initial loop of 60 mm is 20 mm at 150 mm displacement length.

The result obtained can be verified by comparing the value of the power received of the tested sensor with the coupling power ratio graph.

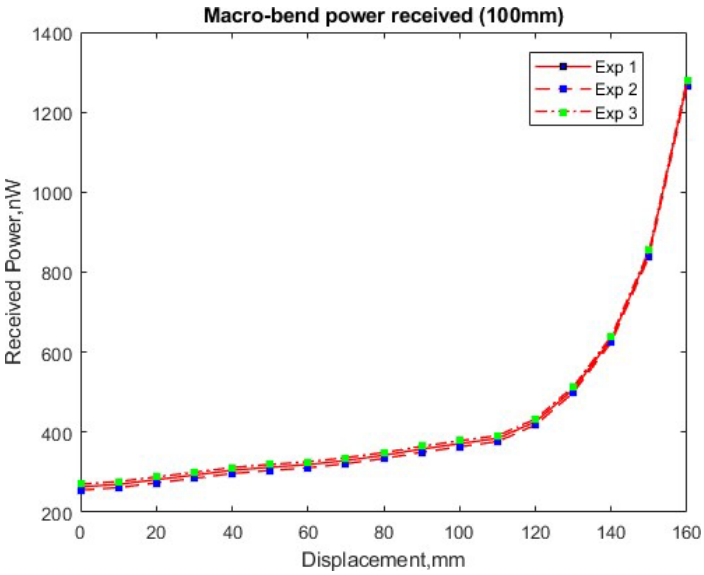


Figure 7: Macro-bend coupling received power at forward-end of receiving fiber.

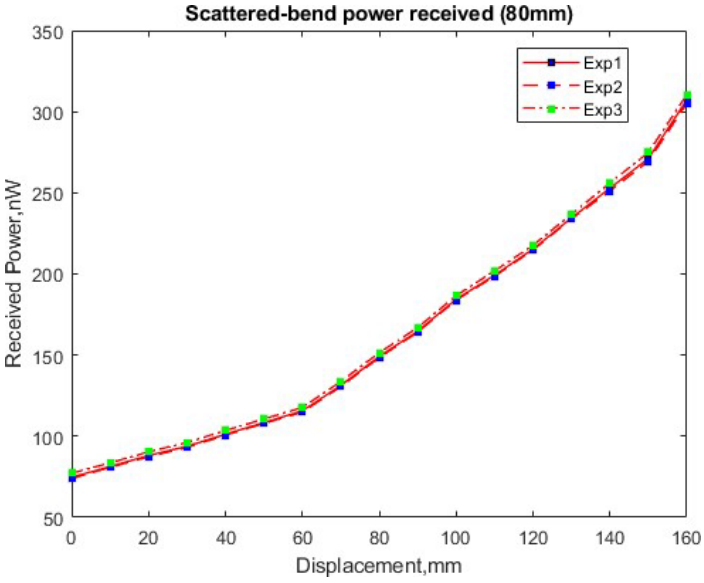


Figure 8: Scattered-bend coupling received power at forward-end of receiving fiber of 80 mm structure.

The power received should be in the range between the initial diameter and smallest diameter based on the coupling power ratio as shown in Figure 3 and Figure 4. The structure with 80 mm and 60 mm initial bend loop also shows a huge power gap between the changes of bend diameter for scattered-bend losses. This is

because the characteristic of both losses is polynomial to exponential-like relation as in Figure 8 to Figure 11. This is also due to the increases of both losses at illuminating fiber which then transferred to receiving fiber by side coupling effect.

As in Figure 8 and Figure 9, both graphs showing the increases of the coupled power of scattered-bend for initial bend loop of 60 mm and 80 mm. In terms of the differences between received power, the coupling power of the macro-bend is particularly high compared to scattered-bend for all bending structures as shown in Figure 10 and Figure 11. This phenomenon happened because the macro-bending loss is producing a much higher loss due to the stress of the fiber which makes the refraction angle inside the fiber core is changes and then the light is radiated out from the core to the cladding. Due to the macro-bend is mostly propagate parallel with the light source the power received at the forward-end of the receiving fiber is higher compared to the back-end.

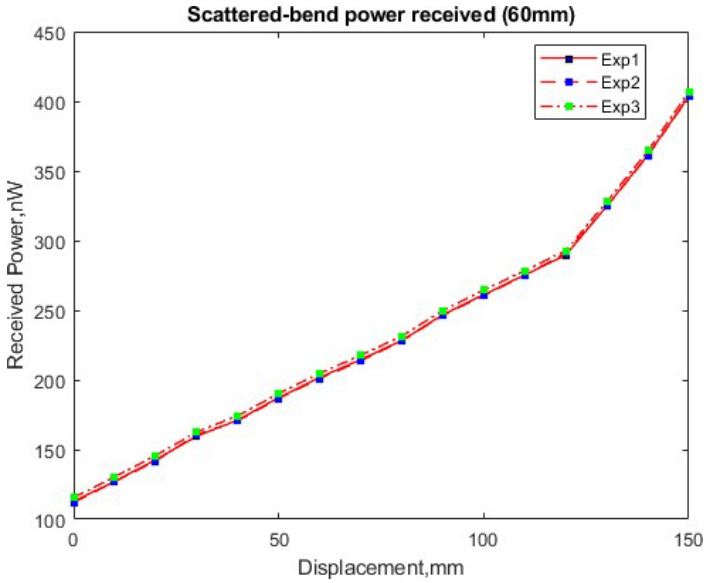


Figure 9: Scattered-bend coupling received power at forward-end of receiving fiber of 60 mm structure.

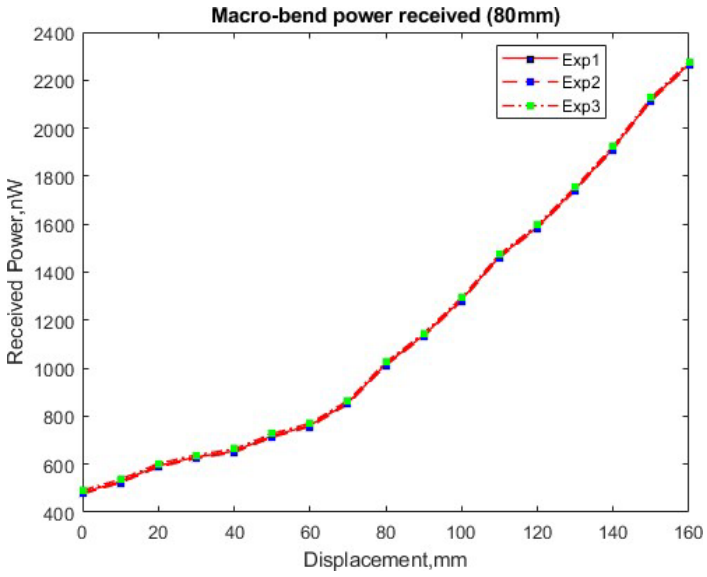


Figure 10: Macro-bend coupling received power at forward-end of receiving fiber of 80 mm structure.

While coupling power of scattering-bend loss is much lower because the scattering losses are considered a minor loss compared to macro-bend. This is because the scattering losses are caused by the density fluctuation and core defect which commonly ignore. Therefore, the coupled power of the scattered-bend in receiving fiber is very low.

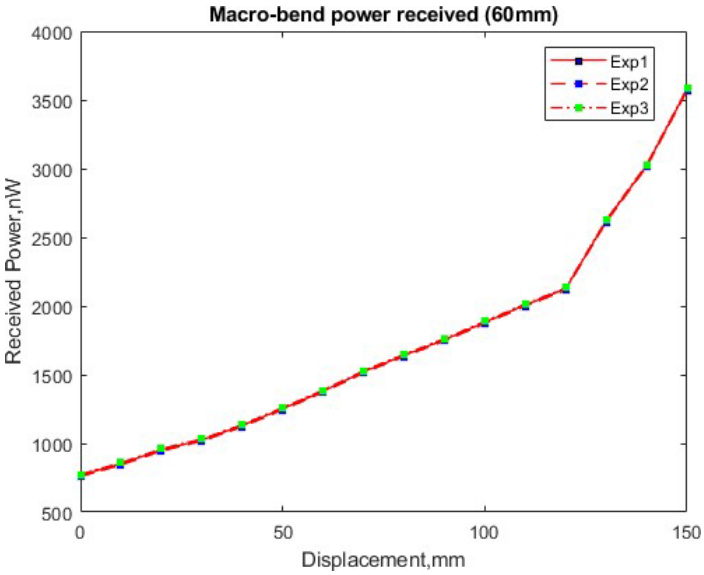


Figure 11: Macro-bend coupling received power at forward-end of receiving fiber of 60mm structure.

In comparison, the coupling power of scattered-bend loss is much stable, and the response showed by the scattered-bend loss coupling power towards the increases of displacement is much more suitable in displacement sensing application compared to macro-bend effect and characteristic of the proposed sensor has high reliability.

4. Sensor Characterization

The sensor characterization is a crucial part of sensing application. A sensor characterization is required to assess if the produced sensor is a good sensor [31]. In this experiment, the characterization parameters of the sensor that have been determined are sensitivity, linearity, resolution, and repeatability error as stated in Table 2. The characteristic will also be compared with other fiber-based displacement sensors in terms of the displacement range. The characterization is analyzed based on static measurement analysis. Based on the tested sensor, the best characteristic sensor is the structure that have an initial bend diameter of 100 mm, because the sensor is much more sensitive with the value of 0.817nW/mm, has a reasonable resolution value of 1.228 mm, and a small repeatability error of 1.858% compared to initial bend loop of 80 mm and 60 mm. Based on the comparison between techniques as in Table 3, the sensor proposed has one of the highest achieved displacement ranges which is up to 160 mm. If the tested sensor is compared with the linearity, the proposed sensor has good linearity at 0 mm to 80 mm and 120 mm to 160 mm where each linearity has a R^2 of 0.9182 and 0.9777. The sensitivity of the sensor is 0.817mW/mm which was calculated by using:

$$S = \frac{\Delta P_2}{\Delta d} \quad (3)$$

Where S is the sensitivity of the sensor, ΔP_2 is changes of received power at the back-end of the fiber, and Δd is changes of displacement. For the repeatability error, the sensor is tested three times at each test using the repeatability testing method where statistical mathematics is used by calculating the pooled standard deviation of the output.

Table 2: Fabricated sensor performance parameters.

Parameter	Measured Value	Reference Value
Range	0 mm – 160 mm	150mm -160mm
Sensitivity	0.817nW/mm	0.1nW/mm to 5nW/mm
Resolution	1.228mm	0.1mm to 1.5mm
Linearity	y = 0.3282x + 42.55 R^2= 0.9182	0.8< R^2 ≤1
Repeatability Error	1.856%	2% to 1%

Table 3: Comparison between the technique in sensor structure.

Technique	Displacement Range
Twisted coupled macro-bend [32]	0mm – 140mm
Diffraction grating Ended [25]	4mm – 14mm
Dual-wavelength compensation [33]	0mm – 10mm
Twisted coupled scattered-bend	0mm – 160mm

5. Conclusions

In this research, a displacement measurement displacement sensor is designed by using a polymer optical fiber (POF) where the sensing part is utilizing the scattered-bend loss by side coupling method utilizing macro-bend effect. A scattered-bend loss is a combination of the scattering losses with the bending loss of the fiber where the losses are caused by density fluctuation and physical bending of the fiber. The measurement of the scattered-bend coupled power is measured at the back-end and the macro-bend coupled power at the forward-end of the receiving fiber. Most of the losses generated propagated parallel with the light source and some of the losses are refracted toward the back-end of the receiving fiber. This phenomenon explains the reason coupled power received at the forward-end of the receiving fiber is higher compared to the power received at the back-end of the fiber. The fabricated sensor can detect a measurement of displacement up to 160 mm with a sensitivity of 0.817 nW/mm, resolution of 1.228 mm, and repeatability error of 1.856%. The fabricated sensor also has a simple structure and analysis, low cost and easy to set up. The sensor also has a high potential advantages on the industrial application such as civil structuring, building surveyor, architecture, earth movement, landslides and medicine. The future works may include the utilization of IoT subsystem to be part of the system where the data collected can be analyzed beforehand and send to the users.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work would not have been possible without the financial and facility support of the Department of Electrical and Electronics Engineering, Faculty of Engineering, National Defence University of Malaysia.

References

[1] L.S. Supian, D.H.M. Razali, S.A. Syed Ali, "Investigation of Scattered-Bend Loss Coupling in Polymer Optical Fiber (POF) - Based Displacement Measurement Sensor," in 2022 IEEE 9th International Conference on Photonics (ICP), IEEE: 13–14, 2022, doi:10.1109/ICP53600.2022.9912445.

[2] A. Zendehnam, M. Mirzaei, A. Farashiani, L. Horabadi Farahani, "Investigation of bending loss in a single-mode optical fibre," Pramana, **74**(4), 591–603, 2010, doi:10.1007/s12043-010-0052-5.

[3] S. Addanki, I.S. Amiri, P. Yupapin, "Review of optical fibers-introduction and applications in fiber lasers," Results in Physics, **10**, 743–750, 2018, doi:10.1016/j.rinp.2018.07.028.

[4] A.W. Snyder, "Coupled-Mode Theory for Optical Fibers," Journal of the Optical Society of America, **62**(11), 1267, 1972, doi:10.1364/JOSA.62.001267.

[5] Z. Wang, H. Wu, X. Hu, N. Zhao, Q. Mo, G. Li, "Rayleigh scattering in few-mode optical fibers," Scientific Reports, **6**(1), 35844, 2016, doi:10.1038/srep35844.

[6] K. Tian, Y. Liu, Q. Wang, "Temperature-independent fiber Bragg grating strain sensor using bimetal cantilever," Optical Fiber Technology, **11**(4), 370–377, 2005, doi:10.1016/j.yofte.2005.03.001.

[7] W.J. Bock, J. Chen, P. Mikulic, T. Eftimov, "A Novel Fiber-Optic Tapered Long-Period Grating Sensor for Pressure Monitoring," IEEE Transactions on Instrumentation and Measurement, **56**(4), 1176–1180, 2007, doi:10.1109/TIM.2007.899904.

[8] G. Perrone, A. Vallan, "A Displacement Measurement System Based on Polymer Optical Fibers," in 2008 IEEE Instrumentation and Measurement Technology Conference, IEEE: 647–651, 2008, doi:10.1109/IMTC.2008.4547116.

[9] R. Correia, S. James, S.-W. Lee, S.P. Morgan, S. Korposh, "Biomedical application of optical fibre sensors," Journal of Optics, **20**(7), 073003, 2018, doi:10.1088/2040-8986/aac68d.

[10] N. Albakri, S. Abdullah, L.S. Supian, N. Arsad, S.D. Zan, A.A.A. Bakar, "Assessment of Palm Oil Fruit Bunch Maturity based on Diffuse Reflectance Spectroscopy Technique," in 2018 IEEE 7th International Conference on Photonics (ICP), IEEE: 1–3, 2018, doi:10.1109/ICP.2018.8533172.

[11] L.S. Supian, A.M.A. Amboalang, U.F.A. Rauf, K. Ismail, C.S. Ping, N.F. Naim, "Qualitative Assessment of Cooking Oil using Diffuse Reflectance Spectroscopy Technique," in 2022 International Conference on Green Energy, Computing and Sustainable Technology (GECOST), IEEE: 221–226, 2022, doi:10.1109/GECOST55694.2022.10010506.

[12] S. Thomas Lee, R. Dinesh Kumar, P. Suresh Kumar, P. Radhakrishnan, C.P.G. Vallabhan, V.P.N. Nampoori, "Long period gratings in multimode optical fibers: application in chemical sensing," Optics Communications, **224**(4–6), 237–241, 2003, doi:10.1016/S0030-4018(03)01597-9.

[13] A.B.L. RIBEIRO, J.L. SANTOS, J.M. BAPTISTA, L.A. FERREIRA, F.M. ARAÚJO, A.P. LEITE, "Optical Fiber Sensor Technology in Portugal," Fiber and Integrated Optics, **24**(3–4), 171–199, 2005, doi:10.1080/01468030590922722.

[14] W. Du, X.M. Tao, H.Y. Tam, C.L. Choy, "Fundamentals and applications of optical fiber Bragg grating sensors to textile structural composites," Composite Structures, **42**(3), 217–229, 1998, doi:10.1016/S0263-8223(98)00045-2.

[15] Y. Koike, T. Ishigure, M. Sato, E. Nihei, "Polymer optical fibers," in 1998 IEEE/LEOS Summer Topical Meeting. Digest. Broadband Optical Networks and Technologies: An Emerging Reality. Optical MEMS. Smart Pixels. Organic Optics and Optoelectronics (Cat. No.98TH8369), IEEE: III/13-III/14, doi:10.1109/LEOSST.1998.690041.

[16] L. Bilro, N. Alberto, J.L. Pinto, R. Nogueira, "Optical Sensors Based on Plastic Fibers," Sensors, **12**(9), 12184–12207, 2012, doi:10.3390/s120912184.

[17] W.E. van de Meent, EXPERIMENTAL DEMONSTRATION OF REDUCED BEND LOSSES IN LOW-CONTRAST POLYMER HYBRID WAVEGUIDES, 2015.

[18] D. Sartiano, S. Sales, "Low Cost Plastic Optical Fiber Pressure Sensor Embedded in Mattress for Vital Signal Monitoring," Sensors, **17**(12), 2900, 2017, doi:10.3390/s17122900.

[19] T. Eftimov, Sensor Applications of Fiber Bragg and Long Period Gratings, Springer Netherlands, Dordrecht: 1–23, doi:10.1007/978-1-4020-6952-9_1.

[20] K.O. Hill, B.S. Kawasaki, D.C. Johnson, Y. Fujii, Nonlinear Effects in Optical Fibers: Application to the Fabrication of Active and Passive Devices, Springer US, Boston, MA: 211–240, 1979, doi:10.1007/978-1-4684-3492-7_12.

[21] K. Kurihara, H. Ohkawa, Y. Iwasaki, O. Niwa, T. Tobita, K. Suzuki, "Fiber-optic conical microsensors for surface plasmon resonance using chemically etched single-mode fiber," Analytica Chimica Acta, **523**(2), 165–170, 2004, doi:10.1016/j.aca.2004.07.045.

[22] A.D. Kersey, T.A. Berkoff, W.W. Morey, "High-resolution fibre-grating based strain sensor with interferometric wavelength-shift detection," Electronics Letters, **28**(3), 236, 1992, doi:10.1049/el:19920146.

[23] D.S. Montero, C. Vázquez, "Polymer Optical Fiber Intensity-Based Sensor for Liquid-Level Measurements in Volumetric Flasks for Industrial Application," ISRN Sensor Networks, **2012**, 1–7, 2012, doi:10.5402/2012/618136.

[24] Polymer Optical Fiber-Based Sensors, EPFL Press: 365–408, 2011, doi:10.1201/b16404-14.

[25] M. Lomer, J. Zubía, J. Arrue, J.M.L. Higuera, "Principle of functioning of a self-compensated fibre-optical displacement sensor based on diffraction-grating-ended POF," Measurement Science and Technology, **15**(8), 1474–1478, 2004, doi:10.1088/0957-0233/15/8/007.

[26] L.S. Supian, M.S. Ab-Rahman, N. Arsad, "Polymer optical fiber tapering using chemical solvent and polishing," EPJ Web of Conferences, **162**, 01018, 2017, doi:10.1051/epjconf/201716201018.

[27] N. Uddin, M.R. M, S. Ali, "Performance Analysis of Different Loss Mechanisms in Optical Fiber Communication," Computer Applications: An International Journal, **2**(2), 1–13, 2015, doi:10.5121/caij.2015.2201.

[28] S. Savović, A. Djordjevich, I. Savović, "Theoretical investigation of bending loss in step-index plastic optical fibers," Optics Communications, **475**, 126200, 2020, doi:10.1016/j.optcom.2020.126200.

[29] C.-A. Bunge, R. Kruglov, H. Poisel, "Rayleigh and Mie scattering in polymer optical fibers," Journal of Lightwave Technology, **24**(8), 3137–3146, 2006, doi:10.1109/JLT.2006.878077.

[30] R.T. Schermer, J.H. Cole, "Improved Bend Loss Formula Verified for Optical Fiber by Simulation and Experiment," IEEE Journal of Quantum Electronics, **43**(10), 899–909, 2007, doi:10.1109/JQE.2007.903364.

[31] C.A.F. Marques, D.J. Webb, P. Andre, "Polymer optical fiber sensors in human life safety," Optical Fiber Technology, **36**, 144–154, 2017, doi:10.1016/j.yofte.2017.03.010.

[32] J. Liu, Y. Hou, H. Zhang, P. Jia, S. Su, G. Fang, W. Liu, J. Xiong, "A Wide-Range Displacement Sensor Based on Plastic Fiber Macro-Bend Coupling," Sensors, **17**(12), 196, 2017, doi:10.3390/s17010196.

[33] A. Vallan, M.L. Casalicchio, M. Olivero, G. Perrone, "Assessment of a Dual-Wavelength Compensation Technique for Displacement Sensors Using Plastic Optical Fibers," IEEE Transactions on Instrumentation and Measurement, **61**(5), 1377–1383, 2012, doi:10.1109/TIM.2011.2180975.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 2, 130-138 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Detecting CTC Attack in IoMT Communications using Deep Learning Approach

Mario Cuomo, Federica Massimi, Francesco Benedetto*

Signal Processing for Telecommunications and Economics Lab., Roma Tre University, Rome, Italy

ARTICLE INFO

Article history:
Received: 28 December, 2022
Accepted: 05 April, 2023
Online: 28 April, 2023

Keywords:
Covert Timing Channel
TCP Protocol
Convolutional Neural Network
Siamese Neural Network
K-Nearest Neighbors
E-Health Security

ABSTRACT

Cyber security is based on different principles such as confidentiality and integrity of transmitted data. One of the main methods to send confidential messages is to use a shared secret to encrypt and decrypt them. Even if the amortized computational complexity of the hashing functions is O(1), there are several situations when it is not possible to use them due to the lack of computing power or the need to keep completely hidden the communication to other parties in the network. Covert Channels (CCs) are an excellent alternative in all these cases because they hide the private message in legitimate communication channels without the need to allocate additional resources to communicate. For this reason, they are difficult to identify because they are fully camouflaged in legitimate traffic. Unfortunately, CC technique is also used by hackers to exfiltrate network data and initiate cyber-attacks against devices in the system: Internet of Medical Things (IoMT) are one of the most vulnerable devices affected by this type of attack. It is therefore essential to create a system that can autonomously identify the presence of a malicious CCs to safeguard the health of patients. This paper describes an approach to create a Covert Timing Channel (CTC) based on TCP packets between client and server and how it is possible to detect the hidden communication using an innovative pipeline composed by several Machine Learning (ML) and Deep Learning (DL) models, such as Convolutional Neural Network (CNN), Siamese Neural Network (SNN) and K-Nearest Neighbors (K-NN). Considering 4 different message types exchanged in CTC, the proposed pipeline achieved 94% accuracy in identifying covert messages in the channel.

1. Introduction

In a world that is becoming increasingly and wirelessly connected, network security is now a critical task that must be seriously considered. It is necessary to avoid cybercriminals gaining illegal access to valuable data and sensitive information. It is important to note that the amount of data that devices produce, and the number of resources used, increase as more devices are connected. When an unauthorized user gets hold of data, he can cause several problems such as stolen assets, identity theft and reputational damage – not only to the individual but also to the entire network. A vulnerability can be described as a situation where a subject A (item, process or person) manages to exploit the privileges of a subject B to carry out operations not initially granted to him. Therefore, it is necessary to proactively manage risks, threats, and vulnerabilities.

Many studies have recently focused on Information Technology (IT) resilience. It describes the ability of a system to continue to deliver the expected results despite the occurrence of incidents, such as natural disasters and especially cyber-attacks.

In this scenario, Internet of Things (IoT) devices are particularly vulnerable to network attacks and the situation becomes extremely dangerous when we consider e-health devices. E-health data represents one of the most important personal information. It is important to design the system as confidential as possible, with high-level security policies. Even if various regulations for data management have been drafted over the years – such as the General Data Protection Regulation (GDPR, https://gdpr-info.eu/) – it is not uncommon to read news of improper exfiltration of data by unauthorized users: according to Protenus Breach Barometer, in 2022 there were 50M+ Patient records breached, 905 Incidents, 44% Increase in hacking incidents (https://www.protenus.com/breach-barometer-report).

*Corresponding Author: Francesco Benedetto, francesco.benedetto@uniroma3.it

The Internet of Medical Things (IoMT) [1] is that set of technologies aimed at using smart devices – connected to each other, even via the internet – in the medical field. If this technology guarantees an improvement in healthcare management, on the other hand new security challenges are expected: it is necessary to ensure correct authentication and authorization procedures (applying minimum privilege as much as possible), maintaining the confidentiality of data (both at rest and in transit, with encryption and obfuscation techniques) and integrity (making sure that the data is not modified by malicious users).

For obtain confidential communications, cryptography is used. It is a technique for encoding messages: symmetric encryption is based on sharing a shared secret; asymmetric cryptography is based on a pair of keys – public key and private key – used respectively to sign a message and to verify its integrity [2] . Given the low computing power, the encryption algorithms used in the IoMT are different from those used in servers, and they are classified into three categories: centralized (*i*), non-centralized (*ii*), low weight (*iii*) [3]. The centralized approach (*i*) uses a central node – often a server and not an IoT device – to encrypt the message. The sender sends the message in clear text over a secure channel to the centralized server which encrypts it and sends it over the potentially insecure channel to the receiver. The central node requires a lot of computing power, and it is a single point of failure (see Figure 1).

In the decentralized approach (*ii*) the encryption is distributed on the various link-by-link nodes: one node receives the encrypted message, decrypts it, re-encrypts it and transmits it to the next node. There is additional encryption level between end systems (see Figure 2).

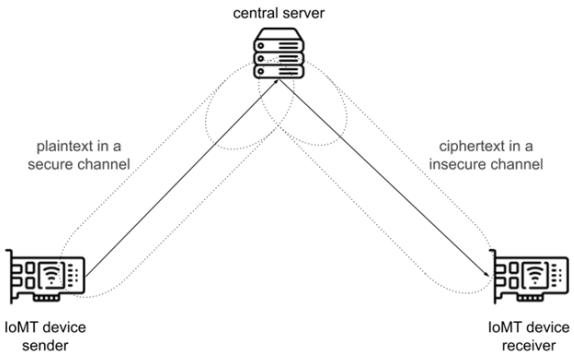


Figure 1: Centralized approach in IoMT cryptography

To minimize the effort of the devices, over the years various approaches have been proposed which are based on symmetric encryption in which each network node authenticates the others [4]. These algorithms belong to the low weight security approach (*iii*).

Obfuscation techniques, in accordance with the principles of least privilege, aim to make data inaccessible when it is not needed. Unlike encryption algorithms that use a key, to understand the plaintext you only need to know the algorithm for generating the obfuscated data [5]. More sophisticated techniques allow to completely anonymize the data, no longer allowing the re-identification of the data after anonymization [6].

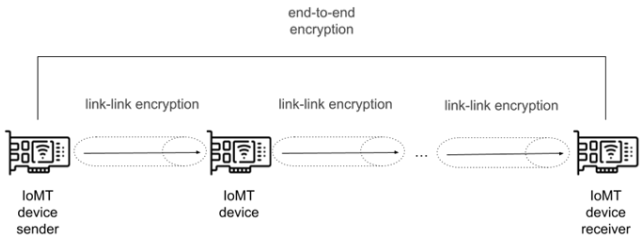


Figure 2: Decentralized approach in IoMT cryptography

The techniques and attacks used by hackers to compromise a system have become increasingly sophisticated so that human observation is less and less useful in identifying a compromise.

It is precisely here that Machine Learning (ML) in cybersecurity comes into play: in fact, the use of ML lends itself greatly to solving this type of problem. The systems can analyze patterns and learn from them to help prevent similar attacks and respond to behavior change. Generalization is the capacity of an ML model to fit correctly to additional, previously unobserved data taken from the same distribution as the model's original data. Because ML can learn from past data, it can recognize odd communications and user behaviors. As a result, it is possible to prevent threats and respond to active attacks in real time, not only reducing the amount of time spent on routine tasks but also allowing us to use the network's resources more strategically.

Artificial Intelligence (AI) approaches are widely used in healthcare: to preserve the privacy of health data from access by unauthorized users, several frameworks have been proposed for analyzing user behavior in a secure environment. Behavioral anomalies – such as unusual login and actions – are flagged by the system as suspicious and blocked [7]. Several vendors – e.g., Microsoft and Google – have implemented these security systems called Security Information and Event Management (SIEM). They are capable of monitoring and identify possible vulnerabilities and acting to mitigate them. Each SIEM is always composed of at least three main modules which are data collection (*i*), learning the normal flow without malicious actions (*ii*) and generating the classification report (*iii*) [8].

Over the years SIEM's training baseline has evolved, moving from using legacy Machine Learning models (e.g., Support Vector Machine, Decision Tree, K-NN [9]) to using Deep Learning (DL) models [10]: researchers have proposed systems running MLP-based Neural Networks [11], DNNs with optimization techniques (Principal Component Analysis and Gray Wolf Optimization) [12], Convolutional Neural Networks with Long Short-Term Memory [13].

Cyberattacks from malicious users are increasingly accurate and range across the entire ISO-OSI stack: Perception-level attacks can involve Denial of Service (DoD) of physical devices or RFID spoofing and cloning; Application layer attacks can create a Man in The Middle (active or passive) by sniffing network traffic by

installing Malware or Medical information injections. At network level, attackers can invalidate DNS or ARP tables to redirect traffic to their destinations [14]. In this scenario, Covert Channels (CC) assume a great importance. Covert Channels are channels used to transmit information using existing system resources that were not designed to carry data. They make it possible not to show the communication taking place between two interlocutors in order not to alarm a third agent – potentially malicious and looking to exfiltrate data. The main characteristics of a CC are stealthies, low bandwidth and indistinguishability. Due to their ability to evade detection, they pose a serious threat to cyber security because attackers can use them for malicious scopes [15]. There are various types of Covert Channels: Covert Timing Channel (*i*), Covert Storage Channel (*ii*), Covert Behavioral Channel (*iii*). Covert Timing Channels (*i*) use a time measurement to signal the value to be sent on the channel; Covert Storage Channels (*ii*) encode information by hiding it in the fields of the network protocol used; Covert Behavioral Channels (*iii*) divide the hidden message to be sent and transmit it in smaller packets and generally using a lower-level protocol.

More recently, ML and statistical methods for detecting CTC attacks communications were presented such as temporal analysis (*i*), traffic analysis on the channel (*ii*), the observation of side channels (*iii*) and the study of entropy (*iv*).

In the temporal analysis (*i*), the computing times of the devices are analyzed to identify any anomalies: if the response time of a device varies abruptly, the presence of a CTC involving that device can be assumed. Unfortunately, this approach is subject to jitter: legitimate traffic on the channel stresses the devices and a false CTC alarm can be raised [16].

The traffic analysis on channel (*ii*) analyzes the traffic but takes into account factors like the frequency of packet sending, their volume, and occasionally even their content. By gathering the network traffic exchanged, a statistical model of the system is constructed in a secure environment, and from these the significant communication frequencies are discovered. The network traffic is evaluated in the detection phase, and the presence of a CTC can be suspected if there are several significant frequencies [17].

The execution of Covert Channels has unintended consequences for the systems; the variation of the side channels can be investigated (*iii*) to spot any network anomalies like CTC attacks. It is extremely challenging to isolate the various processes from one another in a system with highly interconnected components. A message exchange over unconventional channels is referred to as a CC when both the sender and the recipient are aware of it, and a SC when the message is sent by the sender involuntarily, such as through cache access, data movement between the CPU and memory, or the processor emitting electromagnetic waves. The identification of CCs by SC is still under study and there is no proof of its correctness [18].

The system's entropy can be affected by a CTC, and as a result, this measurement can offer helpful information for detection. Information Theory uses entropy as a metric to quantify the degree of disorder or uncertainty in a system. Entropy is a measure of how random and erratic messages are exchanged in the network about the problem of detecting a Covert Timing Channel. A legitimate communication system, as opposed to one that sends messages according to a set of rules and behaviors (such as the CTC-TR), is unpredictable and has a high degree of entropy. In [19], the authors demonstrate a method based on entropy (*iv*) using conditional entropy, which is an estimation of the system's entropy's value obtained from available data.

In recent years, ML models have been used in Covert Timing Channel identification systems: starting with network traffic that has been detected and then being cleaned up by removing unnecessary data, AI models are trained on the resulting data. KNN, SVM, and Naive Bayes are three of the most popular ML models [20].

With the aid of VGG-16 and Squeeze Net, the first instances of the application of DL approaches for the identification of CTCs and their validity can be seen [21].

In this paper, the hidden communication is embedded in the legitimate traffic by means of a CTC, obtained by modulating the inter-arrival packet delays. In practice, the malicious process modulates the inter-arrival delay of the transmitted e-health data, by transmitting one 0 bit (or 1 bit) when the delay is less (greater) than a pre-defined threshold. Here, we move one step further by proposing a type of CTC, based on the inter-arrival delays of TCP packets. In addition, we implement an ML and DL framework to detect what kind of message is transmitted on the channel in CTC (see Figure 3).

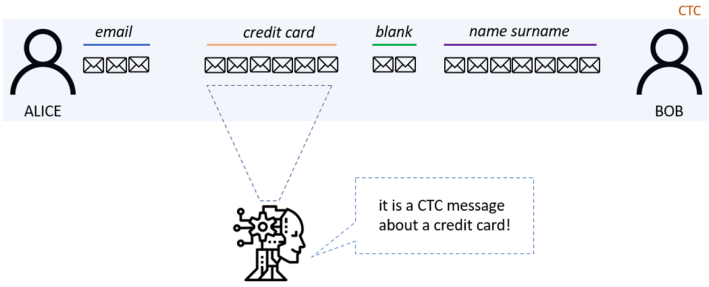


Figure 3: CTC scenario under consideration

The obtained results thus confirm the validity of such approach for ML and DL detection of hidden communications – the current State of Art in detecting CTCs is based on the analysis of the statistical variation of traffic on the network: as soon as this changes and exceeds a chosen threshold, an alert of a possible attack is sent; unfortunately this approach fails to detect highly stealthy traffic: a CTC uses the same throughput as the legitimate communication channel. For this reason, a DL model that can capture the insight hidden in the channel itself is needed.

The remainder of our work is organized as follows. Section 2 illustrates a Covert Timing Channel model and how the dataset of hidden communications is generated. Section 3 shows the ML and

DL methods used for the detection of illegitimate traffic, while Section 4 discusses the simulation results and shows the performance of the methods according to the main metrics used for evaluating the proposed pipeline.

2. Covert Timing Channel and Dataset Generation

For simplicity, we indicate two interlocutors as Alice and Bob. They establish a CTC communication. There is also a third user on the channel, Cindy, who is listening and has the aim of recovering the type of message that Alice and Bob are exchanging (see Figure 4). Let's consider Alice as a user with an active role in the communication: she is the only one who sends messages in the channel, while Bob is listening for them. It's not hard to think a real-life use case where an edge device sends messages to a central system hub to notify an event.

To understand how this channel works, it is necessary to start from a basic concept: what Alice sends on the channel is completely unrelated to what she is communicating. What's really related is how Alice is sending the message. As previously described, in a CTC the message is encoded in the packet interarrival time. We therefore distinguish between a covering message and a covered message: the first one is the message that Cindy recovers by sniffing the traffic, the second is the one that only Bob can reconstruct.

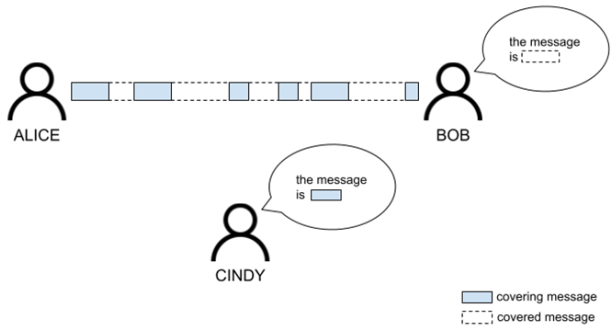


Figure 4: Scenario under consideration

Let's see how it is possible to encode the message that Alice wants to send using the packet interarrival times. Let x be the message to send. The first operation that Alice performs is the conversion of x into binary: each ASCII character can be represented using 7-bit sequence according to Table 1.

Table 1: Encoding ASCII to 7 bits string

	Leftmost three bits							
Righmost four bits	000	001	010	011	100	101	110	111
0000	NUL	DLE	Space	0	@	P	’	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	“	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EOT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	‘	7	G	W	g	w
1000	BS	CAN	(	8	H	X	h	x
1001	HT	EM	)	9	I	Y	i	y
1010	VF	SUB	*	:	J	Z	j	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	^	n	~
1111	SI	US	/	?	O	_	o	DEL

Let $binaryx$ be the binary string representing x. Note that $binaryx$ will be 7 times the length of the initial string x. At this time $7|x|$ packets will be sent on the channel – one for each bit of $binaryx$. Each packet will be sent waiting a specifying time after the previous depending on whether we want to send a 0 bit or a 1 bit (e.g., 0 waiting 10 milliseconds, 1 waiting 50 milliseconds). It is necessary to have a specific packet to notify the beginning and the ending of the message.

The packets sent by Alice are simple TCP frame for communication at level 4 of the ISO-OSI stack and each of them contains a character of the covering message. It is obvious that the more the covering message makes sense, the less Cindy will be suspicious of the presence of a hidden communication between Alice and Bob.

x = hello
$seed\_covert\_message$ = apple

$binaryx$ = 1101000 1100101 1101100 1101100 1101111
$covert\_message$
= $appleappleappleappleappleappleappleapple$

Note that if the $seed\_covert\_message$ is shorter than the message to be sent, it must be repeated several times to cover completely it (see Figure 5).

Figure 5: Example of CTC messaging packets

The use case we have considered is the one where an edge device notifies to a central hub several messages relating to events of 4 different kinds: the request to generate an access token for a user starting from the $email$ or $name\ and\ surname$, checking the validity of a $credit\ card$ or a $blank\ communication$ used as heartbeat. For simplicity, both the edge device and the central hub are active on the same LAN by establishing a socket between them.

We created a dataset with 4 types of messages (see Table 2), developing the CTC described using python and sniffing communication using Wireshark.

Our analysis did not focus on internal packet analysis (known as Deep Packet Inspection) but rather we considered packet

interarrival time as a classification vector. Considering the interarrival times of the packets we have created the representative spectrograms of the communications [22]. To produce the spectrograms, the packet interarrival times were collected and considered as sampling instants of a chirp signal. A chirp signal has the characteristic that its frequency varies - increasing or decreasing - over time. A linear increase was considered. The Short-Time Fourier Transform (STFT) was applied to the chirp signal, obtaining a matrix representation in which each column contains an estimate of the short-term frequency content located in the time of the signal itself. The matrix is the spectrogram of the communication encoded in RGB space. Each communication is therefore not represented by a flow of packets but by a single spectrogram which contains its characteristics (Figure 6).

Table 2. Number of instances for each class

MESSAGE CLASS	# INSTANCES
blank	500
credit_card	500
email	500
name_surname	500

The input for Deep Learning models described later will be images of the size $224 \times 224 \times 3$ and for our experiment we divided it in training set and test set with $1:4$ ratio.

Figure 6: Example of Spectrogram

3. Machine and Deep Learning Models

Over the years several Machine Learning models have been proposed for the classification task and some of them have been considered for our experiments.

We have implemented the following models.

- Random Forest (Figure 7)
- K-NN (see Figure 8)
- Convolutional Neural Network (Figure 9)
- Siamese Neural Network (Figure10)

The Random Forest (RF) model is based on the construction of several Decision Trees (DT) – other classification models - and the final output is obtained by combining the outputs of the individual DTs by applying the majority vote approach. A DT is a tree whose internal nodes represent feature splits, and the leaves represent classes. The idea is to traverse the tree from the root and based by values of the instance to classify and the node splits arrive on a leaf node and assign the corresponding class. To decide how to create the tree – such as which feature use at which level – are used different criteria such as entropy, information gain and Gini index.

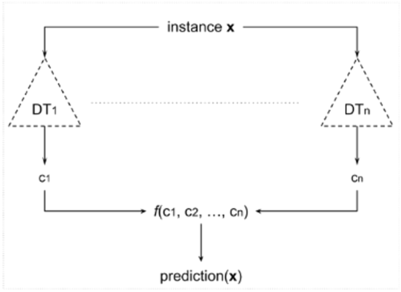


Figure 7: Schema of Random Forest model

To better understand how K-NN works it is necessary to introduce the Nearest Neighbour (NN). The idea of the NN is very simple: 2 instances of the same class are very similar to each other. Similarity can be calculated as the distance – consider the Euclidean distance – between the vectorial representations of the two instances. To classify an input, therefore, it is sufficient to retrieve the item closest to it and assign it the same class. KNN retrieves the K closest items and assigns to the input the class that has majority votes.

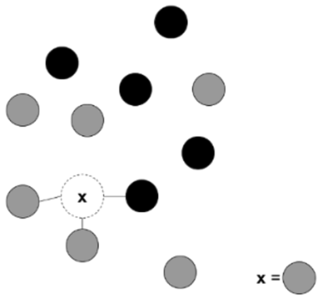


Figure 8: Example of K-NN prediction

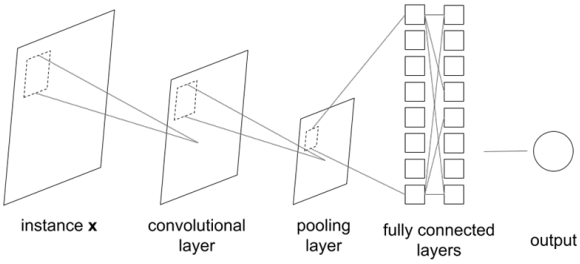


Figure 9: General schema of a CNN

A Convolutional Neural Network is a particular type of artificial neural network (ANN) widely used for many visual tasks such as image recognition, object classification and pattern recognition. It is composed in sequence of – at least – three types of layers: Convolution Layer (or Kernel), Pooling Layer and Fully Connected Layer. In the Convolutional Layer the image (a matrix of pixels) is input and a smaller matrix than the starting one is output. The output matrix is obtained by sliding an activation map over the input matrix and applying the dot product between it and the selected portion of the image. The goal of the pooling layer is

to reduce the spatial dimension of the representation by extracting the dominant features. In the Fully connected layer, we try to learn non-linear combinations of the characteristics of the representation obtained. The Fully Connected level is trying to learn the nonlinear function that connects input to output.

A Convolutional Siamese Network [23] has two images as input and returns the similarity between them. Internally it is composed of two – or more – CNN that share the same weights. When classifying an image with a convolutional network the last layer is almost always a layer with a SoftMax function: we obtain a vector of k elements, and k[i] contains the probability of confidentiality in assigning class i to the image of inputs.

In a Siamese Network the last layer is eliminated, and another one is added to calculate the difference between the two representations. The last level is a single neuron with a sigmoid activation function (0 to 1).

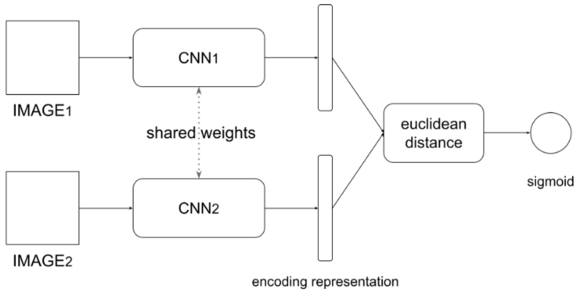


Figure 10: Siamese Network using CNNs

4. Results

To evaluate the performance of the system, some of the metrics mainly considered were used: Precision, Recall, F1 - Score, Accuracy and Specificity, that are calculated using True (T) or False (F) Positive (P) or Negative (N). A correctly classified instance is a True Positive; a misclassified instance can be a False Negative or a False Positive. Considering the $credit\_card$ class as Positive class: a spectrogram of a credit card is a True Positive if the system assigns it the $credit\_card$ class; a spectrogram of a name and surname is a False Positive if the system assigns it the $credit\_card$ class; a credit card spectrogram is a False Negative if the system does not assign it the $credit\_card$ class. These values can be obtained by considering the confusion matrix in Table 3.

Table 3: Confusion Matrix of a binary classification

	Really positive	Really negative
Positive predicted	TP	FP
Negative predicted	FN	TN

Precision – as the name suggests – describes how accurate the system is in identifying true positives. If the system has high precision, it means that it is rarely wrong in identifying class of spectrogram: when the system claims the spectrogram is about email information, that is it.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall indicates the ratio of positive instances that are identified by the system. If the system has high recall about $credit\_card$ messages, it means that almost all instances about this type of communication have been identified.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-Score combines precision and recall into a single metric. This metric is the harmonic mean between the two.

$$\text{F1} - \text{Score} = \frac{\text{TP}}{\text{TP} + \frac{\text{FN} + \text{FP}}{2}}$$

Accuracy indicates how close a predicted value is to the actual one: informally, it is the fraction of predictions that are accurate.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Specificity measures the proportion of true negatives and indicates the proportion of truly negative instances that are correctly identified. High specificity means that the model is correctly identifying most negative instances.

$$Specificity = \frac{TN}{FP + TN}$$

Using the elbow method, it was possible to identify the optimal value of K in the KNN algorithm. A similar method was applied to understand the maximum depth value (MD) of the decision trees constructed for the Random Forest model.

The best performances were achieved considering the application of an ensemble learning technique called Stacking (STC): the idea is to build a meta-classifier that learns from the classifications of the individual classifiers using a personal weight matrix. The meta classifier uses Logistic Regression [24].

The Convolutional Neural Network was built with a single convolutional layer inside it. The input of size $224 \times 224 \times 3$ first crosses the convolutional layer characterized by 16 feature maps of size 3×3. The new matrix is then computed by the MaxPooling layer characterized by matrices of size 2×2. To minimize overfitting a Dropout layer is then applied with a percentage of 20% - at each passage of the training data some random nodes are chosen, and they don't update their weights both in forward and backpropagation. The last layers are composed by a full Dense layer of 128 units which is converted to one of length 4 by the Softmax. Using this configuration, the following performances were achieved (see Figure 11).

The confusion matrix shows how generally the accuracy is high for all classes and there is no evidence of the imbalance between them. It is important to note the misclassification among $credit\_card$ and $name\_surname$ classes (see Figure 12).

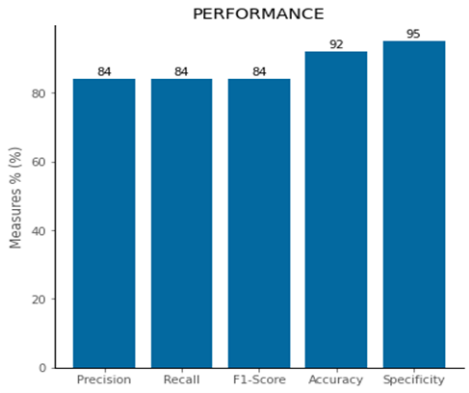


Figure 11: Performance of described CNN

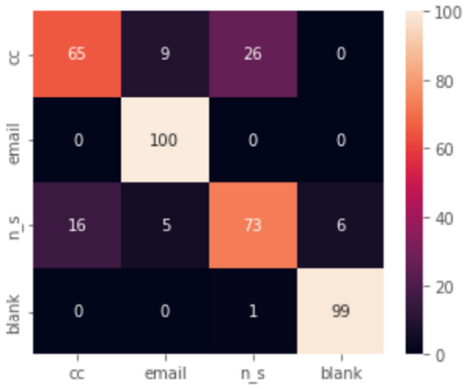


Figure 12: Confusion Matrix of described CNN

Even if the performance of the Neural Network is slightly lower than the application of ensemble techniques, it is very robust to noise. Due to the construction of the Covert Timing Channel and the dataset, the communications are noise-free: we have applied Gaussian noise to the images (see Figure 13).

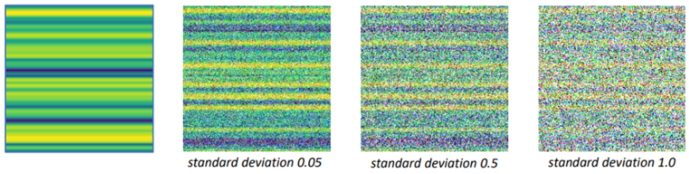


Figure 13: Noise effect with several Standard Deviation

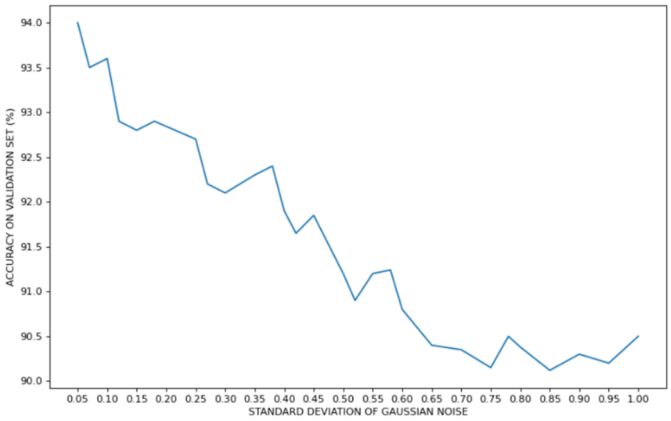


Figure 14: CNN Accuracy, based on Standard Deviation of Gaussian Noise

Gaussian noise is statistical noise having a probability density function equal to that of the normal distribution. We analysed how the performance varies as the standard deviation value of the noise varies (see Figure 14).

By testing the results of the various approaches introduced previously - RF, KNN, STC - the convolutional network is the one that maintains the highest accuracy value even in the case of spectrograms strongly affected by noise. It was therefore decided to use CNN as the first model in the classification pipeline (see Figure 15).

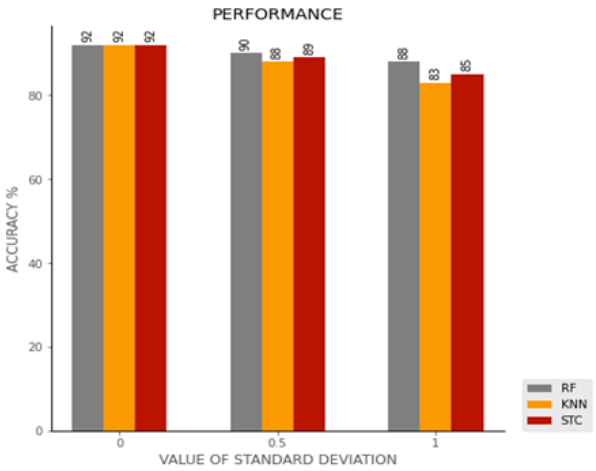


Figure15: Accuracy of ML and DL with several Standard Deviation

To try to improve the performance of Convolutional Neural Network we analysed a Siamese Neural Network. As we can check observing confusion matrix, the network didn't learn the difference between *credit_card* and *name_surname* instances. We trained the network using an input that is constituted by a pair of images. This network consists of two identical subnets that share weights during training. The idea is to train it to understand the level of similarity between two inputs. Internal networks feature a first layer of ReflectionPad2d which modifies the input tensor. Then, there are 3 convolutional layers characterized by a dimension of the convolutional kernel equal to 3×3 and pairs $(\text{input\_size}, \text{output\_size})$ respectively $(1,4)$, $(4,8)$ and $(8,8)$. Relu is used as activation function followed by two layers of batch normalization.

To easily understand how this network works it is sufficient to think in the following way: two images cross the two internal networks simultaneously and these produce a vectorial representation of them. We calculate the vector distance between them and with a sigmoid neuron we return 0 or 1 – 0 if the images are similar, 1 otherwise. There are several loss functions for training and in our case ConstrastiveLoss was used.

Once we obtained a very performing Siamese network to discriminate between two classes, it is used it in the following way [24]: we recovered the most significant spectrograms of each class by applying KMeans, and we built the dissimilarity space with which we trained a Random Forest model. The idea is to apply

the Neural Network and this new model in cascade: when the CNN predicts in output that the instance is a *credit_card* or *name_surname*, then the RF and the Siamese Neural Network are asked for confirmation. Using this trick, accuracy improved by two percentage points (see Figure 16).

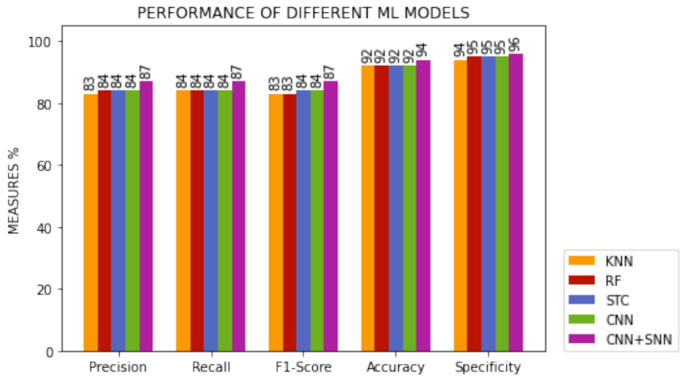


Figure 16: Performance of described ML and DL Models

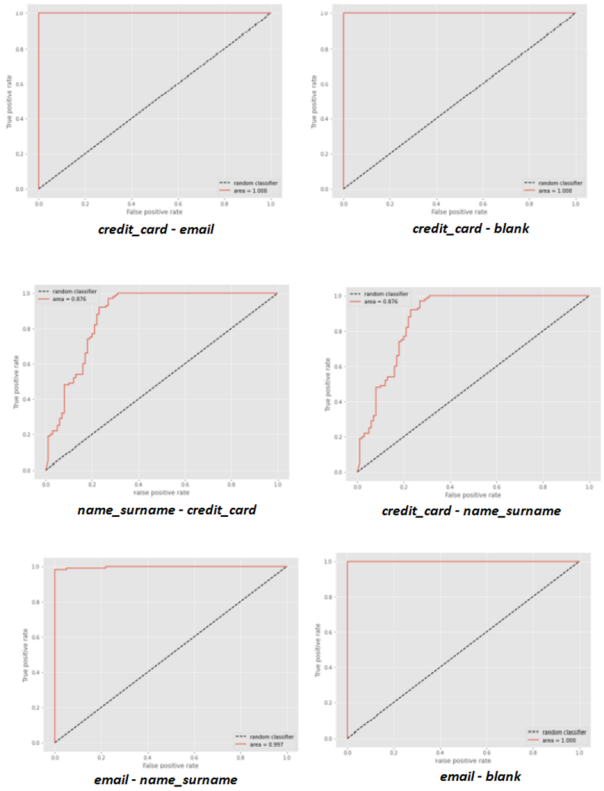


Figure 17: Binary ROCs of proposed pipeline

The value of the AUC - Area Under the Curve - was examined in addition to the metrics previously mentioned. The area under the ROC curve (AUC) is a measurement of its size. The trend of True Positives as a function of False Positives is displayed on a ROC curve – Receiver Operating Characteristics – with different threshold values. In the case of a binary classification, a sigmoid CNN generates a real value in the range $[0, 1]$: based on a selected threshold, the network determines how to classify the input instance.

AUC measures the classifier's ability to distinguish between Positive and Negative classes: the higher the AUC, the more effective the model. It is useful to understand how True Positives and False Positives change depending on the chosen threshold. TPR (True Positive Rate) and FPR (False Positive Rate) are the two metrics that are used: FPR is the decrease in Specificity compared to 1 while the TPR is the same as the Recall. Figure 17 shows ROC curves in binary classifications.

5. Conclusion

This paper showed how to create a simple CTC and how it is possible to apply Machine Learning (Random Forest and K-NN) and Deep Learning (Convolutional Neural Network and Siamese Network) approaches to classify hidden communications in TCP-based Covert Timing Channels in the e-health field. We proposed an innovative pipeline composed of a single CNN and a SNN to improve the accuracy of the classification. We have compared the performances of different methods and improved them with ensemble and combination techniques. The best performance was achieved by our pipeline with an accuracy of 94%. Even if the performances presented are slightly lower than those of the State of Art – which use ML and statistical models obtaining performances of 96% [20] - the work paves the way for the use of DL models for the identification of CTCs. The further contribution presented is the noise resistance of the pipeline: if there is noise, modelled with Gaussian distribution and different value of standard deviation applied to spectrograms, the pipeline performance remains efficient with 90% accuracy. The detection of the prototypes of each class demonstrates how it is possible to identify a representative spectrogram of each message in transit in the CTC. The prototype can be thought as a hashing of the attack and can be used in SIEM systems to compare the state of the network against each known hashed attack type.The current model has been tested on simple messages: as the number of classes increases, the size of the dissimilarity space increases, which could lead to longer training and identification times. At the same time, the simulated noise is only fictitious: it was inserted afterwards.

We are satisfied with the performances obtained in general but not so much with those relating to the Siamese network. Our future studies will focus on the study of this network and how to combine it with other DL models. Other studies are focusing on the internal analysis of the packets exchanged and using different CTCs.

References

[1] F. Hu, D. Xie, S. Shen, "On the Application of the Internet of Things in the Field of Medical and Health Care," in 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, IEEE: 2053–2058, 2013, doi:10.1109/GreenCom-iThings-CPSCom.2013.384.

[2] G.J. Simmons, "Symmetric and Asymmetric Encryption," ACM Computing Surveys, **11**(4), 305–330, 1979, doi:10.1145/356789.356793.

[3] S.K. Kharroub, K. Abualsaud, M. Guizani, "Medical IoT: A Comprehensive Survey of Different Encryption and Security Techniques," in 2020 International Wireless Communications and

Mobile Computing (IWCMC), IEEE: 1891–1896, 2020, doi:10.1109/IWCMC48107.2020.9148287.

[4] Y. Sun, F.P.-W. Lo, B. Lo, "Lightweight Internet of Things Device Authentication, Encryption, and Key Distribution Using End-to-End Neural Cryptosystems," IEEE Internet of Things Journal, **9**(16), 14978–14987, 2022, doi:10.1109/JIOT.2021.3067036.

[5] S.S. Albouq, A.A.A. Sen, A. Namoun, N.M. Bahbouh, A.B. Alkhodre, A. Alshanqiti, "A Double Obfuscation Approach for Protecting the Privacy of IoT Location Based Applications," IEEE Access, **8**, 129415–129431, 2020, doi:10.1109/ACCESS.2020.3009200.

[6] L. SWEENEY, "k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, **10**(05), 557–570, 2002, doi:10.1142/S0218488502001648.

[7] A. Sundas, S. Badotra, S. Bharany, A. Almogren, E.M. Tag-ElDin, A.U. Rehman, "HealthGuard: An Intelligent Healthcare System Security Framework Based on Machine Learning," Sustainability, **14**(19), 11934, 2022, doi:10.3390/su141911934.

[8] J. Asharf, N. Moustafa, H. Khurshid, E. Debie, W. Haider, A. Wahab, "A Review of Intrusion Detection Systems Using Machine and Deep Learning in Internet of Things: Challenges, Solutions and Future Directions," Electronics, **9**(7), 1177, 2020, doi:10.3390/electronics9071177.

[9] C. Janiesch, P. Zschech, K. Heinrich, "Machine learning and deep learning," Electronic Markets, **31**(3), 685–695, 2021, doi:10.1007/s12525-021-00475-2.

[10] Y. Rbah, M. Mahfoudi, Y. Balboul, M. Fattah, S. Mazer, M. Elbekkali, B. Bernoussi, "Machine Learning and Deep Learning Methods for Intrusion Detection Systems in IoMT: A survey," in 2022 2nd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), IEEE: 1–9, 2022, doi:10.1109/IRASET52964.2022.9738218.

[11] H. Rathore, L. Wenzel, A.K. Al-Ali, A. Mohamed, X. Du, M. Guizani, "Multi-Layer Perceptron Model on Chip for Secure Diabetic Treatment," IEEE Access, **6**, 44718–44730, 2018, doi:10.1109/ACCESS.2018.2854822.

[12] S.P. R.M., P.K.R. Maddikunta, P. M., S. Koppu, T.R. Gadekallu, C.L. Chowdhary, M. Alazab, "An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture," Computer Communications, **160**, 139–149, 2020, doi:10.1016/j.comcom.2020.05.048.

[13] S. Khan, A. Akhunzada, "A hybrid DL-driven intelligent SDN-enabled malware detection framework for Internet of Medical Things (IoMT)," Computer Communications, **170**, 209–216, 2021, doi:10.1016/j.comcom.2021.01.013.

[14] A. Djenna, D. Eddine Saidouni, "Cyber Attacks Classification in IoT-Based-Healthcare Infrastructure," in 2018 2nd Cyber Security in Networking Conference (CSNet), IEEE: 1–4, 2018, doi:10.1109/CSNET.2018.8602974.

[15] H. Okhravi, S. Bak, S.T. King, "Design, implementation and evaluation of covert channel attacks," in 2010 IEEE International Conference on Technologies for Homeland Security (HST), IEEE: 481–487, 2010, doi:10.1109/THS.2010.5654967.

[16] A. Chen, W.B. Moore, H. Xiao, A. Haeberlen, L. Thi Xuan Phan, M. Sherr, W.Z. Zhou, Detecting Covert Timing Channels with Time-Deterministic Replay, USENIX Association, 2014.

[17] F. Chen, Y. Wang, H. Song, X. Li, "A statistical study of covert timing channels using network packet frequency," in 2015 IEEE International Conference on Intelligence and Security Informatics (ISI), IEEE: 166–168, 2015, doi:10.1109/ISI.2015.7165963.

[18] C. Shepherd, J. Kalbantner, B. Semal, K. Markantonakis, "A Side-channel Analysis of Sensor Multiplexing for Covert Channels and Application Fingerprinting on Mobile Devices," 2021.

[19] S. Gianvecchio, Haining Wang, "An Entropy-Based Approach to Detecting Covert Timing Channels," IEEE Transactions on Dependable and Secure Computing, **8**(6), 785–797, 2011, doi:10.1109/TDSC.2010.46.

[20] M.A. Elsadig, A. Gafar, "Covert Channel Detection: Machine Learning Approaches," IEEE Access, **10**, 38391–38405, 2022, doi:10.1109/ACCESS.2022.3164392.

[21] F. Massimi, F. Benedetto, "Deep Learning-based Detection Methods for Covert Communications in E- Health Transmissions," in 2022 45th International Conference on Telecommunications and Signal Processing (TSP), IEEE: 11–16, 2022, doi:10.1109/TSP55681.2022.9851366.

[22] S. Al-Eidi, O. Darwish, Y. Chen, G. Husari, "SnapCatch: Automatic Detection of Covert Timing Channels Using Image Processing and Machine Learning," IEEE Access, **9**, 177–191, 2021, doi:10.1109/ACCESS.2020.3046234.

[23] J. BROMLEY, J.W. BENTZ, L. BOTTOU, I. GUYON, Y. LECUN, C. MOORE, E. SÄCKINGER, R. SHAH, "SIGNATURE VERIFICATION USING A 'SIAMESE' TIME DELAY NEURAL NETWORK," International Journal of Pattern Recognition and Artificial Intelligence, **07**(04), 669–688, 1993, doi:10.1142/S0218001493000339.

[24] T.N. Rincy, R. Gupta, "Ensemble Learning Techniques and its Efficiency in Machine Learning: A Survey," in 2nd International Conference on Data, Engineering and Applications (IDEA), IEEE: 1–6, 2020, doi:10.1109/IDEA49133.2020.9170675.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 1, 131-137 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Active Simulation of Grounded Parallel-Type Immittance Functions Employing VDBAs and All Grounded Passive Components

Pratya Mongkolwai[1], Pitchayanin Moonmuang[2], Worapong Tangsrirat[2,*], Taweepol Suesut[2]

[1]*Department of Instrumentation Engineering, Faculty of Engineering, Rajamangala University of Technology Rattanakosin, Nakhon Pathom 73170, Thailand*

[2]*Department of Instrumentation and Control Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand*

ARTICLE INFO

Article history:
Received: 08 October, 2022
Accepted: 15 January, 2023
Online: 24 February, 2023

Keywords:
Voltage Differencing Buffered Amplifier (VDBA)
Immittance function
Impedance simulator

ABSTRACT

This communication proposes a grounded immittance function simulator that, depending on the proper choice of the passive components, can simulate parallel-type impedances of the R-L, R-C, and L-C forms. Only two grounded passive components and two voltage differencing buffered amplifiers (VDBAs) are used to implement the suggested circuit. All three simulated equivalent elements, namely R_{eq}, L_{eq}, and C_{eq}, can be electronically adjusted through the VDBA's transconductance gain. The impact of the non-ideality of the VDBA device on the developed simulator is examined in detail. The voltage-mode bandpass filter has been implemented using the suggested active LC parallel impedance simulator to show that it performs as predicted. To prove the theory, the proposed circuit is simulated using the PSPICE tool. The findings of the experimental measures are also presented to demonstrate the circuit's feasibility.

1. Introduction

Electronic devices have assimilated into our daily lives in the world today. The development of novel technologies will be influenced by the published findings. In several analog signal processing solutions, the different active devices, such as current conveyor (CC), operational transconductance amplifier (OTA), current feedback operational amplifier (CFOA), and current differencing buffered amplifier (CDBA), have gained widespread attention. Similarly, since 2008, the voltage differencing buffered amplifier (VDBA) has been recognized as one of the most versatile and practical devices [1]-[2].

The VDBA element has a tunable transconductor as the input section and a voltage buffer as the output section. Because of this feature, this active element can be used in a variety of voltage-mode, current-mode, and mixed-mode analog circuits and applications [2–6]. Passive elements, such as resistors, capacitors, and inductors, were used in a variety of applications, including analog active filter circuits, sinusoidal oscillator design, and impedance cancellation circuit. However, when applied in the implementation of an integrated circuit (IC), the behavior of passive elements was constrained by its enormous size and suffered from electronic tuning properties. As a consequence, an IC that mimicked the behavior of a passive element was implemented using an active element [7-9]. The parallel-type R-L simulators that were suggested in the literature [10–12] needed at least three active components. Similar to that, three or more passive components are required to realize the circuits in [11–12]. The circuits in [13] also need a high-voltage operation.

Therefore, the contribution of this work is to propose a grounded parallel R-L, R-C, and L-C impedance simulator, which depends on the appropriate selection of the passive element being used. The suggested simulator circuit uses only two VDBAs, two grounded passive components, and allows electronically control of the equivalent simulated elements via the transconductance gains of the VDBAs. In this study, the VDBA non-ideality effect on the actual immittance simulator is examined. With 0.18-μm CMOS technology, the proposed R-L, R-C, and L-C impedance simulator circuit in frequency domain was simulated using PSPICE program. Time-domain analysis and temperature-dependent simulation are also carried out in the parallel R-C simulator. The theoretical analysis is validated by experimental laboratory measurements using commercially available IC LT1228. Additionally, the active L-C simulator has also been used to apply a second-order voltage-

*Corresponding Author: Worapong Tangsrirat, Email: worapong.ta@kmitl.ac.th

mode bandpass filter in order to validate the viability. All results, both from simulations and experiments, are discovered to be in accordance with the theoretical predictions.

2. Fundamental of VDBA

Figure 1 depicts the electrical symbol of the VDBA element. This functional block has two input terminals (p and n) that meet high input impedance criteria and two output terminals (z and w), which have high and low impedances, respectively. Under ideal operating condition, the effective transconductance gain (g_m) of the VDBA converts the differential voltage between v_p and v_n (v_p - v_n) into an output current (i_z) at terminal z. The voltage drop (v_z) at the z terminal is transferred to the output voltage (v_w) at the w terminal. From its ideal operating condition, the following matrix equation can be used to characterize the terminal relationship of the VDBA [1-2]:

$$\begin{bmatrix} i_p \\ i_n \\ i_z \\ v_w \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ g_m & -g_m & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} v_p \\ v_n \\ v_z \\ i_w \end{bmatrix} \quad (1)$$

In general, the g_m value in (1) can be changed by electronic means via the external bias voltage or current.

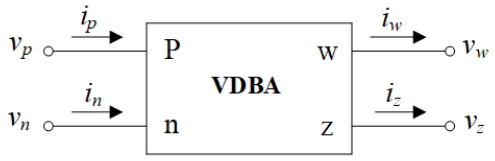


Figure 1: Schematic symbol of the VDBA.

In non-ideal assumption, the characteristic of VDBA can be modified as [3]:

$$\begin{bmatrix} i_p \\ i_n \\ i_z \\ v_w \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \alpha g_m & -\alpha g_m & 0 & 0 \\ 0 & 0 & \beta & 0 \end{bmatrix} \begin{bmatrix} v_p \\ v_n \\ v_z \\ i_w \end{bmatrix} . \quad (2)$$

In above expression, $\alpha = (1 - \varepsilon_{gm})$ and $\beta = (1 - \varepsilon_v)$, where $|\varepsilon_{gm}| << 1$ and and $|\varepsilon_v| << 1$ stand for transconductance inaccuracy coefficient and the voltage tracking error, respectively.

A CMOS model of VDBA consisting of the differential amplifiers with active load (M_1-M_4 and M_7-M_{10}), and the source follower (M_{11}) is shown in Figure 2. For the CMOS VDBA in Figure 2, the relationship between g_m and the bias current I_B can be characterized as follows [4]:

$$g_m = \sqrt{\mu C_{ox} \left(\frac{W}{L}\right) I_B} . \quad (3)$$

Here, μ is the effective carrier mobility, C_{ox} is the gate-oxide capacitance per unit area, and W and L are the effective channel width and length of M_1 and M_2 transistors, respectively.

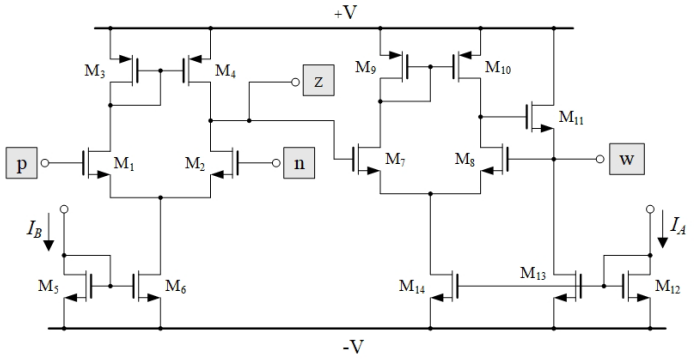


Figure 2: CMOS model of the VDBA used in this work.

3. Proposed Parallel-Type Immittance Function Simulator

According to Figure 3, the suggested grounded parallel-type immittance simulator is made up of two VDBAs and two grounded passive components. Based on ideal condition consumption, the input admittance (Y_{in}) of the circuit is derived as:

$$Y_{in} = \frac{i_{in}}{v_{in}} = g_{m1} g_{m2} Z_A + \frac{1}{Z_B} . \quad (4)$$

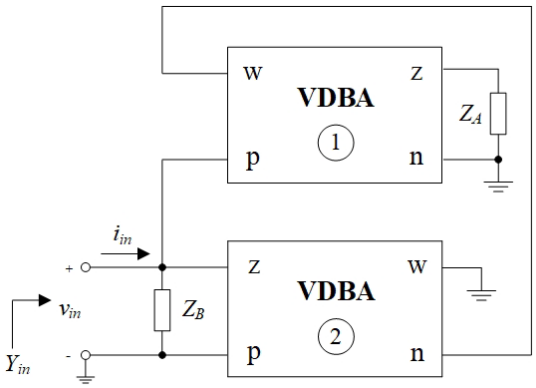


Figure 3: Proposed grounded parallel-type immittance function simulator.

The proposed parallel R-L, R-C, and L-C immittance simulator was made achievable by selecting the appropriate passive components, which describes the realized circuit. Its simulated impedances are summarized in Table 1, which illustrates that all synthetic simulator values can electronically be changed by the transconductance g_{mi} of the i-th VDBA (i = 1, 2). Since all of the passive components are grounded, the configuration is attractive from further integration point of view. Another attractive feature of the design is that it does not need any special component equality for its realization.

Under the non-ideal operation given in (2), the results of reevaluating the proposed circuit in Figure 3 can be summarized in Table 2.

Table 1: Equivalent Circuit and Corresponding Equivalent Values for Figure 3 in Ideal Case

Z_A	Z_B	Equivalent circuit	Equivalent values
$1/sC_A$	R_B	parallel R-L	$R_{eq} = R_B$,

			$L_{eq}=\frac{C_A}{g_{m1}g_{m2}}$
R_A	$1/sC_B$	parallel R-C	$R_{eq}=\frac{1}{R_A g_{m1}g_{m2}}$, $C_{eq}=C_B$
$1/sC_A$	$1/sC_B$	parallel L-C	$L_{eq}=\frac{C_B}{g_{m1}g_{m2}}$, $C_{eq}=C_A$

Table 2: Equivalent Element Values for Figure 3 in Non-Ideal Case

Z_A	Z_B	Equivalent circuit	Equivalent values
$1/sC_A$	R_B	parallel R-L	$R_{eq}=R_B$, $L_{eq}=\frac{C_A}{\alpha_1\alpha_2\beta_1 g_{m1}g_{m2}}$
R_A	$1/sC_B$	parallel R-C	$R_{eq}=\frac{1}{\alpha_1\alpha_2\beta_1 R_A g_{m1}g_{m2}}$, $C_{eq}=C_B$
$1/sC_A$	$1/sC_B$	parallel L-C	$L_{eq}=\frac{C_B}{\alpha_1\alpha_2\beta_1 g_{m1}g_{m2}}$, $C_{eq}=C_A$

The sensitivity coefficients of the simulated equivalent values, R_{eq} L_{eq} and C_{eq}, are each affected by the active and passive circuit components, and the finding values are produced as follows.

For parallel R-L;

$$S_{R_B}^{R_{eq}}=1\ ,\ S_{\alpha_1,\alpha_2,\beta_1,g_{m1},g_{m2}}^{L_{eq}}=-1\ ,\ S_{C_A}^{L_{eq}}=1\ . \qquad (5)$$

For parallel R-C;

$$S_{\alpha_1,\alpha_2,\beta_1,R_A,g_{m1},g_{m2}}^{R_{eq}}=-1\ ,\ S_{C_B}^{C_{eq}}=1\ . \qquad (6)$$

For parallel L-C;

$$S_{\alpha_1,\alpha_2,\beta_1,g_{m1},g_{m2}}^{L_{eq}}=-1\ ,\ S_{C_B}^{L_{eq}}=1\ ,\ S_{C_A}^{C_{eq}}=1\ . \qquad (7)$$

All of the sensitivity coefficients from above (5) to (7) have magnitudes that are less than or equal to one. As a result, the sensitivity of all the proposed parallel R-L, R-C, and L-C immittance simulators is quite low.

4. Simulation Results

PSPICE simulation program has been used to simulate the suggested grounded parallel-type impedance simulator in Figure 3. The simulator was designed employing CMOS VDBA of Figure 2 with a model of 0.18-µm process parameters from TSMC. Table 3 lists the computed aspect ratio (W/L) for each transistor. The supply voltages used to bias this circuit were +V= -V= 0.75 V.

Table 3: Calculated transistor dimensions of VDBA in Figure 2

Transistor	W/L (µm/µm)
M_1-M_2, M_5, M_7-M_8, M_{12}-M_{13}	2.4/0.18
M_3, M_9, M_{14}	5/0.18
M_4, M_{10}	5.2/0.18
M_6	3.25/0.18
M_{11}	10/0.18

The following components were chosen for simulations: I_{B1} = I_{B2} = 90 μA for $g_m = g_{m1} = g_{m2}$ = 0.641 mA/V, $R_A = R_B$ = 1 kΩ, and $C_A = C_B$ = 50 pF. Using data from Table 1, the simulated equivalent values of Figure 3 can be derived as:

- for R-L simulator: R_{eq} = 1 kΩ and L_{eq} = 0.12 mH;
- for R-C simulator: R_{eq} = 2.44 kΩ and C_{eq} = 50 pF;
- for L-C simulator: L_{eq} = 0.12 mH and C_{eq} = 50 pF.

The total power consumed in the circuit for this setting was found to be 0.388 mW.

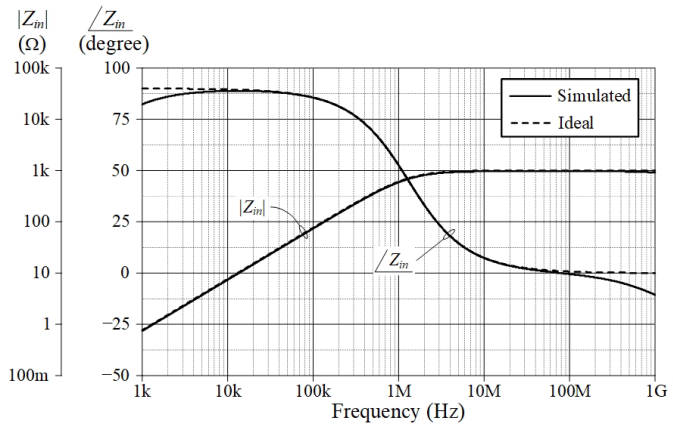
Based on the results of the simulation and theory, Figure 4 depicts the magnitude and phase frequency characteristics of the proposed parallel-type immittance simulator circuit in Figure 3. The frequency corners (f_c) of the R-L and R-C impedance simulators in Figs. 4(a) and 4(b) obtained from the simulation results are found to be roughly 1.29 MHz, which is pretty close to the calculated value of 1.30 MHz. In addition, the simulated f_c value of the L-C impedance simulator was discovered to be 2.04 MHz, which nearly equals to the ideal value of f_c = 2.05 MHz. The input voltage (v_{in}) and current (i_{in}) responses through the R-C impedance simulator are also displayed in Figure 5 as simulated time-domain waveforms. This performance was evaluated by supplying a sinusoidal input signal with a peak value of 50 mV at f = 1 MHz to the simulated RC impedance circuit.

In order to further illustrate the electronic adjustability of the proposed circuit, the parallel L-C simulator has been performed to change $I_B = I_{B1} = I_{B2}$ = 50 μA, 100 μA, and 200 μA, while maintaining $C_A = C_B$ = 50 pF. As a consequence, the simulated equivalent inductance value (L_{eq}) has been altered to 0.22 mH, 0.11 mH, and 54.8 μH, respectively, while the simulated equivalent capacitance value (C_{eq}) remains constant at 50 pF. The results of the simulated frequency responses compared with the theory are given in Figure 6.

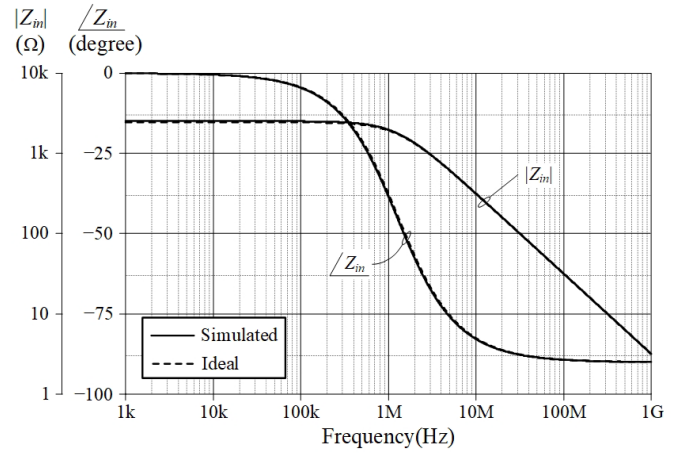
The impact of varying ambient temperature variation on the simulator responses is also being considered. This was accomplished by testing the proposed R-C simulator circuit with changes in ambient temperature ranging from 0˚C to 100˚C with steps of 25˚C. Figure 7 displays the result of its magnitude variations.

5. Experimental Results

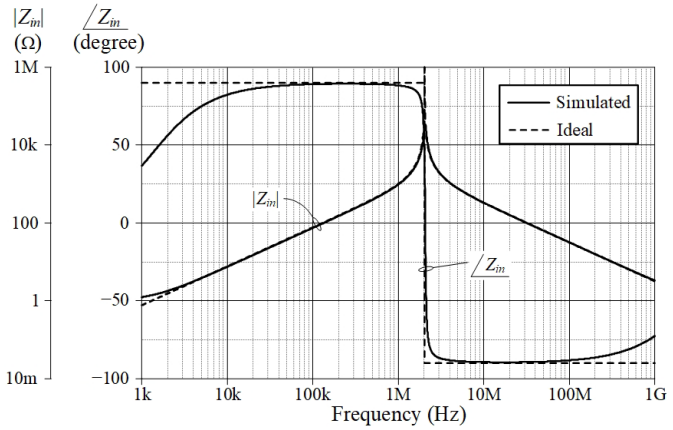
In order to further confirm the feasibility of the proposed idea, the suggested circuit of Figure 3 has been tested in the laboratory utilizing IC LT1228 from Linear Technology [14]. The package information and internal behavior of IC LT1228 are shown in Figure 8. There are two amplifiers: OTA and CFOA. The OTA is used to provide a high-impedance differential input and a current source output with wide output voltage compliance, while the CFOA is utilized to transmit voltage from the z terminal to the o terminal, and the current from the z terminal to the x terminal. According to the following relation, the transconductance gain (g_m) of the LT1228 in this case is reliant on the external bias current (I_B) [14]:



(a)



(b)



(c)

Figure 4: Ideal and simulated frequency responses of the proposed parallel-type immittance simulator circuit in Figure 3. (a) R-L impedance simulator (b) R-C impedance simulator (c) L-C impedance simulator

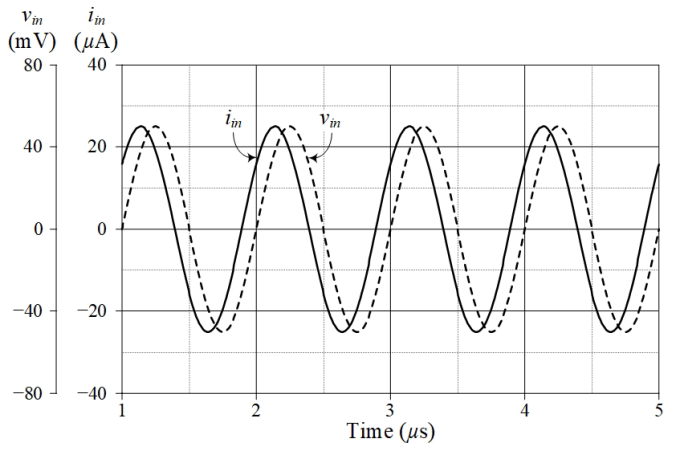


Figure 5: Simulated time-domain responses for v_{in} and i_{in} of the R-C simulator.

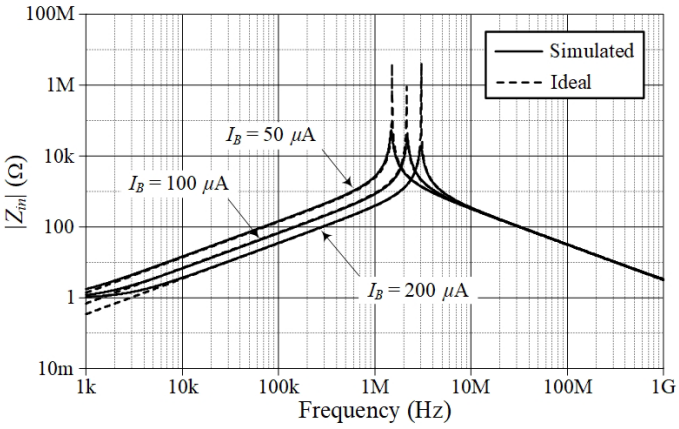


Figure 6: Simulated frequency responses of the L-C simulator with varying I_B.

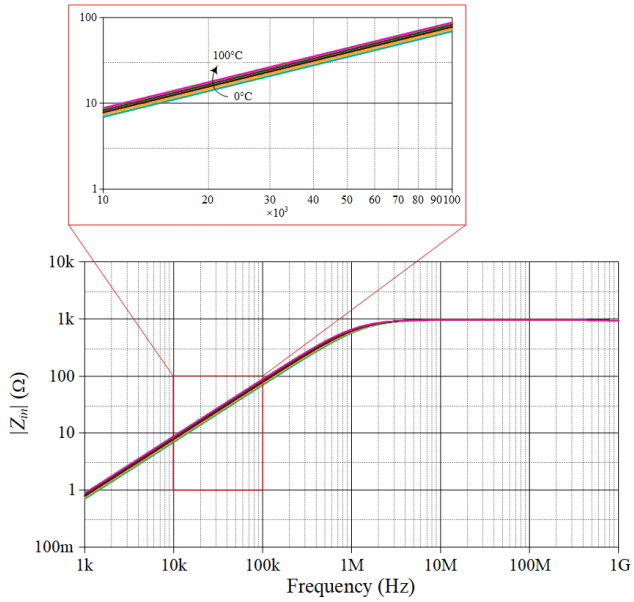
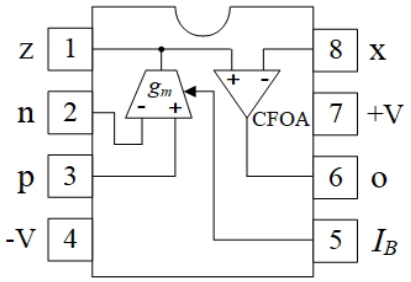


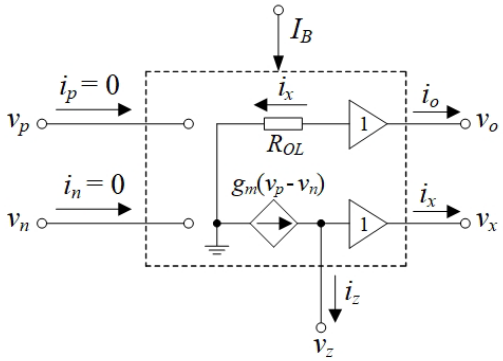
Figure 7: Simulated frequency responses of the R-C simulator at different temperature (0˚C, 25˚C, 50˚C, 75˚C, and 100˚C).

$$g_m = 10I_B \,. \tag{8}$$

In the case of parallel R-L impedance simulation, the active and passive components for the experimental measurement were taken as follows: $g_m = g_{m1} = g_{m2} = 0.5$ mA/V ($I_B = I_{B1} = I_{B2} = 50$ μA), $R_B = 1$ kΩ, and $C_A = 1$ nF, resulting in $R_{eq} = 1$ kΩ, and $L_{eq} = 4$ mH. With symmetrical supply voltages of ±5 V, the LT1228 was biased. Figure 9 shows the grounded parallel lossy inductor's measured magnitude and phase responses for the selected components.

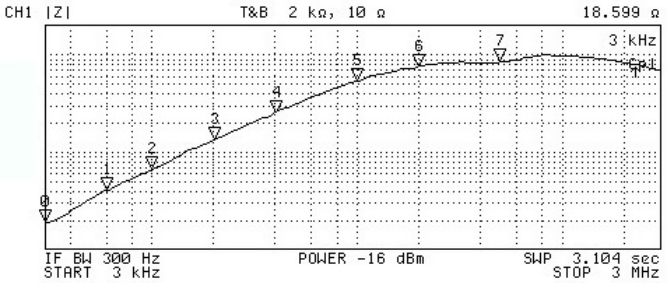


(a)



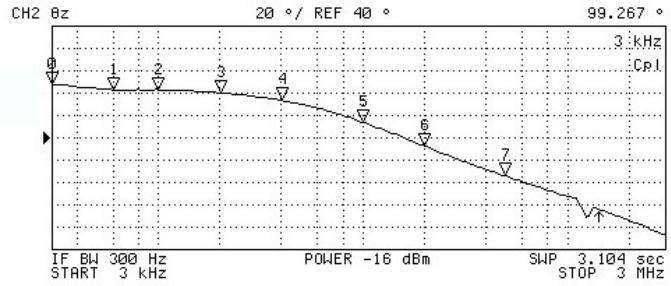
(b)

Figure 8: IC LT1228 (a) package information (b) its internal behavior.



N	SWP PARAM	VAL
0	3 kHz	18.599 Ω
1	6 kHz	40.525 Ω
2	10 kHz	65.485 Ω
3	20 kHz	132.23 Ω
4	40 kHz	251.76 Ω
5	100 kHz	536.14 Ω
6	200 kHz	751.61 Ω
7	500 kHz	838.94 Ω

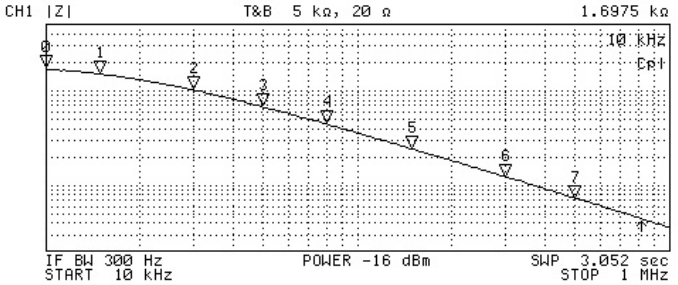
(a)



N	SWP PARAM	VAL
0	3 kHz	94.267 °
1	6 kHz	82.793 °
2	10 kHz	83.121 °
3	20 kHz	80.294 °
4	40 kHz	73.487 °
5	100 kHz	53.942 °
6	200 kHz	32.572 °
7	500 kHz	5.8288 °

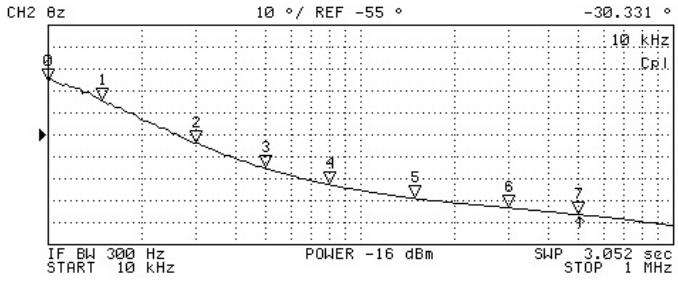
(b)

Figure 9: Measured frequency responses of the parallel R-L simulator (a) magnitude response (b) phase response



N	SWP PARAM	VAL
0	10 kHz	1.6975 kΩ
1	15 kHz	1.4837 kΩ
2	30 kHz	1.0093 kΩ
3	50 kHz	670.32 Ω
4	80 kHz	439.41 Ω
5	150 kHz	240.99 Ω
6	300 kHz	122.06 Ω
7	500 kHz	72.889 Ω

(a)



N	SWP PARAM	VAL
0	10 kHz	-30.331 °
1	15 kHz	-39.565 °
2	30 kHz	-58.712 °
3	50 kHz	-70.298 °
4	80 kHz	-77.599 °
5	150 kHz	-83.904 °
6	300 kHz	-88.317 °
7	500 kHz	-91.252 °

(b)

Figure 10: Measured frequency responses of the parallel R-C simulator (a) magnitude response (b) phase response

The parallel R-C simulator was then tested with the following parameters: $g_m = g_{m1} = g_{m2} = 1$ mA/V ($I_B = I_{B1} = I_{B2} = 100$ µA), $R_A = 500$ Ω, and $C_B = 4.7$ nF, yielding $R_{eq} = 2$ kΩ, and $C_{eq} = 4.7$ nF. Figure 10 shows the measured frequency responses for the equivalent input impedance of the simulator. The experimental results shown in Figures 9 and 10 demonstrate the suggested circuit's practicality in application areas.

6. Application Example

The second-order voltage-mode bandpass filter in Figure 11(a) is intended to emphasize operational performance as an illustration of an application. The bandpass filter realization utilizing the proposed L-C parallel immittance simulator in Figure 3 is shown in Figure 11(b). The voltage transfer action of the filter is written as:

$$\frac{V_{out}(s)}{V_{in}(s)} = \frac{s\left(\frac{1}{RC_{eq}}\right)}{s^2 + s\left(\frac{1}{RC_{eq}}\right) + \frac{1}{L_{eq}C_{eq}}} \quad . \tag{9}$$

The natural angular frequency (ω_o) and the quality factor (Q) of the filter in Figure 11 are determined from (9), respectively, by:

$$\omega_o = 2\pi f_o = \sqrt{\frac{1}{L_{eq}C_{eq}}} \quad , \tag{10}$$

and

$$Q = R\sqrt{\frac{C_{eq}}{L_{eq}}} \quad . \tag{11}$$

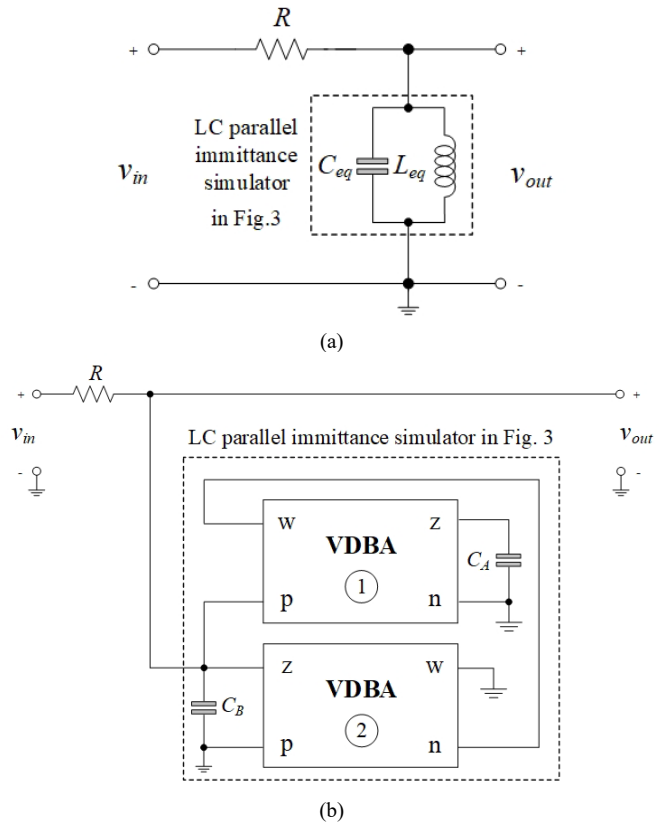


Figure 11: Second-order voltage-mode bandpass filter (a) prototype passive structure (b) utilizing the L-C simulator in Figure 3.

The simulated frequency response of the implemented active bandpass filter is demonstrated in Figure 12 with the following components: $R = 1.5$ kΩ, $g_m = g_{m1} = g_{m2} = 0.675$ mA/V ($I_B = I_{B1} = I_{B2} = 100$ μA), and $C_A = C_B = 100$ pF. With $L_{eq} = 0.22$ mH and $C_{eq} = 100$ pF, the filter is designed to obtain $f_o = \omega_o/2\pi = 1.07$ MHz and $Q = 1$. The resulting responses demonstrate that the circuit can operate correctly between 500 kHz and 100 MHz.

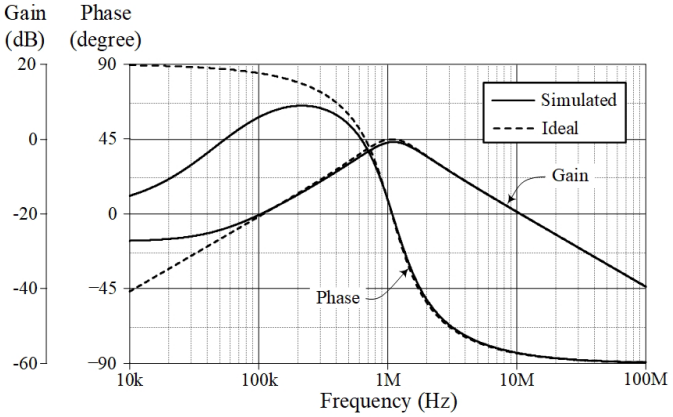


Figure 12: Ideal and simulated results of the bandpass frequency responses in Figure 11(b).

7. Conclusions

A grounded parallel RL, RC, and LC impedance simulator has been designed with VDBAs and two grounded passive components. Through the use of the transconductance parameter in VDBA, the simulated equivalent values, i.e., R_{eq}, L_{eq}, and C_{eq}, can all be electronically altered. The circuit has been simulated using the PSPICE program, which is based on 0.18-µm CMOS technology, to demonstrate its viability. In-depth laboratory tests have also been conducted to verify the practical usability of the simulator circuit. The design of a second-order voltage-mode bandpass filter using the proposed simulator is given.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by Rajamangala University of Technology Rattanakosin (RMUTR). The support by the School of Engineering, King Mongkut's Institute of Technology Ladkrabang, contact no. 2562-02-01-004, is also gratefully acknowledged.

References

[1] D. Biolek, R. Senani, V. Biolkova and Z. Kolka, "Active elements for analog signal processing: classification, review, and new proposals", Radioengineering, **17**(4), 15–32, 2008.

[2] R. Sotner, J. Jerabek, N. Herencsar, N. "Voltage differencing buffered/inverted amplifiers and their applications for signal generation", Radioengineering, **22**(2), 490-504, 2013.

[3] W. Tangsrirat, "Actively Floating lossy inductance simulators using voltage differencing buffered amplifiers," IETE Journal of Research, **65**(4), 446–459, 2018, doi.org/10.1080/03772063.2018.1433082

[4] F. Kaçar, A. Yeşil, A. Noori, "New CMOS realization of voltage differencing buffered amplifier and its biquad filter applications," Radioengineering, **21**(1), 333–339, 2012.

[5] N. Roongmuanpha, T. Pukkalanun, W. Tangsrirat, "Practical realization of electronically adjustable universal filter using commercially available IC-based VDBA," Engineering Review, 41(3), 1-14, 2021, doi.org/10.30765/re.1547.
[6] M. Faseehuddin, N. Herencsar, S. Shireen, W. Tangsrirat, S. H. M. Ali, "Voltage differencing buffered amplifier-based novel truly mixed-Mode biquadratic universal filter with versatile input/output features", Applied Sciences, **12**(3), 2022, doi.org/10.3390/app12031229.
[7] P. Moonmuang, T. Pukkalanun, W. Tangsrirat, "Floating/grounded series/parallel R-L, R-C and L-C immittance simulators employing VDTAs and only two grounded passive elements," AEU - International Journal of Electronics and Communications, **145**, 154095, 2022, doi.org/10.1016/j.aeue.2021.154095.
[8] P. Moonmuang, W. Tangsrirat, "Single VDTA-based tunable floating lossy inductance simulation circuits," IETE Journal of Research, doi: 10.1080/03772063.2021.1900752.
[9] N. Roongmuanpha, W. Tangsrirat, "Practical floating capacitance multiplier implementation with LT1228s," Informacije MIDEM- Journal of Microelectronics, Electronic Components and Materials, **51**(1), 85-94, 2021, doi.org/10.33180/InfMIDEM2021.106
[10] A. Paul and D. Patranabis, "Active simulation of grounded inductors using a single current conveyor", IEEE Transactions on Circuits and Systems, **28**(2), 164-165, 1981, doi.org/10.1109/TCS.1981.1084947
[11] H. Kuntman, M. Gülsoy and O. Çiçekoğlu, "Actively simulated grounded lossy inductors using third generation current conveyors", Microelectronics Journal, **31**(4), 245-250, 2000, doi.org/10.1016/S0026-2692(99)00108-1
[12] O. Çiçekoğlu, A. Toker and H. Kuntman, "Universal immittance function simulators using current conveyors", Computers and Electrical Engineering, **27**(3), 227-238, 2001, doi.org/10.1016/S0045-7906(00)00018-5
[13] F. Kaçar and H. Kuntman, "CFOA-based lossless and lossy inductance simulators", Radioengineering, **20**(3), 627-631, 2011.
[14] Linear Technology, "100MHz current feedback amplifier with DC gain control", LT1228 datasheet, 1994.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 136-143 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Tunable Resistorless Phase Shifter Realization with a Single VDGA

Orapin Channumsin[1], Jirapun Pimpol[1], Tattaya Pukkalanun[2] , Worapong Tangsrirat[2,*]

[1]*Faculty of Engineering, Rajamangala University of Technology Isan, Khonkaen Campus, Khonkaen 40000, Thailand*

[2]*School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand*

ARTICLE INFO

Article history:
Received: 30 January, 2023
Accepted: 12 May, 2023
Online: 12 June, 2023

Keywords:
Voltage Differencing Gain Amplifier (VDGA)
Phase shifter
Electronically tunable
All-pass filter
Voltage-mode circuit

ABSTRACT

This paper describes the design of a phase shifter with electrically adjustable parameters employing only one voltage differencing gain amplifier (VDGA) and one floating capacitor. This circuit requires no external resistors, resulting in a resistorless design and a low component count. The proposed circuit implements a first-order all-pass filter response with electronic control of its passband gain, pole frequency, and phase difference via bias current modification. Non-ideal effects of the VDGA on the phase shifter circuit are also examined. PSPICE simulation results using TSMC 0.25-μm real process parameters and practical test results using readily available LM13700s are incorporated to validate the theoretical conclusions. The results indicate that the simulations and experiments yielded phase shift deviations of 2.22% and 3.11%, respectively. The pole-frequency errors for simulations and experiments were 0.31% and 0.63%, respectively. The applicability of the suggested phase shifter is illustrated by the design of the voltage-mode quadrature oscillator.

1. Introduction

The design and synthesis of the phase shifter circuit, also known as a first-order all-pass filter, has received a great deal of interest [1]. In general, the phase behavior of the phase shifter circuit can be adjusted from 0° to 180° or from 180° to 0°, while its amplitude remains unchanged over the entire frequency range of interest. Due to this, the phase shifter circuit is employed in a number of communication and instrumentation systems, such as universal biquad filters, high quality factor frequency-selective filters, and quadrature and multiphase oscillators [2]-[20]. However, voltage-mode phase shifter circuits with one or more active components are the most often suggested circuits in [2]-[3], [5]-[16], [18]-[20]. Additionally, many of the works in [2]-[3], [5], [10]-[16], [18],[20] are inaccessible electronically. Moreover, all of these configurations are realized with the use of external passive resistors.

The main objective of this contribution is, therefore, to design a simple and compact phase shifter circuit with only one active and one passive component. Without an extra passive resistor, the suggested structure consists merely of one voltage difference gain amplifier (VDGA) and one floating capacitor. The benefits of the design include the facility, low power consumption, and small integrated chip area. Furthermore, the important features of the proposed phase shifter, including passband gain (H_0), pole frequency (ω_p), and phase response (ϕ), are electronically tunable through the transconductance gains of the VDGA. A thorough investigation is also done into the non-ideal gain effects of the VDGA on the circuit performance. In addition, a new voltage-mode quadrature oscillator is proposed to highlight the advantages of the designed phase shifter. The designed circuit and its application are simulated using PSPICE software using TSMC 0.25-μm CMOS process technology, and the simulation results are consistent with the theoretical analysis. The experimental measurement results from the laboratory breadboard using commercially available LM13600s are also given to prove the features of the proposed circuit. In addition, a summary of the performance comparison of the proposed circuit and those that the previous works [2]-[20] is provided in Table 1. The observations show that the suggested circuit has more features than recently published circuits, which is commendable.

*Corresponding Author: Worapong Tanasrirat, Email: worapong.ta@kmitl.ac.th

https://dx.doi.org/10.25046/aj080315

Table 1: Comparative study of the proposed circuit with the similar previous works.

Ref	Year	No. of active components	No. of passive components	Resistorless structure	Variable-gain control	Electronic tunability	Power dissipation (W)	Pole frequency (Hz)	Supply voltages (V)	Technology
[2]	2005	CCII+ = 2	R = 2, C = 2	no	yes	no	NA	15.9 k	±12 (experiment)	AD844
[3]	2006	DDCC = 1	R = 1, C = 1	no	no	no	NA	265.4 k	±3.3 (simulate)	1.2 µm
[4]	2017	MMCC = 1, CFA = 1	R = 1, C = 1	yes	no	yes	NA	9.91 M	NA	AD835, AD844
[5]	2000	CDBA = 1	R = 1, C = 1	no	no	no	NA	1.59 M	±2.5 (simulate)	0.8 µm
[6]	2001	CCCII+ = 1	R = 1, C = 1	yes	no	yes	NA	10 M	±2.5 (simulate)	0.35 µm
[7]	2015	VDGA = 1	R = 1, C = 1	yes	yes	yes	1.45 m	429 k	±1.5 (simulate)	0.35 µm
[8]	2017	VDBA = 1	R = 1, C = 1	yes	no	yes	0.37 m	1.06 M	±0.75 (simulate)	0.25 µm
[9]	2019	LT1228 = 1	R = 2, C = 1	yes	no	yes	NA	100 k	±5 (simulate)	LT1228
[10]	2010	FDCCII = 1	R = 2, C = 1	no	no	no	NA	1.59 M	±3.3 (simulate)	0.35 µm
[11]	2012	DDCC = 2	R = 1, C = 1	no	no	no	NA	15.91 M	±2.5 (simulate)	0.5 µm
[12]	2011	FDCCII = 1	R = 1, C = 1	no	no	no	NA	2.65 M	±1.3 (simulate)	0.35 µm
[13]	2012	DDCC = 2	R = 1, C = 1	no	no	no	NA	1.17 M	±2.5 (simulate)	0.5 µm
[14]	2019	Fig.2: CFOA = 2 Fig.3: CFOA = 3	Fig.2: R = 5, C = 1 Fig.3: R = 6, C = 1	no	yes	no	Fig.2: 0.26, Fig.3: 0.39	7.59 k	±10 (experiment)	AD844
[15]	2020	VCII+ = 2	R = 3, C = 1	no	yes (TM)	no	1.22 m	636.6 k	±0.9 (simulate)	0.18 µm
[16]	2020	EXCCII = 1	R = 2, C = 2	no	no	no	0.7 m	3.18 M	±1.2 (simulate)	0.25 µm
[17]	2021	ICCII+ = 2	Fig.2: R = 1 , C = 1 Fig.3: C = 1	Fig.2: no, Fig.3: yes (active resistor)	no	Fig.2: no, Fig.3: yes	3.29 m (simulate)	7.96 M (simulate), 159 k (experiment)	±0.75 (simulate), ±9 (experiment)	0.13 µm, AD844
[18]	2021	CFOA = 2	Fig.1: R = 3, C = 1 Fig.2: R = 4, C = 1	no	yes	no	NA	33.829 k	±8 (simulate/ experiment)	AD844
[19]	2021	FDCCII = 1	C = 1	yes (active resistor)	no	yes	2 m	6.37 M	±1.25 (simulate)	0.25 µm
[20]	2022	DVCC = 2	Fig.1 : R = 1, C = 1 Fig.2 : R = 3, C = 1	no	Fig.1: no Fig.2: yes	no	0.6	62.41 k	±5 (simulate)	AD844
This work	2022	VDGA = 1	C = 1	yes	yes	yes	1.56 m (simulate)	1.59 M (simulate), 159 k (experiment)	±0.75 (simulate), ±5 (experiment)	0.25 µm, LM13600

Abbreviation:
NA = Not Available
CCII = second-generation current conveyor, DDCC = differential difference current conveyor,
MMCC = Multiplication Mode Current Conveyor, CFA = current feedback operational amplifier,
CDBA = current differencing buffered amplifier, CCCII+ = plus-type current-controlled current conveyor,
VDBA = voltage differencing buffered amplifier, FDCCII = fully differential current conveyor,
CFOA = current feedback operational amplifier, VCII+ = plus-type second-generation voltage conveyor,
EXCCII = extra-X second generation current conveyor, ICCII+ = plus-type second-generation current conveyor,
DVCC = differential voltage current conveyor, TM = transimpedance-mode

2. Proposed Circuit Configuration

The VDGA was first introduced in [21], as illustrated in Figure 1. The VDGA device is a six-port versatile active building block described by the following matrix equation [21]-[22]:

$$\begin{bmatrix} i_z \\ i_{zc} \\ i_x \\ v_w \end{bmatrix} = \begin{bmatrix} g_{mA} & -g_{mA} & 0 \\ -g_{mA} & g_{mA} & 0 \\ 0 & 0 & -g_{mB} \\ 0 & 0 & \beta \end{bmatrix} \begin{bmatrix} v_p \\ v_n \\ v_z \end{bmatrix} . \qquad (1)$$

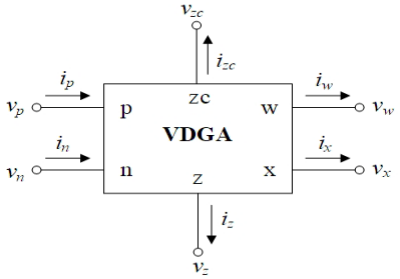


Figure 1: Circuit symbol of the VDGA.

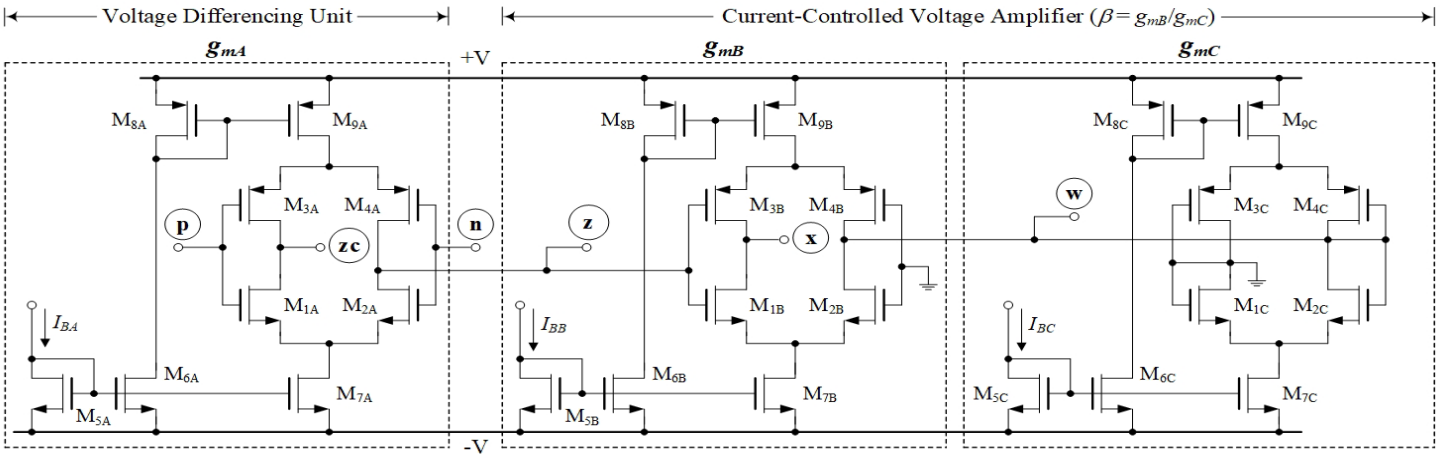


Figure 2: CMOS internal structure of the VDGA.

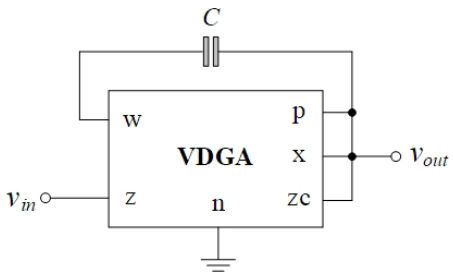


Figure 3: Proposed single VDGA-based resisotrless phase shifter circuit.

In (1), g_{mk} ($k = A, B, C$) and β represent the transconductance gain and the voltage transfer gain of the VDGA, respectively. This element has two high input impedance voltages (v_p and v_n), three high output impedance currents (i_z, i_{zc} and i_x), and a zero output impedance voltage (v_w).

The values of g_{mk} and β, when implemented in CMOS technology such as that depicted in Figure 2 [22]-[23], can be expressed as follows:

$$g_{mk} = \sqrt{K\left(\frac{W}{L}\right)I_{Bk}} \quad , \tag{2}$$

and

$$\beta = \frac{g_{mB}}{g_{mC}} \quad . \tag{3}$$

where $K = \mu_0 C_{ox}$ is the transistor transconductance, μ_0 is the carrier mobility, C_{ox} is the gate-oxide capacitance per unit area, and W/L is the width-to-length ratio of the transistor. From Figure 2, the CMOS VDGA is made up of three sections of transconductance amplifiers (M_{1A} - M_{9A}, M_{1B} - M_{9B} and M_{1C} - M_{9C}). Each transconductor contributes its own transconductance gain g_{mk} that is electronically controllable. Consequently, external bias currents I_{Bk} can be used to adjust the parameters g_{mk} and β of the VDGA.

Figure 3 depicts the realization of the phase shifter circuit that requires only one VDGA and one floating capacitor without an external resistor requirement. Despite the fact that the capacitor C employed in this realization is floating, a second poly-layer technique is provided by advanced integrated circuit (IC) technology, making it simple to implement [24]. A preliminary analysis of the proposed configuration in Figure 3 gives the voltage transfer function shown below

$$\frac{V_{out}(s)}{V_{in}(s)} = \beta\left(\frac{\frac{sC}{g_{mC}} - 1}{\frac{sC}{g_{mA}} + 1}\right) \quad . \tag{4}$$

Assuming $g_m = g_{mA} = g_{mC}$, the passband gain (H_0), pole frequency (f_p) and phase response (ϕ) of the configuration are obtained as:

$$H_0 = \beta \quad , \tag{5}$$

$$f_p = \frac{\omega_p}{2\pi} = \frac{g_m}{2\pi C} \quad , \tag{6}$$

and

$$\phi = \pi - 2\tan^{-1}\left(\frac{\omega C}{g_m}\right) \quad . \tag{7}$$

Thus, the transconductances g_{mk} or by changing the external bias currents I_{Bk} can be modified to alter the values of H_0, ω_p and ϕ. Also noticed is the fact that the gain β can be controlled to provide orthogonal H_0 control.

3. Effects of Non-Ideal Gains

Ideally, the VDGA features are thought to be perfect. However, due to device mismatch, transfer errors may occur in CMOS implementations of VDGA, deviating from the expected behavior. The impact of the VDGA non-idealities on the functioning of the suggested circuit must thus be investigated. In view of VDGA's non-ideal gains, (1) may be changed and expressed as:

$$\begin{bmatrix} i_z \\ i_{zc} \\ i_x \\ v_w \end{bmatrix} = \begin{bmatrix} \alpha_A g_{mA} & -\alpha_A g_{mA} & 0 \\ -\alpha_A g_{mA} & \alpha_A g_{mA} & 0 \\ 0 & 0 & -\alpha_B g_{mB} \\ 0 & 0 & \delta\beta \end{bmatrix} \begin{bmatrix} v_p \\ v_n \\ v_z \end{bmatrix} \quad , \tag{8}$$

where $\alpha_k = 1 - \varepsilon_\alpha$ represents the transconductance inaccuracy coefficient, and $\delta = 1 - \varepsilon_\delta$ represents the parasitic voltage transfer

gain. Here, ε_α ($|\varepsilon_\alpha| << 1$) and ε_δ ($|\varepsilon_\delta| << 1$) are the undesirable parameters deviating from unity due to the transfer errors of the VDGA. For the non-ideal analysis of the suggested phase shifter circuit in Figure 3, the modified parameters H_0, f_p, and ϕ can be given by the following expressions:

$$H_0 = \delta\beta \quad , \tag{9}$$

$$f_p = \frac{\alpha_A g_m}{2\pi C} \quad , \tag{10}$$

and

$$\phi = \pi - 2\tan^{-1}\left(\frac{\delta\omega C}{\alpha_B g_m}\right) \quad . \tag{11}$$

It is evident from (9)-(11) that the circuit parameters H_0, f_p and ϕ are slightly affected by the unwanted factors α_k and δ of the VDGA. However, this effect can be diminished by modifying the transconductance gain g_{mk} for the circuit shown in Figure 3. According to (2), the value of g_{mk} can be modified conveniently by altering the external bias current I_{Bk}.

4. Simulation Verification

In order to evaluate the performance of the proposed phase shifter circuit in Figure 3, the CMOS-based VDGA in Figure 2 was simulated with the TSMC 0.25-µm transistor model in PSPICE computer simulation program. Symmetrical supply voltages of +V = -V= 0.75V were used to bias the VDGA. The transistor sizes (W and L) used in the VDGA realization are listed in Table 2.

Table 2: Transistor sizes used in VDGA realization of Figure 2.

Transistor	W (µm)	L (µm)
M_{1k}-M_{2k}	15	0.25
M_{3k}-M_{4k}	23	0.25
M_{5k}-M_{7k}	4.5	0.25
M_{8k}-M_{9k}	5.5	0.25

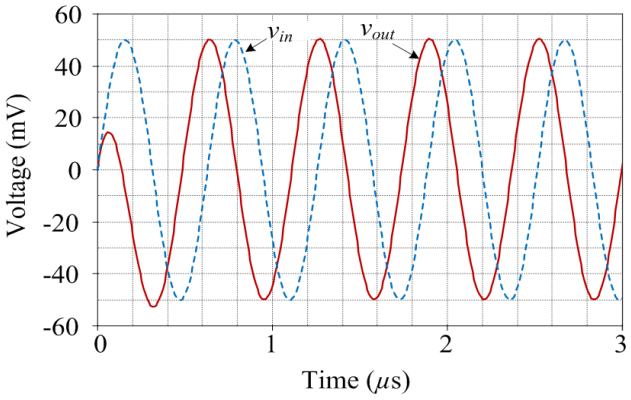


Figure 4: Simulation results of the transient waveforms of the proposed phase shifter in Figure 3.

The active and passive component values are specified as: $g_{mA} = g_{mB} = g_{mC} = 1$ mA/V, ($I_{BA} = I_{BB} = I_{BC} = 100$ μA), and $C = 0.1$ nF for the proposed resistorless phase shifter with $H_0 = 1$ and $f_p = 1.59$ MHz. Figure 4 shows the simulated transient responses of the proposed circuit for an input signal with a sinusoidal frequency of 1.59 MHz and an amplitude of 50 mV (peak). In contrast to the theoretical value of $\phi = 90^\circ$, the simulation results show a phase difference between v_{in} and v_{out} of $\phi = 92^\circ$.

Figure 5 also shows the simulation outcomes for the gain and phase frequency characteristics in comparison to the ideal curves. The simulated f_p is approximately 1.585 MHz, resulting in a frequency error of 0.31 %. The simulation results clearly show that they closely match the theoretical predictions, demonstrating the usefulness of the suggested circuit. It is discovered that the simulated power dissipation of the circuit is around 0.82 mW, when the input v_{in} is kept grounded.

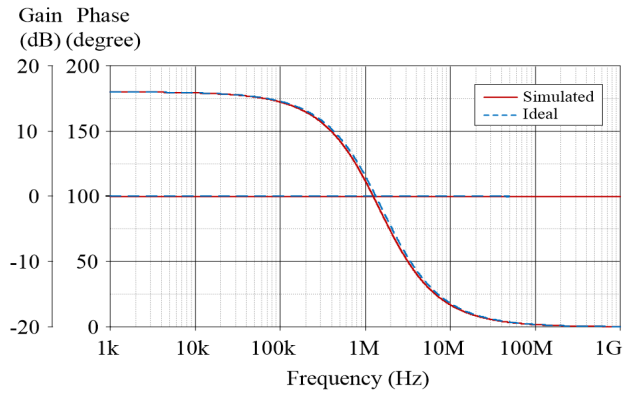


Figure 5: Ideal and simulated frequency characteristics of the proposed phase shifter in Figure 3.

Figure 6 depicts the electronic tuning of H_0 without altering the ϕ-value by controlling the g_{mB}-value. The values of the circuit components for these settings are listed in Table 3. The sinusoidal input waveform in these figures is 20 mV (peak) at f = 1.59 MHz. While $g_{mA} = g_{mC}$ = 1 mA/V ($I_{BA} = I_{BC}$ = 100 µA) remains constant, the values of g_{mB} are altered between 0.707 mA/V, 1 mA/V, and 1.414 mA/V (I_{BB} = 50 µA, 100 µA, and 200 µA). These facts lead to the β-value being, respectively, 0.707, 1, and 1.414.

Table 3: Component values for electronic tuning of H_0 with I_{BB}.

I_{BB} (µA)	g_{mB} (mA/V)	$I_{BA} = I_{BC}$ (µA)	$g_{mA} = g_{mC}$ (mA/V)	β (g_{mB}/g_{mC})
50	0.707	100	1	0.707
100	1.000	100	1	1.000
200	1.414	100	1	1.414

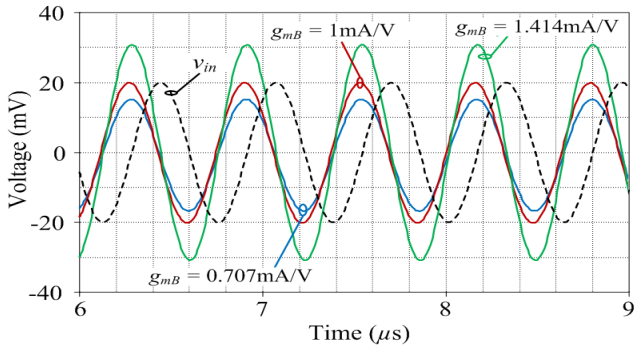


Figure 6: Simulated transient waveforms of the proposed phase shifter with tuning g_{mB} value.

The simulated transient responses of the circuit and its corresponding phase response are also shown in Figures 7 and 8 for three different values of g_m, i.e., $g_m = g_{mk} = 0.707$ mA/V, 1 mA/V and 1.414 mA/V (I_{Bk} = 50 μA, 100 μA, and 200 μA). The computed values of ϕ were determined to be, respectively, 70.5°, 90°, and 109.4°. The measured ϕ values based on the simulation results were 73°, 92.1°, and 110.2°, respectively, which accord quite well with the estimated values.

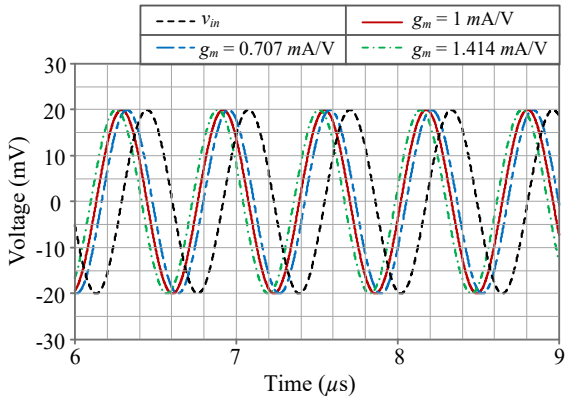


Figure 7: Simulated transient waveforms of the proposed phase shifter with tuning g_{mk} value.

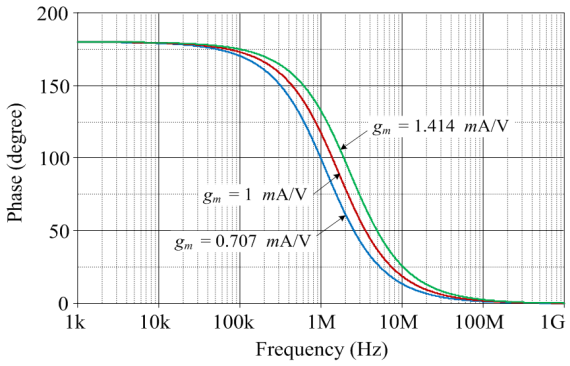


Figure 8: Simulated frequency characteristics of the proposed phase shifter with tuning g_m value.

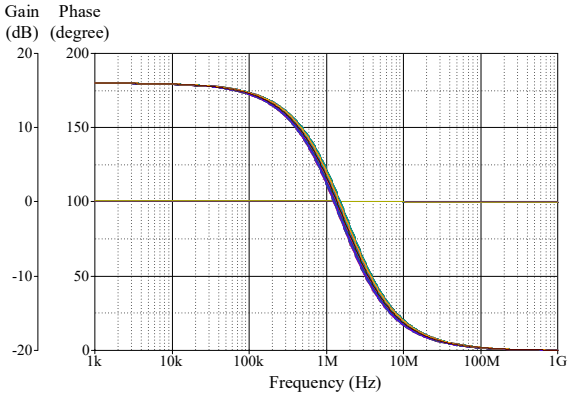


Figure 9: Monte Carlo statistical analysis for the frequency responses of the proposed phase shifter circuit with 5% capacitor tolerance.

In order to demonstrate the impact of capacitor tolerance on the gain and phase responses, a Monte Carlo analysis with a hundred runs is performed for the proposed phase shifter circuit given in Figure 3. It is supposed that the value of capacitor C will change uniformly by 5%. Figure 9 depicts the simulated frequency responses of Monte Carlo statistical analysis. In addition, the results of the Monte Carlo analysis indicate that the mean and sigma of f_p are approximately 1.589 MHz and 2.204 × 10^{10}, respectively. It can be clearly seen from Figure 9 that the capacitance tolerance has a minor effect on the frequency response of the proposed circuit.

5. Experimental Measurements

Experimental measurement was used to validate the practicability of the designed circuit in Figure 3. The schematic for the practical implementation of the VDGA is shown in Figure 10 [25], using readily available IC dual-OTA LM13600s from National Semiconductor [26]. For LM13600s, the DC bias voltages are +V= -V= 5V.

The proposed phase shifter circuit of Figure 3 was constructed with the following component values: g_{mk} = 1 mA/V (I_{Bk} = 100 µA) and C = 1 nF. For time-domain analysis, the circuit was applied with a sinusoidal input of frequency 159 kHz and of amplitude 50 mV (peak). The measured waveforms for v_{in} and v_{out} are shown in Figure 11. The measured ϕ is 92.8°, which is close to the theoretical ϕ of 90°. Accordingly, the measured phase error is about 3.11%. Figure 12 also shows the measured Fourier spectrum of the output waveform v_{out} at 159 kHz.

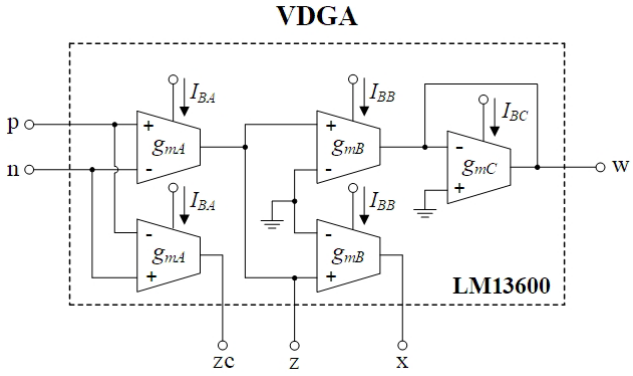


Figure 10: Practical realization of VDGA using readily available IC LM13600s

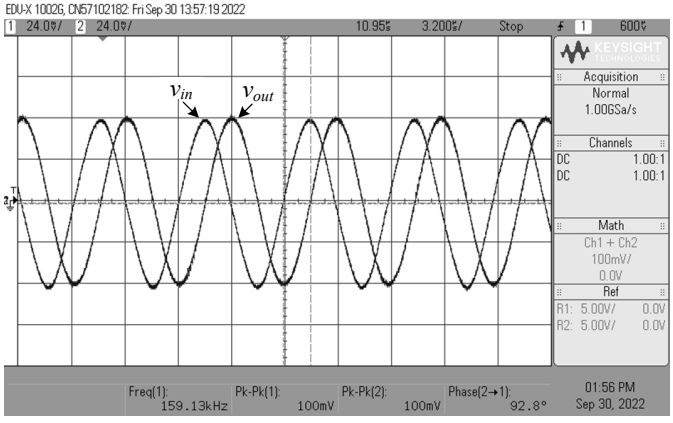


Figure 11: Measured time-domain waveforms of v_{in} and i_{in} for the propose circuit in Figure 3.

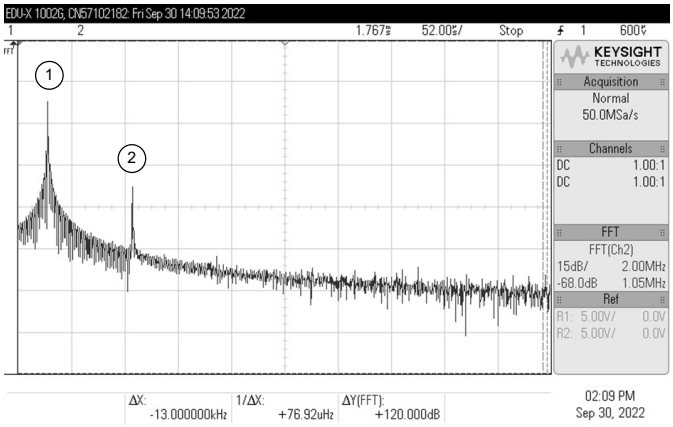


Figure 12: Measured spectrum frequency of v_{out} at 159 kHz (No.1: Frequency = 159 kHz, Gain = -28.91 dB, and No.2: Frequency = 479 kHz, Gain = -29.13 dB)

The next observation on the circuit is carried out on its frequency response characteristic. The measured frequency responses in comparison to the theoretical responses are given in Figure 13. The measured value of f_p is found to be 158 kHz, which corresponds to the frequency deviation of 0.63%. The experimental testing results show that while the gain response is essentially constant up to the working frequency of roughly 4 MHz, the phase characteristic is found to change with frequency, as predicted. The difference between measured and ideal curves in the high-frequency region is predominantly attributable to the gain-bandwidth product of the IC OTA LM13600s used to implement the circuit [26]. Obviously, higher-speed active devices could produce superior frequency responses.

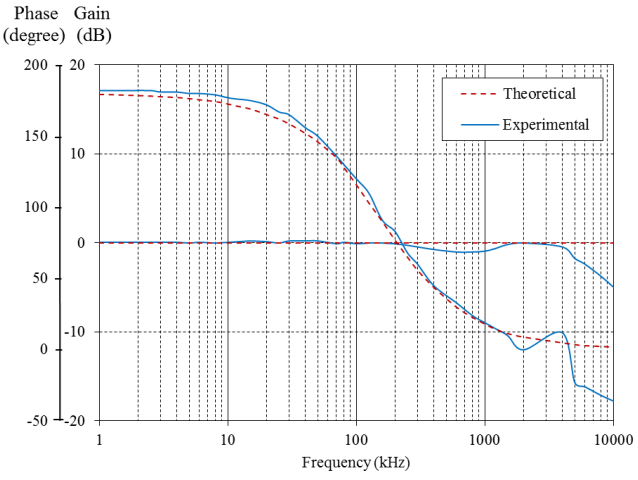


Figure 13: Theoretical and measured frequency responses of the propose circuit in Figure 3.

6. Quadrature Oscillator Application

The quadrature oscillator (QO) circuit can be simply implemented by utilizing the proposed phase shifter circuit, as shown in Figure 14. In the configuration, VDGA2 and C_2 create a simple lossless integrator. The following relationship describes the characteristic equation of the QO circuit:

$$s^2 + g_{mA1}\left(\frac{1}{C_1} - \frac{\beta_1}{C_2}\right)s + \left(\frac{g_{mA1}g_{mB1}}{C_1C_2}\right) = 0 \quad . \qquad (12)$$

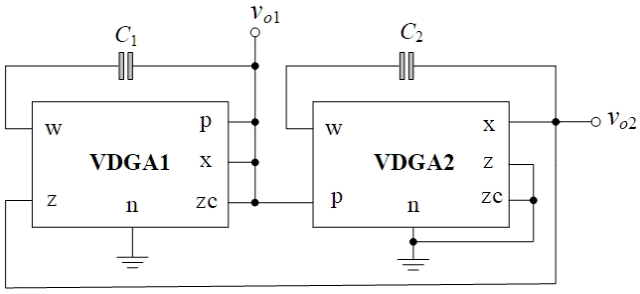


Figure 14: Quadrature oscillator implemented with the proposed circuits.

From the characteristic equation in (12), the condition for oscillation (CO) is satisfied at

$$\frac{\beta_1 C_1}{C_2} \geq 1 \quad , \qquad (13)$$

and the frequency of oscillation (f_o) is obtained as:

$$f_o = \frac{\omega_o}{2\pi} = \frac{1}{2\pi}\sqrt{\frac{g_{mA1}g_{mB1}}{C_1C_2}} \quad . \qquad (14)$$

The equation for the relationship between the produced quadrature signals is

$$\frac{V_{o2}(j\omega)}{V_{o1}(j\omega)} = \left|\frac{g_{mA1}}{\omega C_2}\right| e^{j90^\circ} \quad . \qquad (15)$$

Obviously, both quadrature voltages v_{o1} and v_{o2} are ideally shifted by a phase (ϕ) of 90°. It may also be observed that g_{mA1} and C_2 have a direct impact on the amplitude ratio of the quadrature voltages. Therefore, it follows that the output voltage amplitude of the QO can be controlled by the values of g_{mA1} and C_2. When the frequency is altered, equal voltage amplitudes can be achieved by changing g_{mA1} while maintaining C_2.

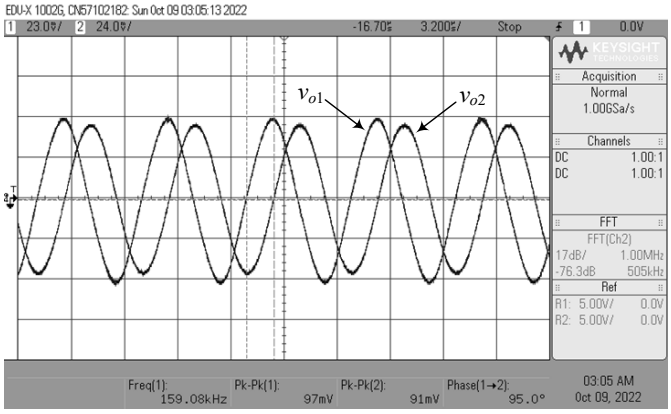


Figure 15: Measured waveforms of the developed QO circuit at v_{o1} and v_{o2} outputs.

The following values were chosen for the circuit elements in order to test the functionality of the QO circuit in Figure 14. The selection of g_{mk} = 1 mA/V (I_{Bk} = 100 μA) and $C_1 = C_2$ = 1 nF for all transconductances and capacitances results in f_o = 159 kHz. Figure 15 shows the typical waveforms measured at v_{o1} and v_{o2}

output terminals with f_o = 159.08 kHz and $\phi = 95^{\circ}$. The resulting deviations for f_o and ϕ are 0.05% and 5.55%, respectively. The differences from ideal values are mainly attributed to the non-ideal gains and parasitic elements of IC LM13600s, which are described in Section 3. In Figure 16, the corresponding Lissajous figure of the QO circuit is also shown. Figure 17 illustrates the measured spectrum frequency of v_{o2} of the QO circuit, with the corresponding frequencies and gains at various spectra listed in Table 4. All the results support the practical usefulness of the proposed phase shifter circuit in implementing the quadrature oscillator.

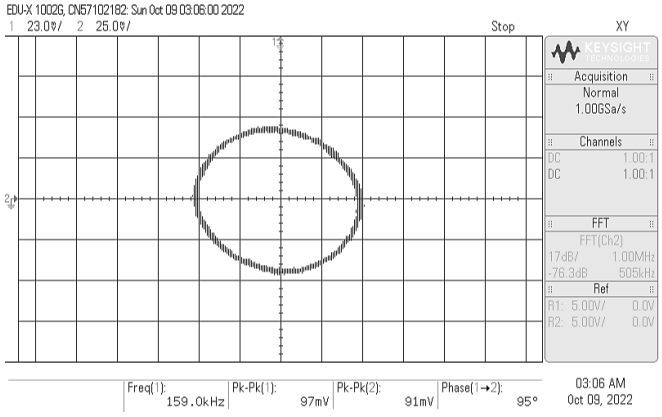


Figure 16: Lissajous figure of the developed QO circuit in Figure 14.

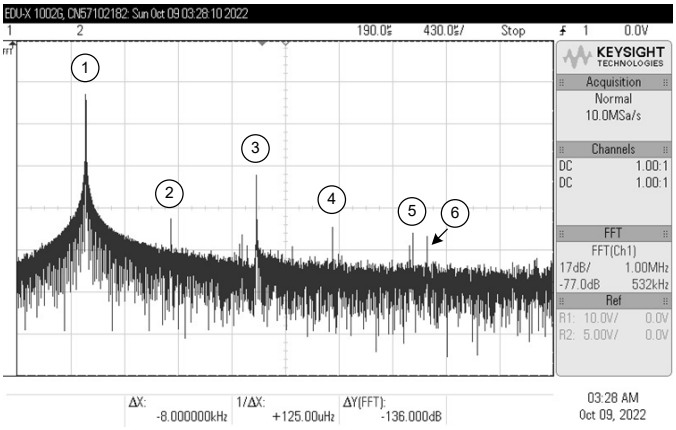


Figure 17: Measured spectrum frequency of v_{o2} at 159 kHz.

Table 4: Frequencies and gains at different spectrums of Figure 17.

No.	Frequency (kHz)	Gain (dB)
1	159	-30.798
2	318	-79.650
3	478	-63.720
4	620	-86.022
5	770	-88.677
6	796	-89.208

7. Conclusions

The paper describes the design of the compact resistorless tunable phase shifter circuit. The described phase shifter circuit requires only one VDGA as an active component and one floating capacitor, resulting in a resistorless architecture and ease of integration. Electronic tuning of the important features of the resulting design, such as the passband gain (H_0), pole frequency (f_p) and phase response (ϕ), is possible by modifying the g_m-values of the VDGA. The non-ideal analysis of the VDGA was also carried out. The voltage-mode quadrature oscillator has been used as an illustrative application for the proposed design. PSPICE simulation data with TSMC 0.25-µm CMOS model parameters have been performed to support the theoretical research. In addition to validating the practical circuit behaviors, experimental measurements with commercially available IC LM13600s have been included.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by King Mongkut's Institute of Technology Ladkrabang [2566-02-01-041]. The Institute of Research and Development Rajamangala University of Technology Isan, and the Faculty of Engineering, Rajamangala University of Technology Isan, Khonkaen Campus, are also acknowledged for their providing support.

References

[1] W. Tangsrirat, T. Pukkalanun and W. Surakampontorn, "Resistorless realization of current-mode first-order allpass filter using current differencing transconductance amplifiers", Microelectronics Journal, **41**(2-3), 178-183, 2010, doi.org/10.1016/j.mejo.2010.02.001.

[2] J. W. Horng, "Current conveyors based allpass filters and quadrature oscillators employing grounded capacitors and resistors," Computers & Electrical Engineering, **31**(1), 81-92, 2005, doi.org/10.1016/j.compeleceng.2004.11.006.

[3] J. W. Horng, C. L. Hou, C. M. Chang, Y. T. Lin, I. C. Shiu, and W. Y. Chiu, "First-order allpass filter and sinusoidal oscillators using DDCCs," International Journal of Electronics, **93**(7), 457-466, 2006, doi.org/10.1080/00207210600711481.

[4] K. Mathur, P. Venkateswaran, and R. Nandi, "All-pass filter based linear voltage controlled quadrature oscillator," Active and Passive Electronic Components, **2017**(4), 1-8, 2017, doi:10.1155/2017/3454165.

[5] A. Toker, E. O. Gune, and S. Ozoguz, "Current-mode all-pass filters using current differencing buffered amplifier and a new high-Q bandpass filter configuration," IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing, **47**(9), 949-954, 2000, doi.org/10.1109/82.868465

[6] A. Toker, E.O. Gune, and S. Ozoguz, "New high-Q band-pass filter configuration using current controlled current conveyor based all-pass filters," in the 8th International Conference on Electronics, Circuits and Systems (ICECS 2001), 165-168, 2001.

[7] J. Satansup, and W. Tangsrirat, "Single VDGA-based first-order allpass filter with electronically controllable passband gain," in the 7th International Conference on Information Technology and Electrical Engineering (ICITEE 2015), 106-109, 2015.

[8] O. Channumsin, and W. Tangsrirat, "Single VDBA-based phase shifter with low output impedance," in the 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2017), 427-430, 2017.

[9] A. Chaichana, S. Siripongdee, and W. Jaikla, "Electronically adjustable voltage-mode first-order allpass filter using single commercially available IC," in International Conference on Smart Materials Applications, Tokyo, Japan, **559**(1), 012009, 2019.

[10] S. Maheshwari, J. Mohan, and D. S. Chauhan, "Voltage-mode cascadable all-pass sections with two grounded passive components and one active element," IET Circuits, Devices and Systems, **4**(2), 113-122, 2010, doi.org/10.1049/iet-cds.2009.0167.

[11] D. S. Chauhan, G. Garg, J. Mohan, and S. Maheshwari, "Two DDCC based cascadable voltage-mode first-order all-pass filters," in International Conference on Advances in Electronics, Electrical and Computer Science (EEC 2012), 290-294, 2012.

[12] B. Metin, N. Herencsar, and K. Pal, "Supplementary first-order all-pass filters with two grounded passive elements using FDCCII," Radioengineering, **20**(2), 433-437, 2011.
[13] B. Chaturvedi, and S. Maheshwari, "An ideal voltage-mode all-pass filter and its application," Journal of Communication and Computer, **9**, 613-623, 2012.
[14] E. Yuce, R. Verma, N. Pandey, and S. Minaei, "New CFOA-based first-order all-pass filters and their applications," International Journal of Electronics and Communications (AEU), **103**, 57-63, 2019, doi.org/10.1016/j.aeue.2019.02.017.
[15] E. Yuce, L. Safari, S. Minaei, G. Ferri, and V. Stornelli, "New mixed-mode second-generation voltage conveyor based first-order all-pass filter," IET Circuits, Devices & Systems, **14**(6), 901-907, 2020, doi.org/10.1049/iet-cds.2019.0469.
[16] J. Jitender, J. Mohan, and B. Chaturvedi, "A novel voltage-mode configuration for first order all-pass filter with one active element and all grounded passive components," in the 6th International Conference on Signal Processing and Communication (ICSC 2020), 2020.
[17] E. Yuce, and S. Minaei, "A new first-order universal filter consisting of two ICCII+s and a grounded capacitor," International Journal of Electronics and Communications (AEU), **137**, 153802, 2021, doi.org/10.1016/j.aeue.2021.153802.
[18] R. Senani, D. R. Bhaskar, and P. Kumar, "Two-CFOA-grounded-capacitor first-order all-pass filter configurations with ideally infinite input impedance," International Journal of Electronics and Communications (AEU), **137**, 153742, 2021, doi.org/10.1016/j.aeue.2021.153742.
[19] J. Jitender, J. Mohan, and B. Chaturvedi, "CMOS realizable and highly cascadable structures of first-order all-pass filters," Walailak Journal of Science and Technology (WJST), **18**(14), 21451, 2021, doi.org/10.48048/wjst.2021.21451.
[20] A. Raj, D. R. Bhaskar, R. Senani, P. Kumar, "Four unity/variable gain first-order cascaded voltage-mode all-pass filters and their fully uncoupled quadrature sinusoidal oscillator applications," Sensors, **22**(16), 6250, 2022, doi.org/10.3390/s22166250.
[21] J. Satansup, W. Tangsrirat, "CMOS realization of voltage differencing gain amplifier (VDGA) and its application to biquad filter," Indian Journal of Engineering and Material Sciences, **20**(6), 457-464, 2013.
[22] O. Channumsin, T. Pukkalanun and W. Tangsrirat, "Single VDGA-based dual-mode multifunction biquadratic filter and quadrature sinusoidal oscillator," Informacije MIDEM-Journal of Microelectronics, Electronic Components and Materials, **50**(2), 125-136, 2020, doi.org/10.33180/InfMIDEM2020.205.
[23] W. Tangsrirat, T. Pukkalanun, O. Channumsin, "Dual-mode multifunction filter realized with single voltage differencing gain amplifier (VDGA)", Engineering Review, **41**(2), 1-14, 2021, doi.org/10.30765/er.1441
[24] R. J. Baker, H. W. Li, and D. E. Boyce, CMOS circuit design, layout and simulation, Chapter 7, IEEE Press, New York, 1998.
[25] N. Roongmuanpha, W. Tangsrirat, T. Pukkalanun, "Single VDGA-based mixed-bode universal filter and dual-mode quadrature oscillator", Sensors, **22**(14), 5303, 2022, doi.org/10.3390/s22145303.
[26] National Semiconductor. Dual operational transconductance amplifiers with linearizing diodes and buffers. LM13600 datasheet 1998.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 1, 138-148 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

A Model for Teaching Mathematics to Gifted Students Based on an Effective Combination of Various Approaches for their Preparation

Zhanna Dedovets[*,1], Mikhail Rodionov[2], Anna Novichkova[3]

[1]*Department of School of Education, The University of the West Indies (UWI), Trinidad and Tobago*

[2]*Department of Computer and Mathematical Education, State Pedagogical University named after S. Ayni (TSPU), Tajikistan*

[3]*Gymnasium named after Abulfazli Balami, Vahdat, Tajikistan*

ARTICLE INFO

Article history:
Received: 27 November, 2022
Accepted: 15 January, 2023
Online: 24 February, 2023

Keywords:
Mathematically gifted students
Model of teaching mathematics
Mathematical Olympiads
Discovery-based method
Partial discovery-based method
Problem-based learning
Project-based learning

ABSTRACT

Currently one of the urgent goals of mathematical education is the organization of effective work with gifted students. Based on the study of various approaches to teaching mathematically gifted students, many years of experience of teachers, students' work, and an analysis of curricula and materials for schools with in-depth study of mathematics, an author's model for the training of gifted students was developed. The novelty of this model is that it ensures a rational combination of various forms of education for gifted children on the basis of differentiation, individualization of the process of teaching mathematics, advanced learning, openness, democracy, reflection, and adequate control. The pedagogical experiment was carried out for two years in the Abulfazl Balami gymnasium for gifted children in the city of Vahdat, Republic of Tajikistan. 41 students and 18 teachers took part in the experiment. The data obtained from the experimental and control groups were subjected to qualitative and quantitative analysis. Over the same time interval there were significant changes in the performance of students in the experimental groups, with 40% of the students moving to a higher level. In the control groups, the change was not significant.

1. Introduction

As is well-known, modern society sets complex tasks for citizens that require non-standard decisions and the manifestation of critical and creative thinking in the constantly changing and non-predictable environment of our world. Accordingly, in a modern school, the discovery-based method as an effective aid for working with gifted children and its optimal organization, combining various formats of such work, is increasingly coming to the fore [1-3].

Despite many studies of the theory and practice of working with gifted children, this issue has not been sufficiently discussed in the methodological and mathematical studies known to us, which indicates the relevance and importance of the present study.

The response to this challenge should consider the specifics of the subject. The process of teaching mathematics to gifted students at school seems to be quite laborious, since here it is necessary, firstly, to ensure the development of the main curriculum, and, secondly, to effectively develop their "non-standard potential". These two goals in existing educational practice are not always coordinated. Elective and basic mathematics courses are taught by different teachers under time pressure. Moreover, this often do not take into account the individual characteristics of students.

As a rule, in the real educational process, the work of a mathematics teacher, both in ordinary and specialized classes, focuses primarily on the development of students' basic competencies, which are taken into account during the current and final certification. The second component of this process - the actual mathematical development - is decided on elective courses and consultations of various kinds by completely different people - "invited lecturers' within their areas of expertise. These two groups of teachers do not always have the opportunity and desire to closely contact each other professionally. As a result, gifted students study, as it were, two different subjects: "basic mathematics" and "Olympiad mathematics', which does not always allow them to effectively realize the developing potential of the studied mathematical content. This was confirmed by our survey of mathematics teachers, most of whom indicated that the spontaneous interaction of the various formats of such training, as

[*]Corresponding Author: Zhanna Dedovets dedzhanna333@gmail.com

https://dx.doi.org/10.25046/aj080117

occurs in the existing system of training gifted children, does not fully ensure its effectiveness. From the foregoing, it is evident that there is a need for theoretical development and practical implementation of a special model for the rational combination of basic and elective courses in the process of teaching mathematics to gifted students. Summarizing, we can conclude that as a rule, the traditional strategy for teaching mathematics to gifted children is one-sided, being limited to an emphasis on elective courses, or individual consultations. At the same time, the learning process itself is spontaneous in nature, not providing for the educational needs of all such students. An analysis of the pedagogical studies known to us revealed that the issue of a rational combination of various formats for mathematical training of gifted schoolchildren was not specifically addressed in these studies.

Thus, the relevance of this study stems from the need for theoretical development and practical implementation of a holistic strategy for working with gifted children in the field of mathematics, which includes the possibility of a rational combination of various learning formats for each gifted student.

2. Mathematically Gifted Students

Many psychological and pedagogical studies have been devoted to the mathematical development of gifted children [4-10]. According to most authors, mathematical giftedness is understood as a kind of intellectual giftedness, which is associated with and develops in special mathematical activity. Its basic characteristics are integrity, multicomponent nature, hierarchy and dynamism [11, 12]. Mathematically gifted schoolchildren are characterized by the ability to think logically, the ability to operate with mathematical symbols, quickly and correctly solve mathematical problems, successfully moving from simple to more complex mathematical constructions.

Such students have a flexible mind, that is, they are able to find a way out of a non-standard mathematical situation and they have a well-developed abstract memory [13]. Approaches to the study of mathematical giftedness, reflected in the literature, are very diverse: they are based on the psychology of individual differences in students, on the special abilities of mathematically gifted students.

The scientists Joy Gilford, Ellis Torrance, Frank Barron and Charles Taylor carried out a number of major studies in the psychology of giftedness and contributed to the unification of theoretical studies on the psychology of individual differences and practical work on the construction of new curricula in the field of differentiated learning [14-18]. They found that giftedness was manifested in the fact that, unlike for ordinary, traditional experiments, students built their own tasks. Scientists have observed the behavior of creatively gifted people in natural situations of communication, work and leisure. They tried to determine the specific manifestations of talent in various activities, as well as the characteristic features of the personality of gifted people, which emerged in behavior, thinking, inclinations and attitudes. The tasks were set to change the idea of giftedness as "a symptom of hereditary degeneration of the epileptoid type" [19]. This process of change took place over a period of more than 30 years from the beginning of the 20th century. The results of their research have shown that by the end of school, many gifted children sometimes experience severe depression. They are forced to hide their giftedness from their peers and adults. Gifted children experience "discrimination" in school due to the lack of differentiated teaching, due to the school's focus on the average student, due to excessive reduction to a uniform system of curricula [20].

Psychologists Sergei Rubinstein and Boris Teplov developed a classification of the concepts of "ability", "giftedness" and "talent". The classification was carried out according to the success of the activity [16]. Abilities are considered as individual psychological characteristics that distinguish one person from another, on which the possibility of success in activity depends, and giftedness is considered as a qualitatively unique combination of abilities (individual psychological characteristics), on which the possibility of success in activity also depends.

In various definitions of the concept of giftedness (source), a number of basic features of giftedness can be traced. A person has:

(1) outstanding (high level) abilities,
(2) developed intelligence,
(3) an increased level of mental development,
(4) creative approach,
(5) the possibility of achieving high results in various activities.

Intellectual giftedness is a developing systemic quality of the personality psyche in the structure of general abilities. The development of this quality requires a holistic didactic approach to working with gifted adolescents [6]. The personal growth of intellectually gifted adolescents depends on the type of educational environment. The environment should contribute to the disclosure and optimal manifestation of the creative nature of the psyche of gifted adolescents. By minimizing the difficulties of a gifted child in contact with his environment, the educational environment contributes to the adequate personal development of gifted adolescents [21].

Professor Gennadiy Sarantsev in his works talks about methods for developing the ability of gifted children to solve non-standard tasks. He considered various heuristic approaches to solving problems and building new curricula in the field of differentiated learning.

Scientists Vadim Krutetsky, Victoria Yurkevich, Irina Levochkina, Elena Kryukova examined in detail the special abilities that characterize mathematically gifted students, as well as ways to recognize them in a child and adolescent.

They emphasize that schoolchildren who are especially gifted in mathematics are characterized by a peculiar mathematical orientation of the mind (the tendency to perceive many phenomena through the prism of mathematical relations, to realize them in terms of logical and mathematical categories).

They single out several components of mathematical talent: the ability to arrive at mathematical generalization; rationality of the decision (the ability to find the shortest way to solve, cut off the excess, not directly related to the achievement of the goal, the task); a sense of "mathematical aesthetics" (the ability to see beauty and elegance in a simple and at the same time witty, concise and economical way of solving a problem); reversibility of thinking; mathematical intuition.

The results of studies of mathematically gifted adolescents show that adolescents gifted in mathematics develop such features of their mental activity as the ability to generalize mathematical material (the ability to see the general in what is outwardly different, singular), the flexibility of thought processes and the desire to find simpler but more effective ways to solve problems [22].

The issue of teaching methods for mathematically gifted schoolchildren is presented in the works of many psychologists and teachers [11, 23]. They consider various methods of working with mathematically gifted students, describe a system of special tasks for gifted students, and describe the necessary conditions for creating a support system for talented students. They discuss the technologies that the teacher should rely on when developing the unique abilities of each child, describe the construction of a program for mathematically gifted students, where they propose to integrate Olympiad tasks into sections of the basic mathematics course (resources). A number of authors describe the development of creative potential in the process of participating in competitions, in the process of solving non-standard problems [24, 25]. They view work with gifted children in the context of differentiated and individualized learning. Such training, as is well known, is implemented on the basis of a full account of the individual and typological characteristics of a person in the form of grouping or ungrouping students and differentiated construction of the learning process in educated groups or individually; learning technology, the purpose of which is to create conditions for the identification of existing inclinations, the effective development of the interests and abilities of students [26].

However, the authors known to us do not specifically consider the correlation and connections of various approaches to differentiated work with gifted students, the conditions for their effective integration within the framework of such work, the difficulties that arise in this, and ways to overcome them. The foregoing determines the relevance and significance of building and creating a model for training gifted students in mathematics, based on a rational combination of various approaches for working with them and, in particular, in both basic and elective mathematical courses.

In the context of our study, the “Working Concept of Giftedness” developed by Diana Bogoyavlenskaya and Vladimir Shadrikov is of great interest [7]. This concept involves the disclosure of giftedness on the basis of the theoretical provisions of psychology and the definition of basic principles in solving the problems of identifying, training and developing gifted children. ‘Giftedness’ is interpreted as a systemic quality that characterizes the child's psyche as a whole. At the same time, it is the personality, its orientation and the system of values that lead to the development of abilities and determine how its potential will be realized.

Giftedness entails a humanistic approach to the education and development of gifted students, that is, special attention is paid to caring for a gifted child, which implies an understanding of not only the advantages, but also the difficulties that his giftedness brings with it.

3. Methodological framework

As a basis for building the authors’ model of work with gifted schoolchildren in mathematics, they used differentiated and individual approaches to learning and the above-mentioned "Working concept of giftedness" [15, 27]. Differentiation of learning is a process involving the division of students into groups according to their mathematical abilities. Individualization of learning - learning aimed at developing the individual abilities of each student - is an integral element of student-centered learning. Such work may include, for example, external studies, elective courses, individual consultations, implementation of project activities. At the same time, it should not lead to the need to separate gifted students from their peers, but should involve the integration of various collective, group and individual learning activities.

As you know, today, work with gifted children is carried out in schools, gymnasiums, and lyceums. In particular, differentiated education is carried out by dividing classes into profiles: physical and mathematical, humanitarian, chemical and biological, and others. At the same time, work with gifted children is carried out both within the framework of school lessons and within the framework of additional courses (elective courses, courses for preparing for Olympiads, individual lessons). A rational combination of these formats makes it possible to implement regular and systematic work to maximize the full disclosure of the creative potential of schoolchildren [26, 28].

To organize such work, the following factors are necessary:

(1) a strategy for teaching gifted students heuristic methods for solving problems,

(2) continuity of basic and elective mathematical courses of a developing orientation,

(3) a system of students-centered methods of working with gifted students.

4. A strategy for teaching gifted students’ heuristic methods for solving non-standard problems

The main forming factor in organizing such work is the strategy of teaching gifted schoolchildren heuristic methods for solving non-standard problems. Non-standard problems are understood as problems that cannot be solved by standard algorithms known from the basic mathematical course. When solving them, it is necessary to use one or another heuristic procedure.

The essence of heuristic methods for solving problems lies in the fact that the student is naturally involved in the process of rediscovering a non-standard condition of the problem and finding a way to solve it, without having a direct opportunity to apply the basic algorithm for solving. The selection of such tasks and the development of an individual learning plan is a serious challenge for a mathematics teacher. This plan should include the possibility of targeted implementation of several individual learning approaches that correspond to different strategies for working with a particular gifted student.

Scholars define an educational plan as a set of learning stages, forms of learning, and combinations of individual topics from

mathematics curricula. In other words, the educational plan is a differentiated educational program that provides the student with a choice of the type and amount of pedagogical support he needs for his self-determination and self-realization [28].

An individual educational plan is built based on the educational needs, individual abilities and capabilities of the student (level of readiness to master the program), as well as existing regulatory documents. The purpose of such training is to purposefully ensure the differentiation and individualization of the education of gifted children, giving it a personality-oriented character.

An individual educational plan consists of a number of "sections" corresponding to various formats of work (basic course, elective course, individual work, group projects, competitions, Olympiads, etc.). All these approaches should be closely related and organically combined with each other.

5. Ensuring the continuity of the basic and elective components of the mathematical teaching of gifted children

In the combination of various formats of work with gifted children in mathematics mentioned above, a special role is played by the continuity of basic and elective mathematical courses of a developing orientation. This factor, as the analysis of the literature and our own pedagogical experience show, has not yet become one of the imperatives of the educational process for the considered contingent of schoolchildren.

In particular, often different mathematical courses in the same group are taught by teachers of different qualifications (school and university), while the material studied in parallel within these courses is often characterized by "diversity", "patchwork" both in content and in developmental aspects. Overcoming such diversity, obviously, involves providing an organic combination of basic (profile) and elective courses. To illustrate the latter point, we present Figure 1, in which the first and third columns present some topics that are quite important in terms of preparing schoolchildren for mathematical Olympiads. In the central part of the diagram, sections are presented that are the result of the "interaction" of the corresponding basic and additional mathematical courses (Figure 1).

Commenting on the structure and content of the above diagram, it is necessary, first of all, to note that work with gifted students is a holistic, systematic process that emerges from the basic course.

Here, when solving developmental problems, students mainly apply the heuristic method at the stage when they discover some general algorithms that are practiced with all students in a collective form. On the elective course, in a differentiated or individual form, the studied material is deepened by varying and combining the mastered algorithms when solving problems of a search nature. Such work, in turn, creates the basis for mastering various heuristic procedures by gifted students, which, in particular, further contributes to their successful participation in Olympiads at various levels.

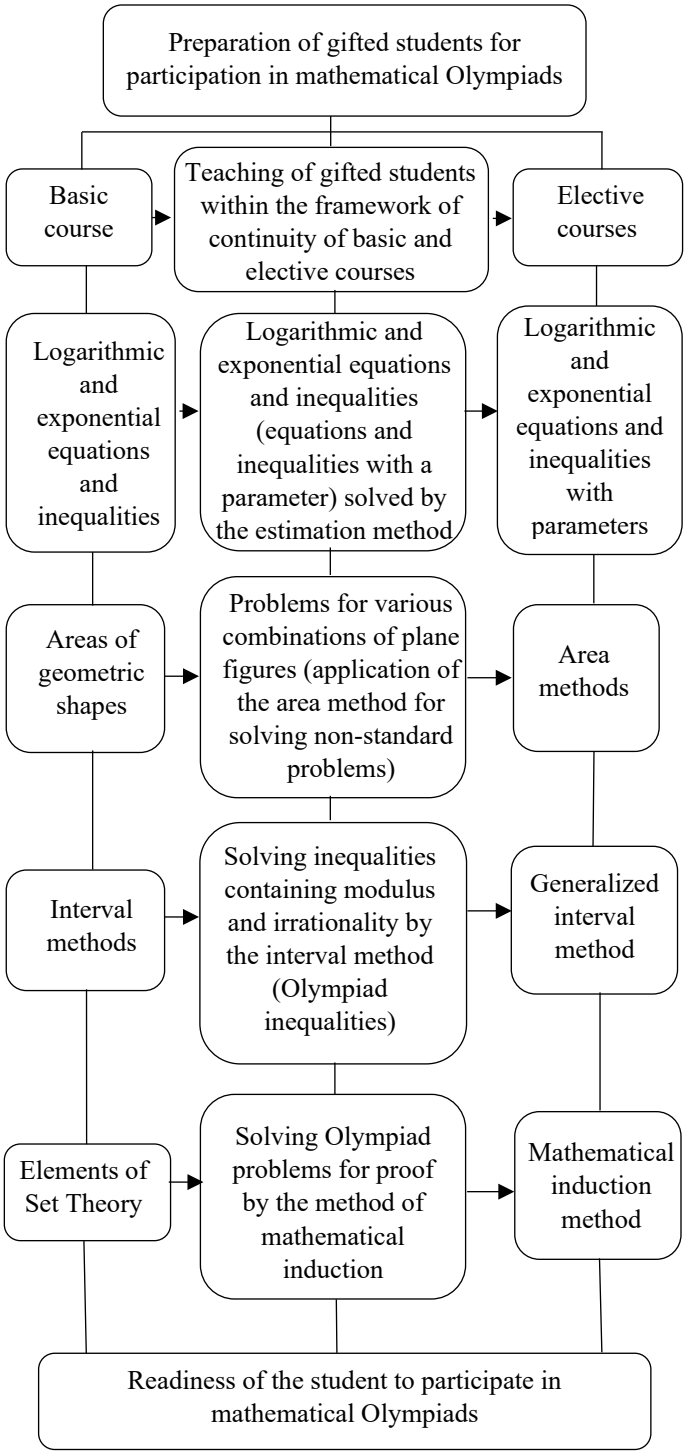


Figure 1: Continuity of Basic and Elective Courses in the Preparation of Students in Grades 4-5 for Olympiads in Mathematics

For example, within the framework of the basic course, the traditional topic "Method of intervals" is studied. Here we touch upon square inequalities, cubic inequalities, inequalities of higher degrees, fractional rational inequalities solved by the interval method based on the sign placement rule. Within the framework of the elective course "Selected Issues of Mathematics", it is advisable to consider the generalized method of intervals, which is used, in particular, in solving irrational and trigonometric inequalities, as well as inequalities containing unknowns under the

module sign, unknowns in the exponent, inequalities containing logarithms of the type:

$$(x+\frac{5}{x})(\frac{\sqrt{x^2-10x+25}-1}{\sqrt{6-x}-1})^2 \geq 6(\frac{\sqrt{x^2-10x+25}-1}{\sqrt{6-x}-1})^2.$$

The consideration of such inequalities is directly based on the material of the basic course, while providing for the variable application of the known algorithm.

At the same time, tasks of a search (Olympiad) nature are beginning to be involved, which are an example of tasks of an even more generalized nature.

Problem: For what values of the parameter a among the solutions of the inequality $(x^2-ax-x+a)\sqrt{x+5} \leq 0$, will there be two solutions, the difference between which is equal to 4?

When solving this problem, obviously, the generalized method of intervals is used and the analysis of all possible cases of the state of the considered mathematical construction is carried out, depending on the value of the parameter a. Accordingly, when analyzing a possible solution path, it is advisable for the teacher, if necessary, to rely on the material of the basic mathematics course, projecting it onto a higher level of generalization of the content.

Irrational and trigonometric inequalities are included in the curriculum for mathematics in high school. Here they are considered to be inequalities, the solution of which, depending on their complexity, is carried out by the method of intervals both at the basic and advanced levels (in the elective course), as well as in preparation for the Olympiads. The method of intervals in various modifications is used at all stages of the study of inequalities, providing a connection, in the context under consideration, between the relevant substantive sections of the basic and elective courses.

The interval method can be used in solving irrational inequalities of a certain type, subject to the appropriate restrictions arising from the properties of the arithmetic root of the nth degree. The algorithm for using this universal method is well known.

6. Model of preparation of gifted children in mathematics

Based on the idea of purposefully ensuring the continuity of basic and elective courses, we have built and implemented a methodological model for working with gifted students in mathematics (Table 1).

Table 1: Model of Preparation of Gifted Children in Mathematics

<table>
<tr><td colspan="2">Purpose:
Creation of conditions for the development of mathematical abilities of gifted students, their self-development, the harmonious development of the personality of a unique child; ability to independently acquire and apply knowledge</td></tr>
<tr><td>Tasks:</td><td>1. Identification of gifted students in the field of mathematics
2. Development of methodological support for the effective development of the mathematical abilities of gifted students, providing for the continuity of various forms of their preparation
3. Implementation of mathematical training of schoolchildren, aimed at achieving socially and personally significant results
4. Approbation and monitoring of the proposed methodological solutions</td></tr>
<tr><td colspan="2">Principles of building the educational process of gifted students</td></tr>
<tr><td colspan="2">The principle of rational combination of various forms of work with gifted children, the principle of differentiation and individualization of the learning process, the principle of student-centered learning, the principle of advanced learning, the principle of openness, the principle of adequate control, the principle of democracy, the principle of reflection</td></tr>
<tr><td rowspan="4">The content of work with gifted children in mathematics</td><td>Profile course
When teaching gifted children in mathematics, the existing programs of specialized courses and relevant textbooks are involved</td></tr>
<tr><td>Elective course
When teaching children gifted in mathematics, topics are considered that deepen the relevant topics of the profile course, and topics that go beyond its scope (for example, the method of mathematical induction, graph theory, etc.)</td></tr>
<tr><td>Project work
When working on an individual project, attention is paid to topics that go beyond the core and basic courses, topics that affect the relationship of mathematics with other areas of knowledge, non-standard solutions to standard tasks are considered (for example, the project “Ten ways to solve one quadratic equation”)</td></tr>
<tr><td>Preparation for the Olympics
When preparing schoolchildren for participation in the Olympiads, first of all, various sets of tasks are considered, which include non-standard tasks (e.g. coloring problems, double counting problems)</td></tr>
<tr><td colspan="2">Forms and methods of organizing work with gifted children in mathematics</td></tr>
<tr><td>Forms of work organization
1.Standard school lesson as part of the basic course
2.Standard school lesson within the profile course
3.Elective courses in mathematics
4. Math events
5. Work on an individual project
6. Individual preparation for the Olympiads
7. As a result of work on all previous formats - participation in subject Olympiad</td><td>Dominant methods of organizing work
1.Discovery-based method
2.Partial discovery-based method
3. Problem-Based learning
4. Project-Based Learning</td></tr>
<tr><td colspan="2">Expected learning outcomes for gifted students</td></tr>
<tr><td colspan="2">1. Successful participation of students in Olympiads and conferences
2. Readiness to pass exams
3. The formed ability of students to independently acquire and apply knowledge in the framework of project-based learning</td></tr>
</table>

4. Increasing motivation to work on solving problem of a discovery-based method and partial discovery-based method nature
Criteria for success in working with gifted students
It is necessary to understand how much the student's giftedness was enhanced, which is manifested in the following indicators of the student's preparation: 1) can work with mathematical text, solve text problems of increased difficulty 2) can solve non-standard problems using heuristic methods 3) can solve problems of an increased level of complexity by various methods, choosing from them the more rational one 4) can effectively solve problems of an Olympiad nature 5) has high-level thinking according to Bloom's taxonomy, has high levels of cognitive ability: a) identifies hidden (implicit) assumptions in the problem, evaluates the significance of the data (analysis) b) uses knowledge from various fields to make a plan for solving a non-standard problem (synthesis) c) has the ability to evaluate particular mathematical material, that is, he can select and study in-depth material based on different criteria (evaluation)

As can be seen from this table, the formulation of this model is guided by the principles of a rational combination of various forms of work with gifted children, differentiation and individualization of the learning process, student-centered learning, advanced learning, openness, democracy, reflection and adequate control.

Let us briefly explain the content of the last three principles in our understanding (the content of the rest is obvious).

The principle of openness implies an approach to the subject of study as potentially open, allowing constant expansion and generalization by connecting initially non-obvious meaningful relationships.

The principle of democracy presupposes the right of the student to voluntarily choose the level and the corresponding form of education that he considers most acceptable.

The principle of reflection implies the need for the teacher to constantly monitor the nature of the interaction of a gifted student with peers, his behavior in situations of success and failure. In the course of devising a problem, it is necessary to carry out an ongoing adjustment of the individual plan for the training and education of a gifted student.

Finally, the principle of adequate control presupposes a variety of diagnostic tools for assessing the mathematical training and development of gifted schoolchildren, which are not limited to existing regulatory documents (control and independent work, exam materials, competitions and Olympiads of various levels).

This diagram shows the dominant methods of working with gifted students, their relationship. We will reveal the essence of each of them. All these methods are based on the active discovery-based and creative activity of students.

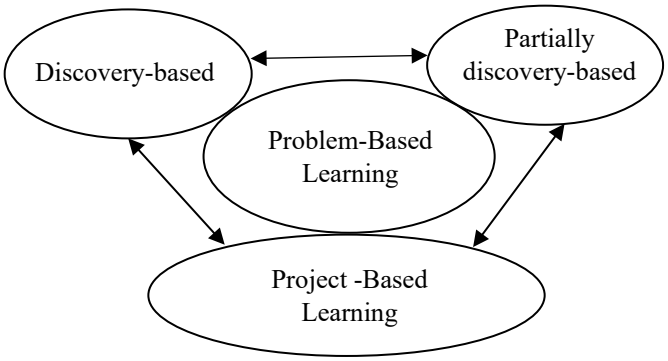


Figure 2: The relationship of various methods of working with gifted students

6.1. Partially discovery-based teaching method

In order to gradually bring students closer to independent problem solving, including non-standard tasks, it is necessary to first teach how to perform individual steps of solving a problem, individual stages of search [1, 2, 6, 29]. In this case, the partial discovery-based method of teaching should be used. It is necessary to offer students independent work with a mathematical text, to provide them with the opportunity to independently derive and formulate the basic mathematical concepts by means of a plan devised using heuristic conversation, well-known sources and their own search work. Another variant of this method is to break down a complex problem into a series of available subproblems, each of which makes it easier to approach the solution of the main problem. To complete a separate part of the task, a gifted student has to draw on knowledge from various branches of mathematics, which he could meet both when studying the basic program, and in elective classes. Thus, within the framework of this method, the teacher constructs a task, divides it into auxiliary ones, outlines the steps to find a solution to the problem. Students, on the other hand, perceive the task, comprehend its condition and solve part of the problem, actualizing the available knowledge and discovering the necessary information, exercising self-control in the process of arriving at a solution and self-motivating. But at the same time, the student's activity does not involve planning the research stages and correlating the stages with each other. The partial discovery-based method is used within the framework of the basic course and elective classes.

6.2. Discovery-based teaching method

The next method that should be introduced into the process of teaching gifted students is the discovery-based method of teaching. Within the framework of this method, the activities of students target the independent resolution of practical problems that require creative solutions–hypotheses [1, 13, 30]. The teacher gradually moves away from drawing up a plan for solving problems, dividing them into subtasks, and invites gifted students to put forward their own hypotheses to find a solution. This method can be used both during the development of a profile course when solving problems of an increased level of complexity (for example, problems with parameters), and in the process of working on a student research project. This method is actively

used in solving Olympiad problems. Initially, the student expresses a hypothesis for solving such a problem, and then builds his own small study to prove or disprove it. Also, this method will be appropriate when mastering new knowledge. An important role is played by independent experiments on the derivation of basic mathematical concepts and statements.

6.3. Problem-based Learning

Within the framework of the data of the problem-based method of teaching, a gifted student always faces a problem that requires a creative, heuristic approach to its solution [25, 31-33]. First, in the basic course, at the very beginning of the elective courses, the teacher himself puts forward small problem situations to the students, dividing the more complex ones. Further, the students themselves meet and recognize them, organizing their research activities to find solutions to these problem situations. With the help of the problem-based method of teaching, the skill of independent search work is formed, which helps in preparing for the Olympiads, when working on an individual research project. The problem-based method is also implemented in the process of participation by schoolchildren in various mathematical festivals and mathematical Olympiads. In these formats, a gifted student is constantly faced with a new, partially or completely unknown mathematical situation that requires a heuristic solution.

6.4. Project-based Learning

One of the most difficult methods for organizing the activities of schoolchildren is the project method. The essence of the project method is the solution of a problem based on the independent activity of students using appropriate methods, means, knowledge, including interdisciplinary, intellectual and practical skills, as well as the realization of creative potential to obtain a specific result [27, 34-35, 39].

In our opinion, the use of projects as applied to gifted students, first of all, helps to maintain constant motivation for an in-depth study of mathematics. Students can choose an interesting topic for themselves, for example, explore different ways to solve one problem and study the proof of a little-known theorem. Thus, the student is constantly developing independently and expanding his knowledge with interesting mathematical facts primarily for himself [36-38].

The project method can also be used as part of a school lesson, an elective course, or an extracurricular activity. It is possible to offer schoolchildren the opportunity to independently acquire knowledge in solving practical problems and problems that require the integration of knowledge from various subject areas. Similar problems are encountered in preparing for examinations in mathematics.

Revealing the structural elements of Figure 2 and their relationship, we can conclude that the proposed system of methods is aimed at developing the research abilities of gifted students and their creative potential in solving non-standard problems. Each method can be implemented or partially implemented in different learning formats. That is, this system of methods is the main integration factor for various formats of training gifted students. Using a system of methods, it is possible to construct a diagram of the relationship between various teaching formats in the preparation of mathematically gifted students (Figure 3).

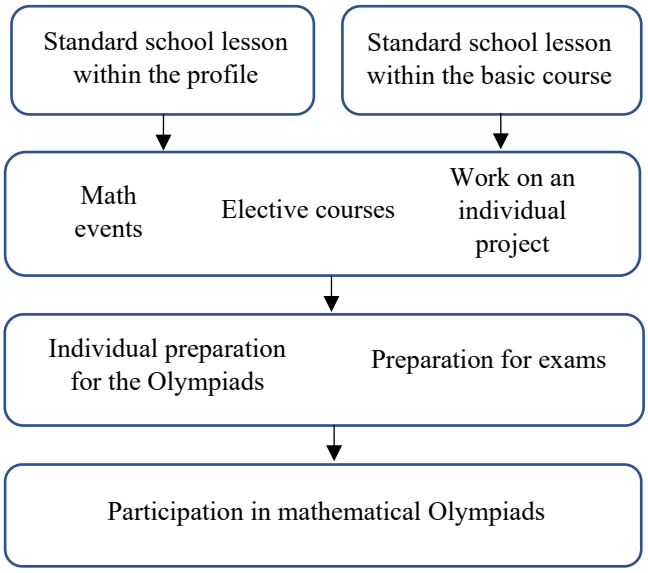


Figure 3: Relationship between different learning forms

7. Implementation of Model of Preparation of Gifted Children in Mathematics

The constructed model was used by us over a period of two years, working as mathematics teachers with gifted students at the Abulfazl Balami gymnasium for gifted children in the city of Vahdat, Republic of Tajikistan.

Let us briefly outline the strategy we employed.

First, during the lesson we observed how students solved partially algorithmic problems and identified which students were potentially capable of mathematics. These students were further involved in attending mathematical circles, where they tried their hand at school-level Olympiads.

If a student developed a sufficiently persistent interest in mathematics, and significant progress in the development of his abilities was evident, then he entered specialized classes. The persistent interest of students was seen in the process of students solving non-standard tasks [33]. The desire to solve such tasks, and, most importantly, the success in applying heuristic methods showed progress in the development of the abilities of the gifted students. Here they more purposefully, in parallel with the lessons of the profile course, continued their studies in the classroom of elective courses. For the most successful students, we developed a special learning plan that involved additional training activities and consultations in addition to elective classes.

In addition to the training of gifted schoolchildren immersed in subject mathematical activity, we made special efforts to ensure that all schoolchildren were constantly involved in joint communicative activities with classmates. In particular, in the classroom they could act as consultants on the subject, constantly participating in school competitions, concerts, sports events. This practice contributed to the development of the communicative abilities of schoolchildren, their emotional and volitional sphere, and reduced the risk of manifestation of possible difficulties in productive interaction with peers and adults around them.

Thanks to such an organization of the educational process, as our experience showed, non-standard abilities of schoolchildren developed and improved quite successfully, which was partially confirmed by the results of examinations and Olympiads.

In more detail, the methodological features of the implementation of this model are reflected in methodological materials we created and published in several articles.

Table 2 presents examples of some individual educational plans that we have built for gifted schoolchildren, as well as the first results of schoolchildren studying within the framework of the constructed model (Table 2).

Table 2: The Results of Work within the Framework of the Constructed Model

Student A - Grade 3 Works within the framework of the model for 1 year	Student B - Grade 4 Works within the framework of the model for 2 years
Forms of study: 1. standard school lesson within the framework of the basic course 2. elective courses 3. mathematical events Results: 1. shows interest in solving non-standard problems 2. Olympiads winner	Forms of study: 1. standard school lesson within the framework of the basic course 2. elective courses 3. individual preparation for the Olympiads 4. participation in mathematical Olympiads Results: 1. can solve problems of an increased level of complexity by various methods, choosing the most rational one 2. has high-level thinking according to Bloom's taxonomy, that is has high levels of cognitive ability 4. has a strong interest in learning mathematics 5. Olympiads winner
Student C- Grade 5 Works within the model for 2 years	Student D- Grade 6 Works within the model for 2 years
Forms of study: 1. standard school lesson within the framework of the basic course - 2. elective courses 3. individual preparation for the Olympiads 4. participation in mathematical Olympiads Results: 1. can solve problems of an increased level of complexity by various methods, choosing the most rational one 2. has a strong interest in the study of mathematics 3. goes to the final rounds of level Olympiads 4. is the winner and prize-winner of the final rounds of level Olympiads	Forms of study: 1. standard school lesson within the framework of the basic course - 2. elective courses 3. individual preparation for the Olympiads 4. participation in mathematical Olympiads 4. preparation for the profile Results: 1. can solve problems of an increased level of complexity by various methods, choosing the most rational one 2. has high-level thinking according to Bloom's taxonomy, that is has high levels of cognitive ability 3. has a strong interest in advanced mathematics 4. goes to the final rounds of level Olympiads 5. is the winner and prize-winner of the final rounds of level Olympiads 6. claims for high scores in the profile

A pedagogical experiment to determine the possibilities of the proposed strategy for the training of gifted students, based on a rational combination of various formats of such training, was carried out over a two-year course of mathematical training of gifted schoolchildren in the Abulfazl Balami gymnasium for gifted children in the city of Vahdat, Republic of Tajikistan and a number of educational organizations in Dushanbe. In total, 41 students of the basic and senior levels of education were included in the experimental sample. The control groups in the study included two classes, in one of which (26 students) the dominant format for preparing gifted students was a series of elective courses devised by the authors. In the other the leading factor in training was individual consultations with specialists (15 students).

The data collection methodology included an analysis of the performance of current diagnostic work, including tasks of increased complexity, as well as taking into account the achievements of schoolchildren participating in the study in mathematical Olympiads of municipal, regional and republican status. In particular, the readiness of schoolchildren to solve non-standard tasks of increased difficulty using heuristic methods, their ability to find and compare different methods of performing the same task and to perform Olympiad tasks of a research nature were assessed. The evaluation of the results was carried out using the traditional five-point scale. The generalized result was considered as an average score for all diagnostic work. Students who received marks in the interval (2;3] were assigned to the first (low) level; in the interval (3;4] to the second (middle) level, and in the interval (4;5] - to the third (highest) level.

The received ordinal data at the pre- and post-implementation stages in the experimental and control groups were subjected to qualitative and quantitative analysis.

At the initial stage, the differences between the groups according to the selected levels proved to be unreliable. Statistical processing using the non-parametric fit method χ^2 - Pearson showed that the empirical values of the Pearson criterion when comparing the distributions of estimates in pairs in three samples proved to be lower than the corresponding critical values for given sample sizes. This fact indicates a relatively similar distribution of gifted schoolchildren by levels of success.

Experimental work within the framework of the ongoing study was carried out for two years. The experimental group studied according to the authors' model described above. For one control group the teacher used elective courses only. For the second control group the teacher used individual consultation only. The educational material in different formats in the control groups did not specifically correlate in any way either, in the program or in the procedural aspect.

As a result of experimental training, control measures were again carried out. A comparison of the dynamics of changes in success in the selected groups of students showed that over the same time interval in the experimental classes, about 40% of schoolchildren moved to a higher level of success, while the number and composition of students at each level in the second and third groups changed less significantly (Table 3).

Table 3: Results of Pedagogical Experiment

Level	After the experiment					
	Experimental group		Control group 1		Control group 2	
	Quantity	%	Quantity	%	Quantity	%
Initial	8	19	10	42,3	8	66,7
Medium	15	35,7	10	38,4	2	6,7
High	19	45,2	6	19,2	5	26,7

Statistical processing using the non-parametric fit method χ^2- - Pearson showed that the empirical values of Pearson's criterion in a pairwise comparison of the distributions of estimates in three samples turned out to be higher than the corresponding critical values for given sample sizes. When comparing the experimental sample and the first control sample, the following empirical value of the χ^2 --Pearson criterion was obtained χ^2_e =8.2. The corresponding critical value at $\rho \leq 0.05$ is significantly less than the empirical one (5.991). Similar results were obtained when comparing the experimental sample and the second control sample. Thus, after the use of the authors' model in the experimental group, students showed significant improvement in comparison with students in the two control group which used other approaches.

Due to the largely individual nature of work with gifted children, the generalized results of this diagnosis, in our opinion, cannot be considered an absolutely reliable indicator of the effectiveness of the study. Therefore, we also carried out an expert assessment of the study materials as an additional diagnostic technique. It was attended by 18 experienced mathematics teachers working in specialized mathematics classes. Their survey showed that the vast majority of teachers confirmed the feasibility and prospects of the proposed methodological solutions.

8. Conclusion

The problem of training gifted students is now becoming particularly relevant. Purposeful provision of such training involves the development of a number of methodological solutions relating, in particular, to the rational correlation in the educational process of the relevant content, methods and teaching resources.

When considering this, we analyzed regulatory and policy documents, programs in mathematics of basic and elective courses for specialized mathematical classes, scientific and scientific-methodical works of leading domestic and foreign experts in the field of developmental psychology, didactics, theory and methodology of mathematical education, as well as existing textbooks, teaching aids, methodological recommendations and software for educational purposes. In addition, a longitudinal observation was made of mathematics courses for gifted schoolchildren in a number of educational institutions and a survey was conducted of mathematics teachers working in specialized classes. The survey revealed the difficulties that arise when studying in these classes. As a result of this work, it was discovered that the majority of the teachers surveyed indicated, for the most part, that there is insufficient interaction of various formats of mathematical training of gifted students, which does not ensure its integrity.

The following relatively new results were obtained.

1. The authors' model of teaching gifted schoolchildren in mathematics has been developed and theoretically substantiated. The model includes components that reflect the content, forms and methods of working with gifted children in the framework of the main and optional courses.

2. The main methods of working with gifted children in mathematics classes which ensure their active search and research and creative activity are disclosed. The "mechanism of their interaction" in the educational process is also disclosed.

3. A holistic strategy for the work of a mathematics teacher with gifted children has been determined, encompassing classes within the framework of basic and elective mathematical courses, work on individual projects, participation in mathematical holidays and preparation for mathematical Olympiads. All of these formats are integrated into the individual educational routes of each of the students, ensuring the quality of their mathematical preparation and a high level of intellectual development.

4. Various options for constructing individual educational routes were identified and tested in the implementation of the developed strategy, depending on the stage of teaching mathematics. The implementation of these options formed the basis for the development of methodological materials and recommendations for the preparation of gifted students, taking

into account the continuity of basic and optional mathematical courses.

A pedagogical experiment to determine the possibilities of the proposed strategy for the training of gifted students based on a rational combination of various formats of such training. was carried out in the course of a two-year subject mathematical training of gifted schoolchildren in the Abulfazl Balami gymnasium for gifted children in the city of Vahdat (Republic of Tajikistan) and a number of educational organizations in Dushanbe. During the experiment, the educational and developmental capabilities of three models of mathematical training of gifted children were compared: one experimental group and two control groups, in which the relationship of related formats of teaching mathematics was not specifically taken into account. As a result of experimental training, a pairwise comparison of the degree of change in success in the selected groups of students showed that for the same time interval in the experimental classes, 40 % of the students moved to a higher level of success, while the number and composition of students at each level in the control groups changed less significantly.

In general, it can be concluded that empirical learning based on the methodological determinants and recommendations outlined above proved to be feasible and quite effective within the framework of the current regulatory formats. This is evidenced, in particular, by a fairly large number of schoolchildren participating in Olympiads of various levels, and high scores in examinations. This result was also confirmed by the results of an expert evaluation of the proposed methodological solutions by mathematics teachers working in specialized mathematical classes.

As a further development of our work, we are considering the development of an adaptive pedagogical technology for working with gifted students, which will contribute to the development of their mathematical abilities, as well as provide for effective learning to solve Olympiad problems. For teachers, this technology will help build an individual learning plan for each gifted student, naturally updating his cognitive activity.

References

[1] A. G. Balım, "The Effects of Discovery Learning on Students' Success and Inquiry Learning Skills," Eurasian Journal of Educational Research, Issue 35, 1-20 Spring 2009, 2009.

[2] L. Alfieri, P.L. Brooks, N.J. Aldrich, H. R. Tenenbaum, "Does discovery-based instruction enhance learning?" Journal of Educational Psychology, **103**(1), 1–18, 2011, doi.org/10.1037/a0021017.

[3] J. Gallagher, Teaching the gifted child (2nd ed.). Boston: Allyn and Bacon. 1975.

[4] S. Assouline, A. Lupkowski-Shoplik, Developing Math Talent. Texas: Prufrock Press, 2011.

[5] F. Barron, Creativity and the gifted. In New directions for gifted education. Report on bicentennial mid-year Leadership Training Institute. Los Angeles: National/State Leadership Training Institute on the Gifted and Talented, 1976, doi.org/10.1177/001698628002400306

[6] C.P. Benbow, L. L. Minor, "Cognitive profiles of verbally and mathematically precocious students: Implications for identification of the gifted," Gifted Child Quarterly, **34**(1), 21–26, 1990, doi.org/10.1177/001698629003400105.

[7] D. Bogoyavlenskaya, V. Shadrikov Working concept of giftedness. (2nd ed). M., Progress, 2003.

[8] G. Davis, S. Rimm, Education of the gifted and talented (5th ed.). Boston, MA: Allyn & Bacon, 2004.

[9] F. Gagné, "Transforming gifts into talents: The DMGT as a developmental theory," In N. Colangelo & G. A. Davis (Eds.), Handbook of gifted education (3rd ed., 60–74). Boston, MA: Allyn & Bacon, 2003.

[10] L. Vygotsky, Imagination and creativity in childhood. St. Petersburg, SOYUZ, 1997.

[11] M. Hoeflinger, "Developing mathematically promising students," Roeper Review, **20**(4), 244–247, 1998, doi.org/10.1080/02783199809553900.

[12] E. P.Torrance, Guiding creative talent. Englewood Cliffs, N. J.: Prentice-Hall, 1962.

[13] G. A. Goldin, "Mathematical creativity and giftedness: perspectives in response," ZDM 49, 147–157, 2017 doi.org/10.1007/s11858-017-0837-9.

[14] F. Barron, Creative Person and Creative Process, Holt, Rinehart & Winston, New York, 1969.

[15] J. Gilford, Three sides of the intellect. Psychology of thinking. M., Progress, 1965.

[16] C. W. Taylor, "Cultivating simultaneous student growth in both multiple creative talents and knowledge," In J. S. Renzulli (Ed.), Systems and models for developing programs for the gifted and talented, 306–351, 1986.

[17] E. P. Torrance, "Growing up creatively gifted: A22-year longitudinal study," Creative Child and Adult Quarterly, **5**(3), 148–158, 1980.

[18] E. Torrance, J. Khatena, "Originality of imagery in identifying creative talent in music," Gifted Child Quarterly, **13**, 3-8, 1969, doi.org/10.1177/001698626901300101.

[19] C. Lombroso, The Man of Genius. London: W. Scott, 1891.

[20] J. Guilford, The nature of intelligence. New York: McGraw-Hill, 1967.

[21] S. Rubinshtein, Fundamentals of general psychology. M, Uchpedgiz. 1940.

[22] A. Karp, "Knowledge as a manifestation of talent: Creating opportunities for the gifted," In B. Sriraman (Ed.), Creativity, giftedness, and talent development in mathematics (209–224). Charlotte, NC: Information Age Publishing, 2008, https://www.diva-portal.org/smash/get/diva2:1390686/FULLTEXT01.pdf.

[23] T. Hirano, "Achieving mathematical excellence in Japan: Results and implications," Journal of Educational Research, **25**(6), 545-551, 1996, doi.org/10.1016/S0883-0355(97)86731-6.

[24] K. Heller, A. Lengfelder, German Olympiad study on math, physics and chemistry. Paper presented at the American Educational Research Association, New Orleans, 2000, doi.org/10.1080/02783193.2011.530202.

[25] C. E. Hmelo-Silver, H. S. Barrows, Goals and Strategies of a Problem-based Learning Facilitator. Interdisciplinary Journal of Problem-Based Learning, **1**(1), 2006, doi.org/10.7771/1541-5015.1004.

[26] R. Campbell, H. Walberg, "Olympiad Studies: Competitions Provide Alternatives to Developing Talents That Serve National Interests," Roeper Review, **33**(1), 8-17, 2011, doi.org/10.1080/13803610701785949

[27] Y. Terada, Boosting student engagement through project-based learning. Edutopia, 2018.

[28] K. Tirri, "Finland Olympiad Studies: What factors contribute to the development of academic talent in Finland," Journal of NACE., **5**(2), 56-66, 2001.

[29] J. Boaler, "Promoting 'relational equity' and high mathematics achievement through an innovative mixed-ability approach," Br. Educ. Res. J. 34, 167–194, 2008 doi.org/10.1080/01411920701532145.

[30] D. P. Wolf, The art of questioning. Academic Connections, 1987.

[31] A. Benson, D. Blackman, "Can research methods ever be interesting?", Active Learning in Higher Education **4**(1), 39-55, 2003, doi.org/10.1177/1469787403004001859.

[32] C. H. Chen, Y. C. Yong, "Revisiting the effects of project-based learning on students' academic achievement: A meta-analysis investigating moderators," Educational Research Review, **26**, 71–81, 2019, https://www.learntechlib.org/p/207141/.

[33] S. Cho, H. Lee, "Korean gifted girls and boys: What influenced them to be Olympians and non Olympians," Journal of Research in Education, 12(1), 106–111, 2002.

[34] D. Kokotsaki, V. Menzies, A. Wiggins, "Project-based learning: a review of the literature," Improving schools., **19**(3). pp. 267-277, 2016, doi.org/10.1177/1365480216659733.

[35] A. Mettas, C. Constantinou, "The Technology Fair: a project-based learning approach for enhancing problem solving skills and interest in design and technology education," International Journal of Technology and Design Education, **18**, 79-100, 2007, doi.org/10.1007/s10798-006-9011-3.

[36] J. Brunstein, Achievement motivation (H. Heckhausen, Ed.). In J. Heckhausen & H. Heckhausen (Eds.), Motivation and action (137–183). Cambridge University Press, 2008,

doi.org/10.1017/CBO9780511499821.007
[37] A. Conley, "Patterns of motivation beliefs: combining achievement goal and expectancy-value perspectives," Educ. Psychol. **104**, 32–47, 2012, doi.org/10.1037/a0026042
[38] C. S. Dweck, Self-theories: Their role in motivation, personality and development. Philadelphia, Psychology Press, 1999.
[39] S. K.W. Chu, S. K. Tse, K. Chow, "Using collaborative teaching and inquiry project-based learning to help primary school students develop information literacy and information skills," Library & Information Science Research, 33, 132-143, 2011, doi.org/10.1016/J.LISR.2010.07.017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 144-153 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Design and Comparative Analysis of Hybrid Energy Systems for Grid-Connected and Standalone Applications in Tunisia: Case Study of Audiovisual Chain

Saidi Mohamed[1,2,*], Habib Cherif[1, 2], Othman Hasnaoui[1, 2], Jamel Belhadj[1, 2]

[1]*Electrical Systems Laboratory (LSE-LR-11ES15)-ENIT, University of Tunis el Manar, Tunis, 1002, Tunisia*

[2]*Université de Tunis, ENSIT, BP 56 Montfleury, 1008, Tunisia*

ARTICLE INFO

Article history:
Received: 06 February, 2023
Accepted: 13 May, 2023
Online: 12 June, 2023

Keywords:
Audiovisual chain
Net Present Cost
Micro-grid
Optimization
Renewable energy

ABSTRACT

In this research paper, a technical and economic-environmental study was developed to investigate the possibility of establishing various hybrid power systems with different operation modes. Grid-connected and standalone hybrid systems (solar-wind with storage batteries and diesel generators) have been realized in order to carry out a comparative analysis study of two configurations. These systems have been investigated through the Hybrid Multi-Energy Resource Optimization software (HOMER), which calculates pollutant gas emissions, simulates and optimizes energy consumption based on energy demand and resources. As part of the economic analysis, the internal rate of return, the net present value and the payback period were estimated. Both configurations have been developed to meet the power consumption of an audiovisual system. The results obtained show that the first system is the most cost-effective to establish, considering in particular the energy production potential and gas emissions. It can be stated that the proffered grid-connected hybrid system is the most suitable and cost-effective system as it offers several advantage. The total net present cost is $5.425million and the total energy cost is approximately $0.0686 per unit.

1. Introduction

Today the most serious problems of the world are the decrease of fossil fuel reserves, and their elevated price, it is therefore necessary to consider a solution to reduce the use of non-renewable resources and to use sustainable alternative energies like solar and wind energy. Renewable energies have progressively substituted nuclear energy and fossil fuels in four different sectors: power generation, heating plants, transports and stand-alone energy production [1]. This is the case of photovoltaic (PV) and wind energy, worldwide, this sector is growing rapidly [2,3], in particular because of the increasing competition of renewable energies, the rise of electricity demand in developing countries and the benefits of this type of energy in terms of pollution reduction. In this regard, it is important to note that gas, oil and coal are always highly popular fuels for power generation. The effects of these energy sources, as well as the increase in the world's population and its energy consumption, have had a negative impact on the environment [4,5]. In addition, among the objectives of the Tunisian state is to limit the impact of environmental damage by exploiting the potential of available renewable resources, which will help manage the risk. Considering the previous context, in previous years a growing number of studies have been conducted on hybrid grid-connected and standalone generation systems, by mixing different energy resources such as wind, solar, hydro and generators [6–8]. For this reason, a design and a technical-economic and environmental study has been carried out in the development of various systems of solar-wind hybrid systems that incorporate additional generation sources such as diesel generators. To evaluate the proposed design [9–11], the energy consumption of the building of an audiovisual channel in Tunisia was used as a case study. The novelty of this work is the incorporation of the ecological component as well as the technical and economic aspects in order to achieve a consistent analysis [12], which facilitates decision for the development of energy production strategies based on hybrid systems.

As a developing country, Tunisia has energy supply problems to support its economic growth. 30% of energy should be renewable by 2030, according to the national energy strategy. The development of multi-source energy systems for commercial enterprises is crucial, including energy sources such as photovoltaic, wind [13–15]. Tunis is the capital city, geographically located in the north of Tunisia, the coordinates of its geographical site are 36°49.43'north10°09.27'east (Figure 1).

*Corresponding Author: Saidi Mohamed, Email msaidi31@yahoo.fr , Tel 0021698924458

Therefore, this site is a good place to design a multi-source renewable energy system, especially for an audiovisual television channel with an available surface area of about 10000 m^2. In this context, this paper presents a comparative study of two modes micro-renewable energy network [7], in order to choose the right solution to reduce the cost of energy consumption for an audiovisual chain and ensure continuous and regular electricity production.

The paper is based on real project in the Tunisian television, in order to implant a new smart-grid for the Tunisian television, which is a high consumer of energy every year.

Our main objective is to define the optimal sizing based on various configurations, and minimize cost parameters such as, total net present cost (TNPC), cost of energy (COE), and unmet electrical load and CO2 emissions [10] using HOMER software.

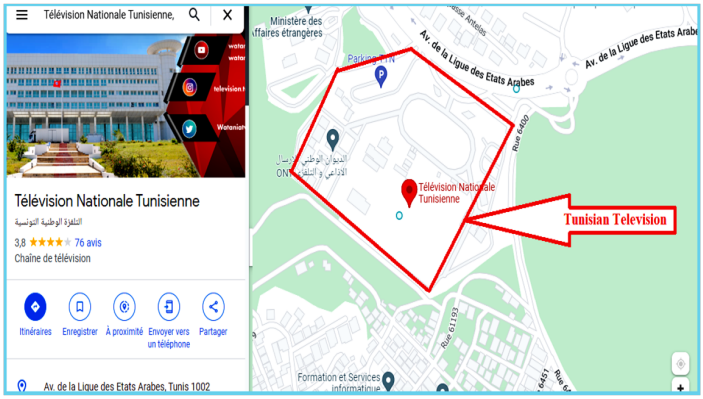


Figure 1: Geographic coordinate of studied area

2. Current Power Supply Architecture

Figure 2 shows that the television is supplied from a medium voltage network (10kV) through six transformers (5x1250kVA and 1x800kVA) connected in parallel; two of them are stand-by. Three diesel generators (3x1390kVA) are used to provide emergency power for certain priority equipment. It is very important to ensure the continuity of broadcasting; in this context four (online) inverters (500kVA and 250 kVA) with a park of batteries (12V and 6V) are used [11].

In this study, two types of loads are defined [8]:

- Sensitive loads, also called critical loads (control rooms, broadcasting center, studios, stage lighting)
- Other/normal loads (air conditioner, offices, lighting, pumps)

3. Methods and Materials

In this research, two configurations of hybrid micro-grid on-grid and standalone were designed and compared in order to find the optimum for an audiovisual chain located in the north of Tunisia (36°49.43'N, 10°09.27'E). The objective of this micro-grid design is to reduce the cost of the company's electricity consumption and to provide energy supply continuity.

3.1. Introduction to HOMER-pro software

The HOMER-Pro software was developed by NREL, USA. It facilitates the process of designing the most economical microgrid in a distributed energy system. It gives precise and independent results. It also investigates all possible configurations and determines the lowest cost solution by combining several system elements and a storage technology.

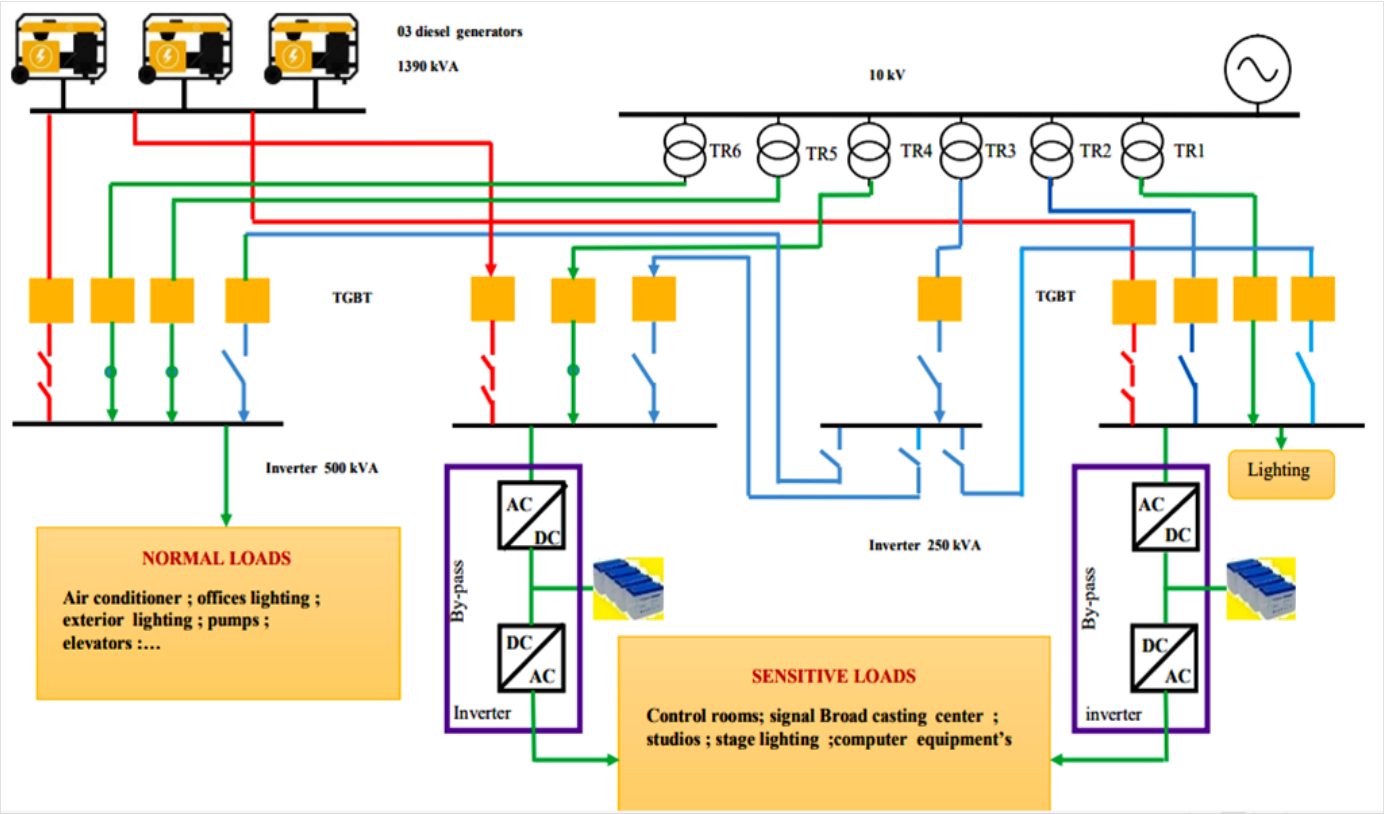


Figure 2: Micro-grid of the power supply and electrical distribution system

Different metrological data are needed to design the hybrid system, in the same way as to describe the possibilities. The workflow in this software is in 3 steps. The process starts with the project input data which includes the load profile, site specific resources and system components, in step 2 HOMER-Pro analyzes the simulation, optimization and sensitive parameters, in step 3 it shows the result which delivers detailed information about the system sizing, performance and financial parameters.

The following Figure 3 presents the diagram of the proposed hybrid energy systems design methodology.

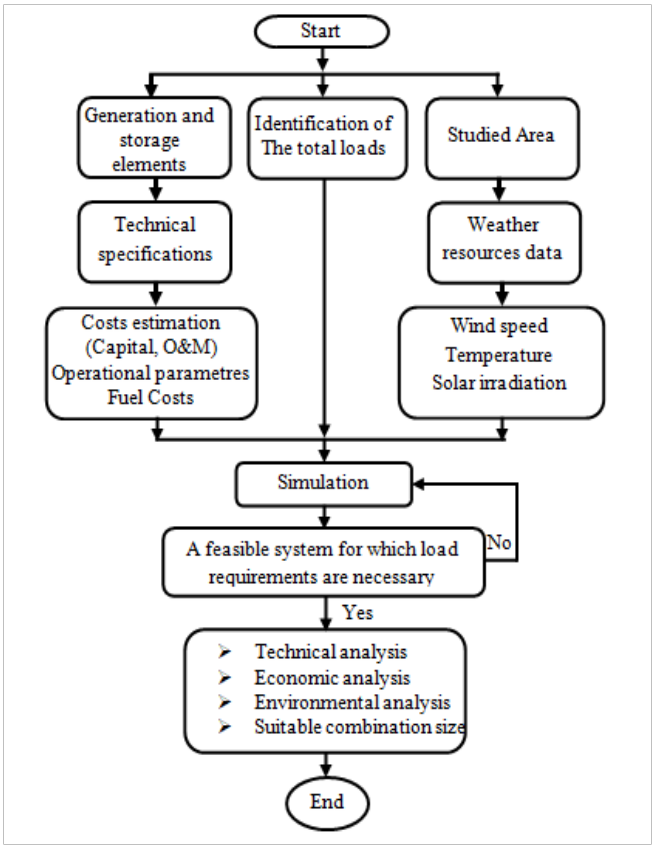


Figure 3: The diagram of the proposed hybrid energy systems design methodology.

In the simulation process, during simulation we define the various system combinations, the number of components needed, the sizes and the dispatch strategy. HOMER-Pro is able to model many system configurations by taking into account any combination of PV panels, hydrogen tanks, wind turbines, hydroelectricity, an electrolyzer, an AC-DC converter and a battery storage. The designed system can be grid-connected or off-grid and can supply different electrical loads. HOMER-Pro analyzes two important aspects: firstly, it evaluates the functionality of the system design and secondly, it determines the life cycle cost of the design, which is the total cost of installation and operation.

The optimization phase defines the most ideal combination, as well as the best configuration that meets the load requirements. In this process, HOMER-Pro simulates different varieties of system configurations, removing those that are not feasible. The feasible configurations are organized according to the lowest net present cost (NPC) and energy cost (COE). This software analyzes different variables to obtain the most feasible configuration for the desired load [16,17]. The decision variables that are analyzed include: generator size, PV module, AC to DC converter, electrolyzer, hydrogen storage tank, number of wind turbines, number of batteries and dispatch strategy.

The software adopts assumptions that affect the design of the system. These assumptions are called sensitivity variables. Sensitivity variables include solar radiation, diesel cost, wind speed, interest rate, grid price, etc. The software finds many combinations of systems which are feasible under some conditions. A sensitivity analysis illustrates the effect of changing inputs on the results [18]. The software is able to analyze multiple sensitive values at once to find the economically suitable result. The software user can include as much sensitivity as desired for the needs of the analysis.

3.2. Energy Potential

In this paragraph, we analyze the wind, solar and ambient temperature resources available in the project area. The ambient temperature and wind values are obtained from the National Institute of Meteorology of Tunisia. The solar radiation is obtained using the HOMER (Hybrid Optimization of Multiple Energy Resources) simulation tool, which is linked to the National Aeronautics and Space Administration (NASA) database.

First, Figure 4 shows the monthly average ambient temperature. The lowest value is indicated in January with 14.5°C and the highest in August with 31.5°C. It is essential to take into consideration the environmental temperature, since it affects the photovoltaic modules' efficiency and performance, which is best achieved at an operating temperature of 25°C.

Figure 5 shows the average monthly solar radiation of the position of the Tunisian TV channel, where we observe that the minimum radiation is presented in december with 2.09 kWh/m²/day and the maximum radiation is presented in July with 7.31 kWh/m²/day.

Figure 6 presents the monthly average wind speed, where it can be seen that the minimum speed is presented in August with 4.950 m/s and the maximum speed is presented in February with 7.07 m/s.

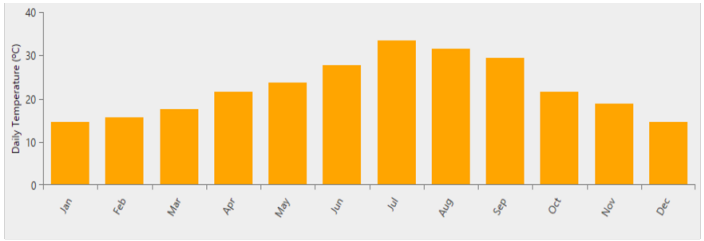


Figure 4: The monthly average of the ambient temperature

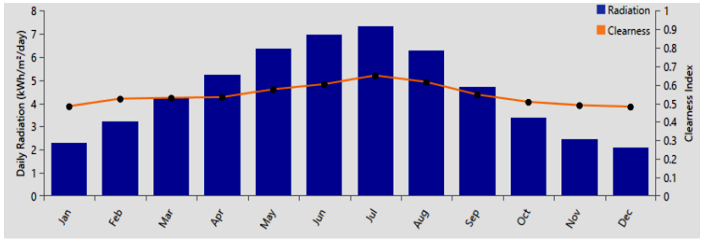


Figure 5: The monthly solar radiation

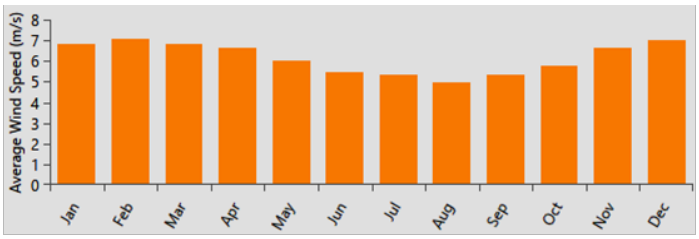


Figure 6: The monthly average wind speed

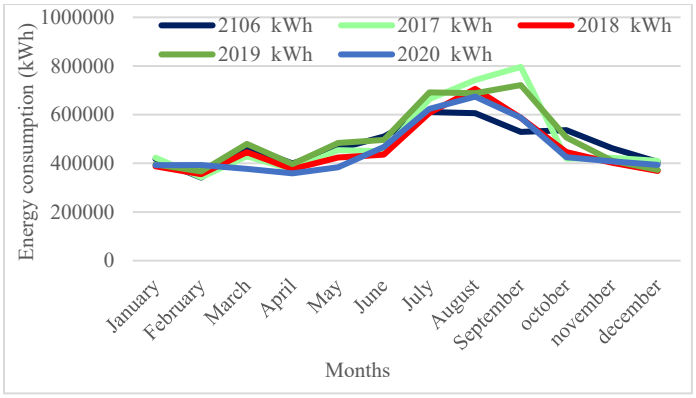


Figure 7: Monthly energy consumption in recent years

3.3. Load profile

Based on Figure 7 and Figure 8, in Tunisian television the average monthly consumption in the years (2016-2017-2018-2019-2020) is a function of production emissions during three distinct seasons: winter, summer and the month of Ramadan. Monthly consumption is almost the same during the months of

October to May (an average consumption of 400,000 kWh per month), then a slight increase (20%) in the month of Ramadan (June) and finally three months of high consumption (June-July-August), about 35% of total annual consumption (air conditioning). It was seen that the peak is about 1442 kW and with an average of 626 kW [7], [10].

The investigated area is located in the north of Tunisia; its geographical site co-ordinate is located at 36°49.43'north10°09.27'east). Consequently, this site is an appropriate place to design a grid connected multisource energy renewable system, especially for an audiovisual television chain.

The options of hybrid systems proposed in this work were designed for energy consumption in broadcasting building of the Tunisian television during 2020. The two hybrid systems were designed for a load of 15000kWh/d. The electrical load profile is shown in Figure 8.

To study the performance of the system over the course of a year, HOMER uses the daily load profile illustrated in Figure 9. The load profile is variable on various days and times due to seasonal and regional time changes for more efficient electricity consumption. The load variations during seasonal changes, and the minimum and maximum loads recorded are illustrated.

4. System description and simulation models

Using HOMER, two models of hybrid system have been established: grid-connected hybrid system and off-grid hybrid system.

Figure 10 illustrates the concept of a grid-connected hybrid system, which includes wind turbines, photovoltaic panels, loads, and inverter, with a diesel generator and a park of batteries that are used to maintain the system in its off-grid state. Figure 11 shows the same concept, but a standalone configuration.

The HOMER simulation software gives the possibility of entering the various electrical components that will have to be used for the device to be effective and which automatically carry out all the possible configurations, taking into account only those that satisfy the required load.

To maximize system performance in various situations, HOMER simulates the configurations listed below with the same loads in the same region based on various costs such as estimated installation cost, operation and maintenance cost, replacement cost, interest rates and energy cost, as well as the analysis of pollutant emissions.

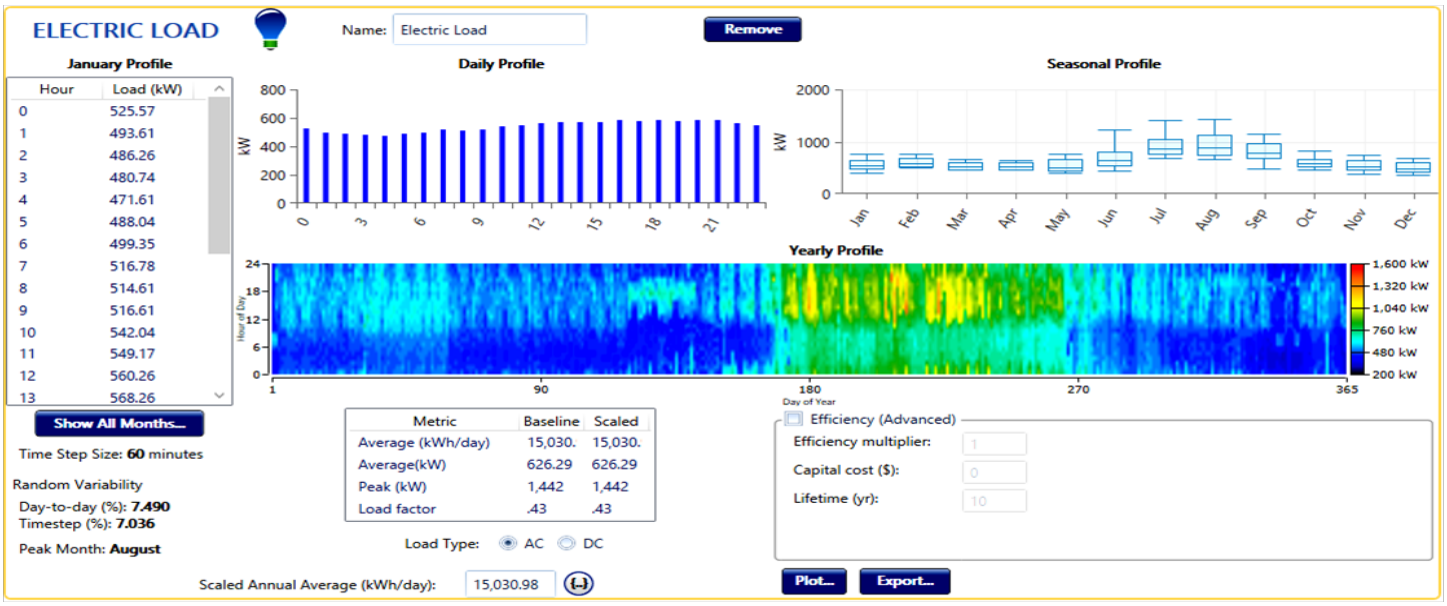


Figure 8: Load profile of audiovisual chain.

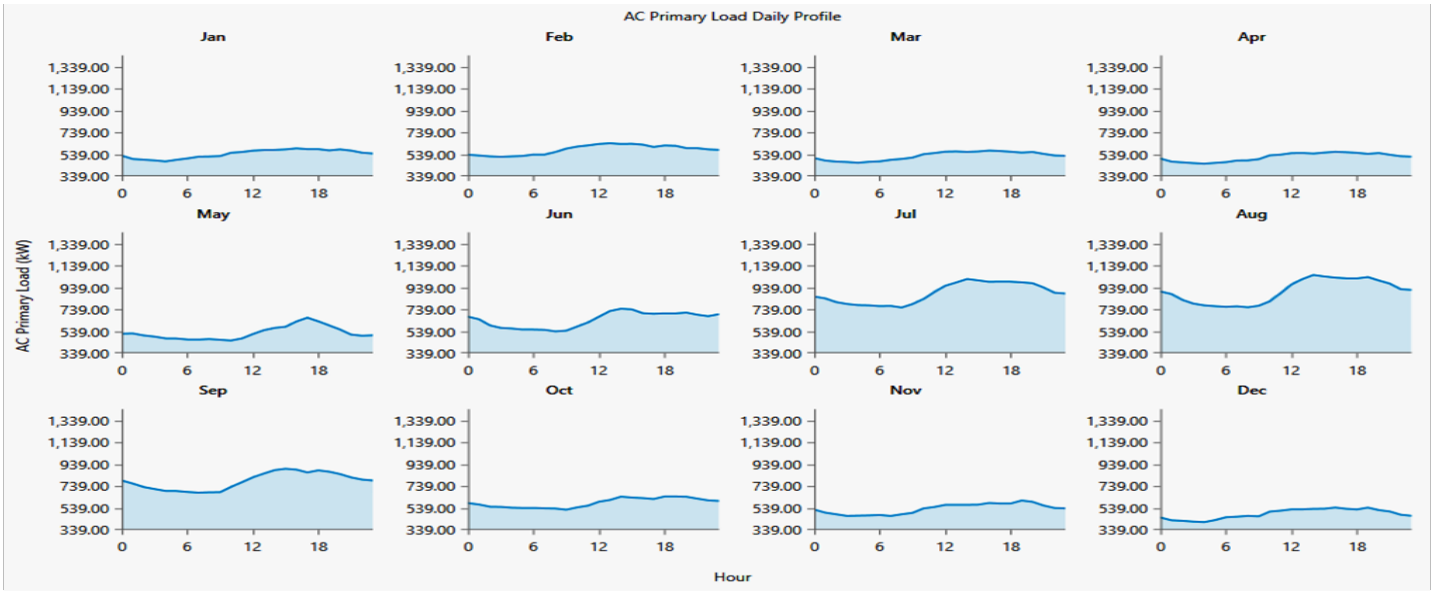


Figure 9: load profile of the studied area

The Figures 10 and 11 are the studied models in this paper which are developed to replace the current power supply configuration illustrated in Figure 2, with the integration of new renewable energy sources.

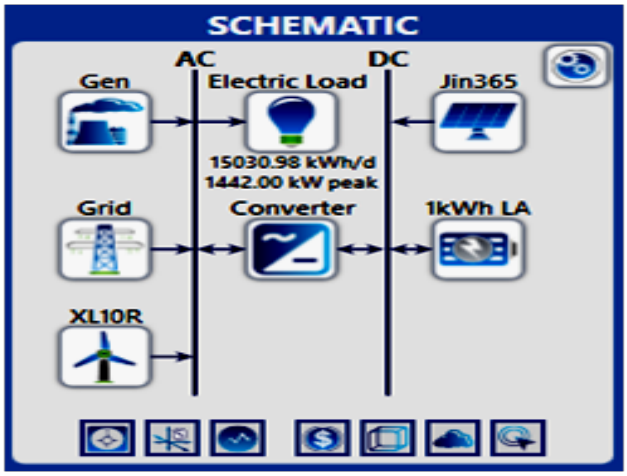


Figure 10: Grid-connected Hybrid System

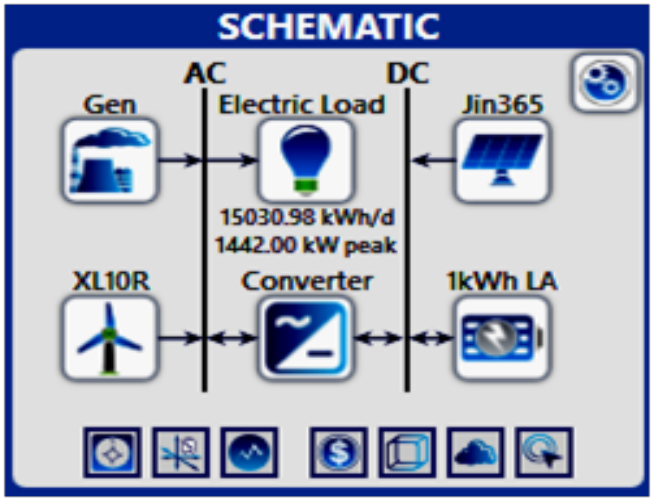


Figure 11: The Standalone Hybrid System.

So, our main task is to design and investigate a new smart-grid based on renewable energy for energy supply of the Tunisian television.

4.1. Wind Turbine

Wind turbine converts the kinetic energy of the wind into AC or DC electricity for particular power curve. The power curve is a graph between power output and wind speed.

Bergey wind power's Excel 10-R model with hub height (30 m) is considered. It has a rated capacity of 10 kW and provides (AC) voltage as an output. The cost of one unit is considered ($15000) while replacement and maintenance costs are taken as ($13500) and ($150/year)

$$Pwt= 0.5\rho V^3 S\ \eta_l C_p \quad (1)$$

where:

V: the wind speed in m/s,

S: rotor swept area in m^2,

η_l: generator efficiency,

C_p: maximum power coefficient,

ρ: air density in kg/m^3,

4.2. Solar PV panel

HOMER-Pro develops a PV module that produces DC electricity when the solar radiation incident upon it. It is best to choose a solar panel when the price of diesel fuel is high and wind speeds are low. The capital cost of a 1kW solar panel is approximately$600; the replacement and operating costs are $600 and $10, respectively. The service life of the PV system is 25 years.

$$P_{PV} = C_{PV} \times P_{PV} \times (I_T/I_S) \quad (2)$$

where,

C_{PV}: PV rating factor.

P_{PV} : PV array rated capacity (kW).

I_T: Solar radiation values that strikes on the surface of the PV Array (kW/m^2).

I_S: Standard radiation value, 1kW/m$^{2.}$

4.3. Diesel Generator project

Diesel generator is used as alternative power sources when there is no access to the grid or interruption of electrical power. Diesel generator is also preferred as backup power, ensuring a robust energy supply. Generator cost is determined for 1kW output power. The capital cost is $ 300, the replacement cost is 270$, O&M cost is 0.03($/op. hour) and the lifetime is 15000 hours [4,5]. The fuel consumption per hour of a diesel-fueled diesel generator is calculated by Equation 3.

$$F=F_0 \times Y_{gen}+F_1 \times P_{gen} \quad (3)$$

where,

F: the fuel consumption rate (L/hr.)

F_o: the fuel curve intercept coefficient (L/hr/kW),

F_1: the fuel curve slope (L/hr/kW),

Y_{gen}: the rated capacity of the generator (kW),

4.4. Battery Storage Bank

For a certain hour, the surplus power produced by the hybrid system can be used to charge the batteries, while the stored energy can be discharged at any time when there is a power deficit.

The battery considered is a generic 12v lead acid battery with 1kwh of energy storage. The estimated lifetime is (5 years) and the cost of one battery is ($250) with a replacement cost of ($250) while the maintenance cost is estimated at ($5/year).

$$Cwh = (E_L \times AD) / (\eta_{inv} \times \eta_{bat} \times DOD) \quad (4)$$

where,

E_L is the average daily load energy (kWh/day),

AD is daily autonomy of the battery,

DOD is battery depth of discharge,

η_{inv} and η_{batt} respectively, represent the inverter and battery efficiency.

4.5. Converter

The energy feed between the AC bus and the DC bus within the microgrid is realized by a bidirectional power converter according to the production, consumption, and storage energy conditions of the micro grid. Converter cost is determined for 1kW output power. The capital cost is 200$, the replacement cost is 180$, O&M cost is 5$ and the life expectancy is 15 years [5–7]. The efficiency of the converter connected to two bus bars is taken as 95%.The inverter changes the DC current at the output of the photovoltaic system and battery to supply the electrical load, which by its nature is alternating current. The design is kept to supply the load 24 hours a day, 7 days a week without any disruption in the power supply due to the uncertain nature of solar and wind generation.

5. Economic analysis

Net Present Cost (NPC): The NPC is the cost of installing and operating of the system over its lifetime which is calculated with the following formula [18,19].

$$NPC= T_{Ann,\ cost}/ CRF\ (j, R_{Project}) \quad (5)$$

$T_{Ann,\ cost}$: It is the sum of the annualized costs of each component of the power system, including capital, operating and maintenance costs. It also includes the replacement cost and the cost of fuel.

j: Interest rate in percentage

$R_{project}$: lifetime in year

CRF: Capital recovery factor is a ratio which is used to calculate the present value of a series of equal annual cash flow.

$$CRF = j \times (1+j)^n / (1+j)^n - 1 \quad (6)$$

where,

j: Interest rate in percentage
n: number of years

The cost of energy is calculated by the following formula

$$COE= T_{Ann,\ cost} / (E_{primary}+E_{def}+E_{grid.sell}) \quad (7)$$

$T_{Ann,\ cost}$: Total annualized cost
$E_{primary}$: No of primary load
E_{def}: No of deferrable load

6. Optimization and Simulation Results

The supply of a load of 15,030.98kWh/d with a peak value of 1442kW was studied in 2 scenarios. In the first scenario, the power demand was provided by a grid-connected microgrid and in the second scenario by standalone microgrid. For both scenarios, PV, wind turbines, diesel generator and a battery park were used.

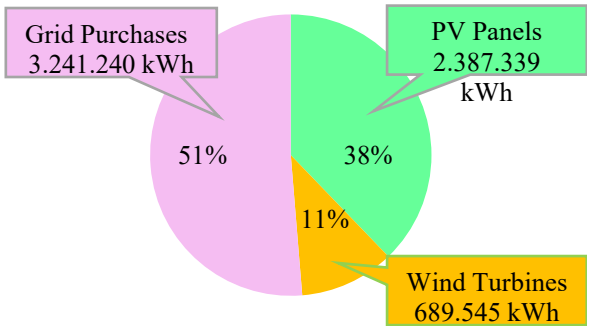


Figure 12: Energy Production summary

6.1. Grid- connected configuration

In this part, the simulation results of grid-connected installations are calculated for different configurations using HOMER. The main energy source is the photovoltaic. In case of lack of energy from PV panels and wind turbines, the electrical

Table 1: Categorized optimization results

Architecture						Cost				
PV (kW)	Wind turbine	Diesel generator (kW)	Batteries	Grid (kW)	Inverter (kW)	NPC ($)	COE ($)	Operating Cost ($/yr)	Initial capital ($)	Renewable fraction (%)
1 ,667	30		300	999.999	1000	5.43M	0.0686	285,183	1.74M	47
1,708			300	999.999	1000	5.47M	0.0712	321,218	1.32M	37.6
1 ,667	30	1600	300	999.999	1000	5.81M	0.0734	277,318	2.22M	47.0
1,708		1600	300	999.999	1000	5.85M	0.0761	313,413	1.80M	37.6
	30		300	999.999	1000	6.15M	0.0867	418,215	740,000	12.6
			300	999.999	1000	6.30M	0.0888	464,519	290,000	0.00372
	30	1600	300	999.999	1000	6.53M	0.0920	410,410	1.2M	12.6
		1600	300	999.999	1000	6.67M	0.0941	456,714	770,000	0.0372

grid, the battery storage system and the diesel generator are used as auxiliary energy sources.

Monthly average electrical power generatedby each of the hybrid system elements is illustrated in figure12, it is clear that the most important quantity of energy is givenfrom the grid (51.3 %) whereasphotovoltaic generate only 37.8% of overal energy and wind turbine produce about 10.9%. The photovoltaic energy production is 2.387.339 kWh/yr and the wind turbines energy production is 689.545kWh/yr.

As shown in Table 1, in current simulation of the designed micro-grid, the optimized results demonstrates that the cost of energy have the minimum value COE of $0.0686/kWh.

Table 2 indicates that pollutant emissions like carbon dioxide, carbon monoxide, and sulfur dioxide are significantly decreased by using the grid-connected hybrid system.

Table 2: Quantity of emission produces by different pollutants

Quantity	Value (kg/yr)
Carbon Dioxide	2.055.538
Carbon Monoxide	0
Unburned Hydrocarbons	0
Particulate Matter	0
Sulfur Dioxide	8.912
Nitrogen Oxides	4.358

According to Table 3, 89.7% of the electricity produced by the renewable system is sent to the AC grid and 10.3% is sold to the grid at specified selling price.

Table 3: Energy consumption summary

component	Consumption kWh/yr	Percentage
AC Primary Load	5.486.307	89.7
Grid Sales	633.321	10.3
Total	6.119.628	100

According to Figure 13 in this micro grid, 48.7% of the total energy is generated from renewable sources. All of this generation is delivered right into the load. The proportion of renewable sources energy directly used by the load is 47%. Thus, 1.16% of the total energy, generation during a year are excess energy confirming to the results of simulation.

Figure 14 illustrates the results of the microgrid evaluation in terms of capital costs and operation and maintenance costs. The PV array system represents the largest capital cost, therefore the grid represents the highest operation and maintenance cost Based on this estimation, the hybrid project will have approximately a $5.425 million cost over its operating lifetime In Figure 15, we observe the cumulative cash flow of this system over the life of the project, as shown in the curve over 25 years, there are significant changes between the initial system and. the proposed system, there is a large decrease in term of cash flow of the final system.

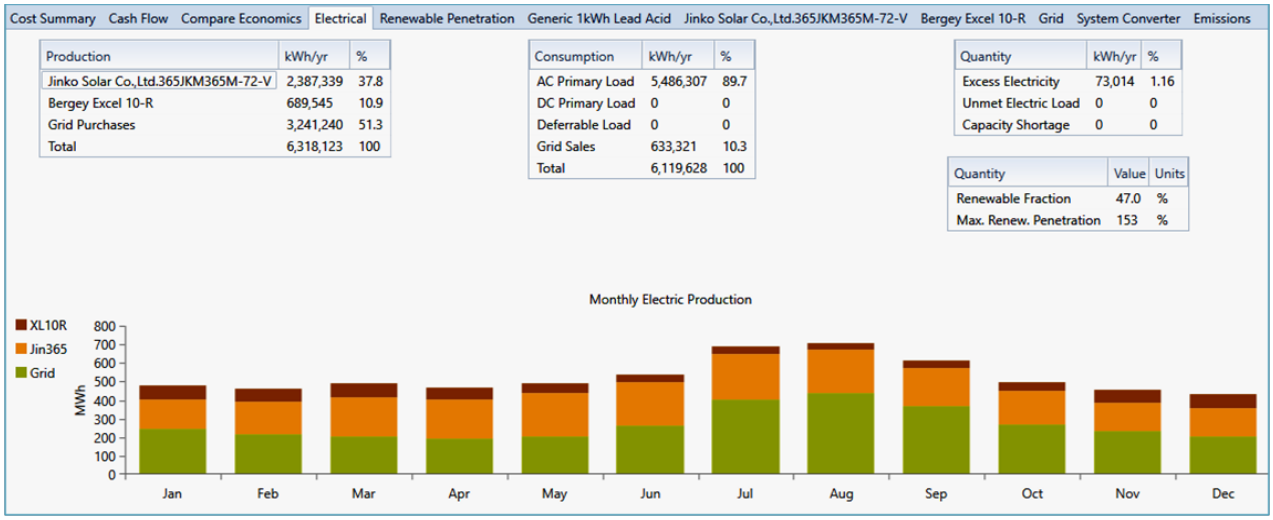
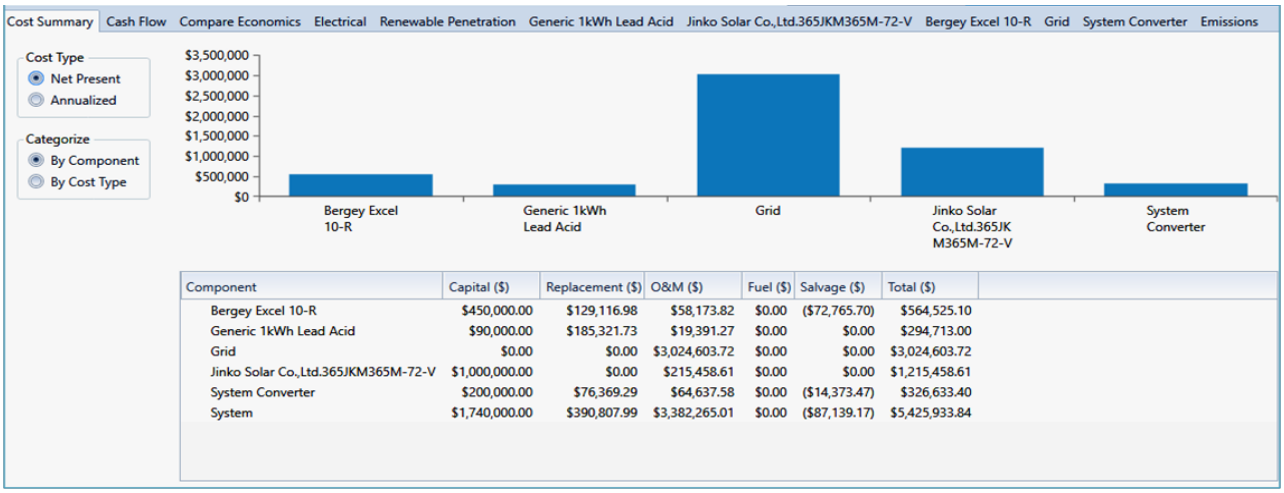


Figure 13: Production and consumption of electricity



Component	Capital ($)	Replacement ($)	O&M ($)	Fuel ($)	Salvage ($)	Total ($)
Bergey Excel 10-R	$450,000.00	$129,116.98	$58,173.82	$0.00	($72,765.70)	$564,525.10
Generic 1kWh Lead Acid	$90,000.00	$185,321.73	$19,391.27	$0.00	$0.00	$294,713.00
Grid	$0.00	$0.00	$3,024,603.72	$0.00	$0.00	$3,024,603.72
Jinko Solar Co.,Ltd.365JKM365M-72-V	$1,000,000.00	$0.00	$215,458.61	$0.00	$0.00	$1,215,458.61
System Converter	$200,000.00	$76,369.29	$64,637.58	$0.00	($14,373.47)	$326,633.40
System	$1,740,000.00	$390,807.99	$3,382,265.01	$0.00	($87,139.17)	$5,425,933.84

Figure 14: Cost status of the system

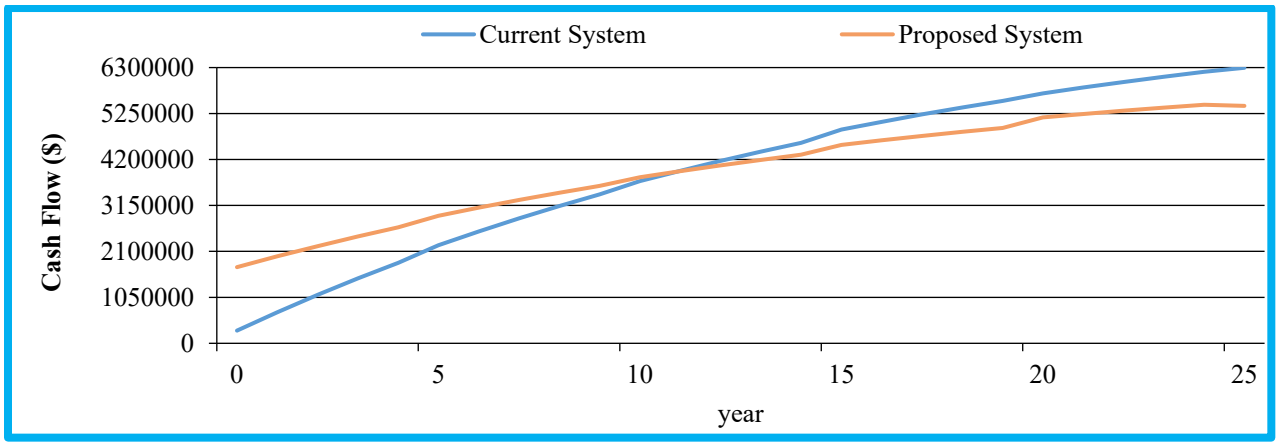


Figure 15: Cumulative cash flow over project lifetime

6.2. Off-Grid configuration

In this section, the simulation results of the different stand-alone combinations are calculated for the various configurations using HOMER. The two main energy sources are photovoltaic and wind turbine. In case of lack of energy from the photovoltaic panels and wind turbines, the diesel generator is used as a complementary energy source, as well as the battery storage system which is used in case of lack of power to ensure the continuity of the TV chain.

Monthly average electrical power generatedby each of the hybrid system elements is illustrated in figure16, it is clear that the most important quantity of energy is givenfrom the diesel generator (49.2 %) whereasphotovoltaic generate only 30.5% of overal energy and wind turbine produce about 20.3%. The photovoltaic energy production is 2.283.006 kWh/yr, the wind turbines energy production is 1.522.149kWh/yr and an energy production of 3.688.661 kWh/yr from the diesel generator.

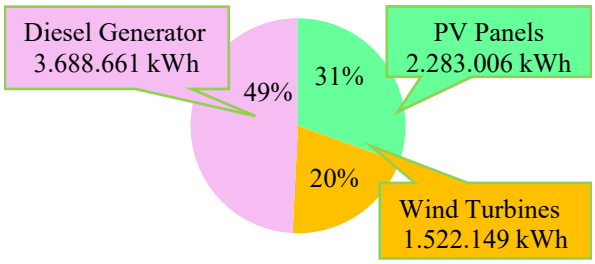


Figure16: Energy Production summary

As shown in Table 4, in current simulation of the designed micro-grid, the optimized results demonstrates that 1505kW of photovoltaic panels and 58 wind turbines of 10kWare used the cost of energy COE is $0.258/kWh.

Table 4: Categorized optimization results

Architecture					Cost				
PV (kW)	Wind turbine	Diesel generator (kW)	Batteries	Inverter (kW)	NPC ($)	COE ($)	Operating Cost ($/yr)	Initial capital ($)	Renewable fraction (%)
1,505	58	1,600	1,200	500	18.3M	0.258	1.21M	2.68M	32.8
270	40	1.600		160	18.4M	0.260	1.33M	1.27M	15.9
	40	1.600			18.5M	0.261	1.35M	1.08M	12.2
	40	1.600	4	1.30	18.5M	0.261	1.35M	1.08M	12.2

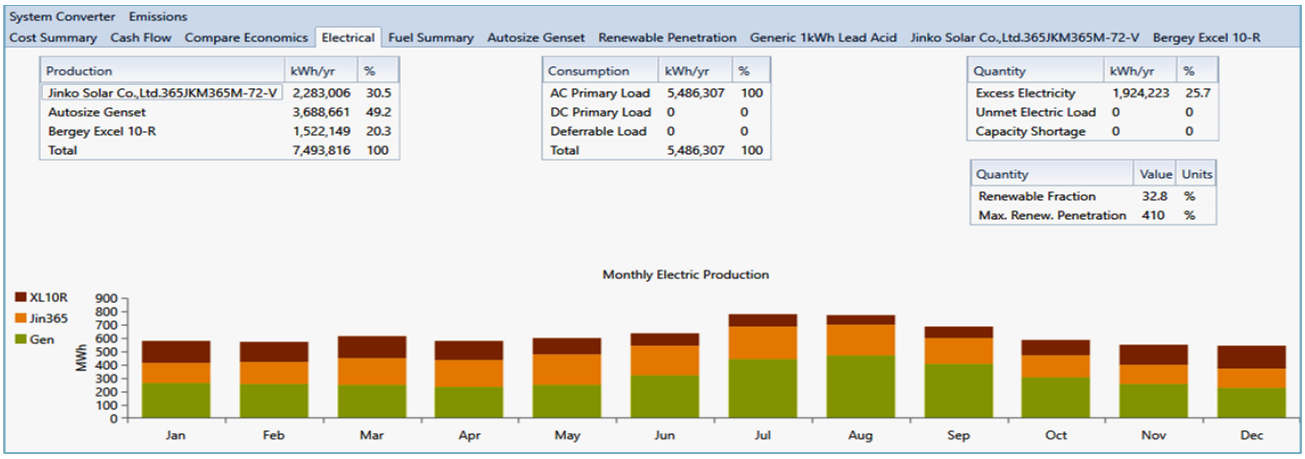


Figure 17: Production and consumption of electricity

In this microgrid, 50.8% of the total energy is produced from renewable sources. All of this generation is delivered directly to the load. According to Figure 17, the proportion of energy from renewable sources directly used by the load is 32.8%. Thus, 100% of the electricity generated by the renewable system is sent to the AC grid and 25.7% of the total energy production during a year is surplus energy confirming the simulation results.

Table 5 illustrates that pollutant emissions such as carbon dioxide, carbon monoxide, unburned hydrocarbons, nitrogen oxides and sulfur dioxide are significantly increased by the use of the stand-alone hybrid system.

Table 5: Quantity of emission produces by different pollutants

Quantity	Value (kg/yr)
Carbon Dioxide	2.779.450
Carbon Monoxide	17.646
Unburned Hydrocarbons	770
Particulate Matter	107
Sulfur Dioxide	6.855
Nitrogen Oxides	16.577

Figure 18 demonstrates the cost status of the off grid hybrid system during its project lifetime. Table 4 shows the results of the evaluation of the isolated micro grid in terms of investment costs and operation and maintenance costs. The renewable energy system (photovoltaic and wind) represents the high investment cost, therefore the diesel generator represents the highest operation and maintenance cost. Based on this estimate, the hybrid project will have a lifetime cost of approximately $18.30 million.

In Figure 19 we can see the cumulative cash flow of this system over the project live time, as seen on the curve during 25 years there is no significant change between the based system and the proposed system.

7. Conclusion

This working study is a comparative study between off-grid and a grid-connected hybrid electrical system with different modes of operations for an audiovisual chain. The results obtained with the HOMER-pro software show a detailed cost analysis structure, cash flow overview, and energy production yield of the proposed hybrid configuration.

An analysis of the situation was carried out in the Tunisian television chain. The study was conducted taking into account energy use, climatic conditions, current prices for all components and accessible areas within the study area. The optimization results show that the on-grid hybrid system is more efficient and cost-effective than the off-grid hybrid system with the same load.

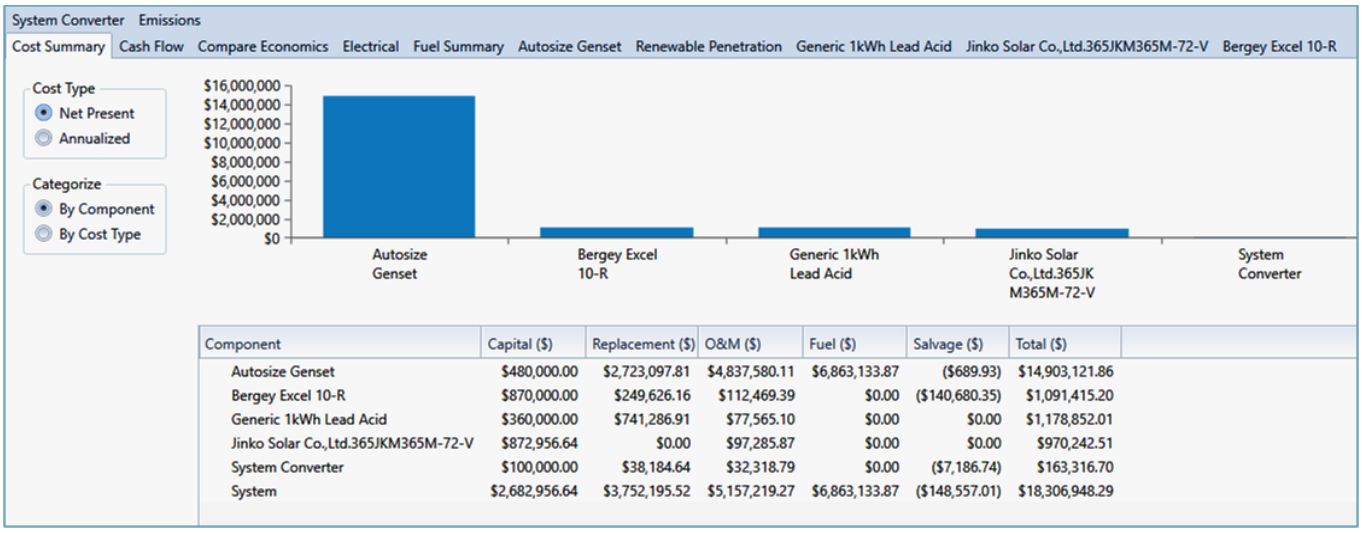


Figure 18: Cost status of the system

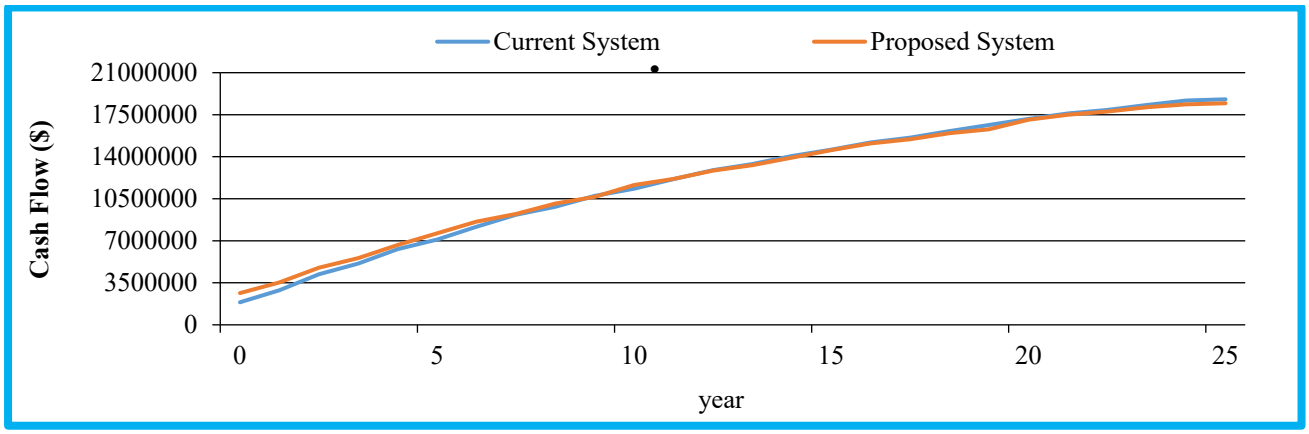


Figure 19: Cumulative cash flow over project lifetime

According to the results of the simulation, it can be observed that of the on-grid hybrid system is the most sustainable hybrid system configuration with the lowest NPC and COE. The total net present cost is $5.425million and the total energy cost is approximately $0.0686 per unit. From the simulation results it is clear that the grid-connected configuration is optimal.

Acknowledgment

This work was supported by the Tunisian Ministry of High Education and Research under Grant LR 11ES15.

References

[1] G. Zhang, C. Xiao, N. Razmjooy, "Optimal operational strategy of hybrid PV/wind renewable energy system using homer: a case study," International Journal of Ambient Energy, **43**(1), 3953–3966, 2022, doi:10.1080/01430750.2020.1861087.

[2] İ. Çetinbaş, B. Tamyurek, M.D. Of, "Design, analysis and optimization of a hybrid microgrid system using HOMER software: Eskisehir osmangazi university example," Int. Journal of Renewable Energy Development (IJRED), **8**(1), 65–79, 2019, doi:10.14710/ijred.8.1.65-79.

[3] M. Usman, M. Khan, A. Rana, S.A. And, "Techno-economic analysis of hybrid solar-diesel-grid connected power generation system," Journal of Electrical Systems and Information Technology, 2017, doi:http://dx.doi.org/10.1016/j.jesit.2017.06.002.

[4] D. Nachtigall, "Improving Economic Efficiency and Climate Mitigation Outcomes through International Co-ordination on Carbon Pricing-Environment Working Paper No. 147," 2019.

[5] L. Oliveros-Cano, J. Salgado-Meza, C. Robles-Algarín, "Technical-economic-environmental analysis for the implementation of hybrid energy systems," International Journal of Energy Economics and Policy, **10**(1), 57–64, 2020, doi:10.32479/ijeep.8473.

[6] F. Rinaldi, F. Moghaddampoor, B. Najafi, R. Marchesi, "Economic feasibility analysis and optimization of hybrid renewable energy systems for rural electrification in Peru," Clean Technologies and Environmental Policy, **23**(3), 731–748, 2021, doi:10.1007/S10098-020-01906-Y.

[7] A.C. Duman, Ö. Güler, "Techno-economic analysis of off-grid PV/wind/fuel cell hybrid system combinations with a comparison of regularly and seasonally occupied households," Sustainable Cities and Society, **42**, 107–126, 2018, doi:10.1016/j.scs.2018.06.029.

[8] S. Mohamed, C. Habib, H. Othman, B. Jamel, "Comparative Analysis of Hybrid Systems for on-grid and off-grid Applications in Tunisia: case study of Audiovisual chain," in 5th International Conference on Advanced Systems and Emergent Technologies, IC_ASET, 450–455, 2022, doi:10.1109/IC_ASET53395.2022.9765947.

[9] S. Goyal, S. Mishra, A.B. And, "A comparative approach between different optimize result in hybrid energy system using HOMER," International Journal of Electrical and Computer Engineering (IJECE), **9**(1), 141–147, 2019, doi:10.11591/ijece.v9i1.

[10] K. Ritu, A. Wadhwani, … A.R., "Techno-Economic Comparison of on Grid and off Grid Hybrid WT/Solar Photo Voltaic Connected Power Generating Unit Using HOMER," in International Conference on Advanced Computation and Telecommunication (ICACAT), 2018, doi:10.1109/ICACAT.2018.8933685.

[11] S. Mohamed, C. Habib, H. Othman, B. Jamel, "Electrical Distribution Architecture and Load Curves Analysis of Audiovisual System," in 2021 IEEE 2nd International Conference on Signal, Control and Communication (SCC). IEEE, 25–30, 2023, doi:https://dx.doi.org/10.1109/SCC53769.2021.9768369.

[12] H.S.A.-E. Mageed, "Cost analysis and optimal sizing of PV-Diesel hybrid energy systems," in American Journal of Renewable and Sustainable Energy, 47–55, 2018.

[13] I. Tizgui, F. El Guezar, H. Bouzahir, A.N. Vargas, "Estimation and analysis of wind electricity production cost in Morocco," in zbw.eu, 58–66, 2018.

[14] A. Said, A. Busaidi, H.A. Kazem, A.H. Al-Badi, M. Farooq Khan, "A review of optimum sizing of hybrid PV–Wind renewable energy systems in oman," International Journal of Students Research in Technology & Management, **2**(3), 93–102, 2015, doi:10.1016/j.rser.2015.08.039.

[15] A.S. Almashakbeh, A.A. Arfoa, E.S. Hrayshat, "Techno-economic evaluation of an off-grid hybrid PV-wind-diesel-battery system with various scenarios of system's renewable energy fraction," Energy Sources, Part A: Recovery, Utilization and Environmental Effects, 2019, doi:10.1080/15567036.2019.1673515.

[16] R.T.A. Al-Rubaye, A.T.A. Al-Rubaye, M.M. Al-Khuzaie, "Optimal Design of Hybrid Renewable Energy System off grid in Al-Diwaniyah, Iraq," IOP Conference Series: Materials Science and Engineering, **454**(1), 2018, doi:10.1088/1757-899X/454/1/012103.

[17] M. Kamran, M. Mudassar, M. Rayyan Fazal, R. Asghar, S. Rukh Ahmed, M. Irfan Abid, M. Usman Asghar, M. Zunair Zameer, C. Author, "Designing and optimization of stand-alone hybrid renewable energy system for rural areas of Punjab, Pakistan," Researchgate.Net, **8**(4), 2018.

[18] A. Al-Sharafi, A. Sahin, T. Ayar, S. Bekir, "Techno-economic analysis and optimization of solar and wind energy systems for power generation and hydrogen production in Saudi Arabia," Renewable and Sustainable Energy Reviews, **69**, 33–49, 2017, doi:https://doi.org/10.1016/j.rser.2016.11.157.

[19] H. Taghavifar, Z.S. Zomorodian, "Techno-economic viability of on grid micro-hybrid PV/wind/Gen system for an educational building in Iran," Renewable and Sustainable Energy Reviews, **143**, 110877, 2021.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 154-160 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Photoluminescence Properties of Eu(III) Complexes with Two Different Phosphine Oxide Structures and Their Potential uses in Micro-LEDs, Security, and Sensing Devices: A Review

Hiroki Iwanaga*

Photoluminescence Material Project, New Business Development Office, Next Business Development Div. Toshiba Corporation, 72-34 Horikawa-cho, Saiwai-ku, Kawasaki 212-8585, Japan

ARTICLE INFO

Article history:
Received: 28 February, 2023
Accepted: 20 May, 2023
Online: 12 June, 2023

Keywords:
Eu(III) complex
Phosphine Oxide
Red Phosphor
micro-LED
Security
Sensing
Display

ABSTRACT

In the field of micro-LED displays, there is strong demand for red phosphors with high photoluminescence intensity, high color purity, and small particle size. Here, we focus on Eu(III) complexes because they produce sharp photoluminescence spectra with high color purity and can be dissolved in polymer, enabling a reduction in particle size to the molecular level. We have previously established novel molecular design concepts for Eu(III) complexes by coordinating two different phosphine oxide structures to one Eu(III) ion in order to enhance photoluminescence intensity and increase solubility in polymers and solvents. Many Eu(III) complexes have been developed based on these concepts and their photoluminescence properties investigated. Eu(III) complexes with two different phosphine oxide structures are important candidates for red phosphors in micro-LEDs.

1. Introduction

Displays require phosphors with high photoluminescence intensity and high color purity. In addition, in micro-LED displays containing ultraviolet (UV) or blue LED arrays and phosphors, where chips are very small, the particle size of phosphors must be sufficiently small to suppress variation in hues among pixels. Therefore, there is a strong demand for a red phosphor that satisfy these conditions. To this end, novel Eu(III) complexes were introduced in a paper originally presented at the 2022 International Conference on Electronics Packaging as a candidate red phosphor for micro-LEDs [1].

In the case of inorganic phosphors, quantum yields decrease with decreasing phosphor particle size because they are present as fine particles in a polymer (Figure 1). Comparison of properties of the Lanthanide complexes and inorganic phosphors are shown in Table 1. The color purity of inorganic phosphors is low because of the large half widths of emission spectra.

Recently, lanthanide complexes, especially Eu(III) complexes, have attracted increasing attention for their application in emission devices, secure media, sensors, and so on [2–8]. Eu(III) complexes are attractive for display use because they produce sharp photoluminescence spectra with high color purity and can reproduce colors in large-area displays.

*Corresponding Author: Hiroki Iwanaga, hiroki.iwanaga@toshiba.co.jp

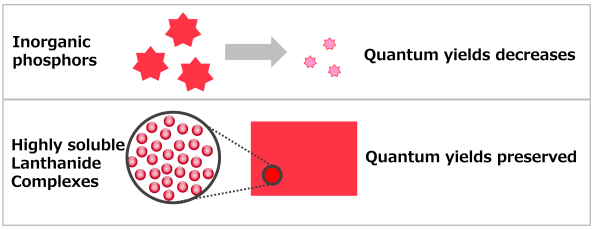


Figure 1: Comparison of inorganic phosphors and highly soluble lanthanide complexes in a polymer.

Table 1: Comparison of properties of the Lanthanide complexes and inorganic phosphors

Phosphors with small particle size	Quantum yield	Color purity
Lanthanide complex	Large	High
Inorganic phosphor	Small	Low

In contrast to inorganic phosphors, particle size is not relevant to the theoretical quantum yield because each molecule of a Eu(III) complex has the function of absorbing and emitting light. From this point of view, Eu(III) complexes are promising candidate red phosphors for micro-LEDs. However, the photoluminescence intensity and solubility of Eu(III) complexes developed to date are insufficient for display use.

An Eu(III) ion itself has very low light absorption and weak emission. However, the emission of lanthanide ions can be enhanced through the antenna effect of ligands. β-diketonates are known to be effective ligands for enlarging photoluminescence intensity of Eu(III) complexes. β-Diketonates absorb light and transfer energy to lanthanide ions efficiently [9, 10]. The photoluminescence intensity of Eu(III) complexes depends largely on the substituents on the β-diketonates because the triplet-state energy levels of β-diketonates are derived from the molecular structures of the substituents. However, it can be difficult to obtain sufficient emission intensity for use in emission devices simply by adjusting the substituents of β-diketonates.

There are two main types in ligands of lanthanide complexes. One is ionic ligands and the other is non-ionic ligands. β-diketonates are prominent ionic ligands and neutralize the charge of lanthanide ions. It is known that photoluminescence intensities are enhanced by the effects of non-ionic ligands in addition to β-diketonates. Phosphine oxide compounds are strong Lewis bases and excellent non-ionic ligands for enlarging photoluminescence intensity [11] (Figure 2).

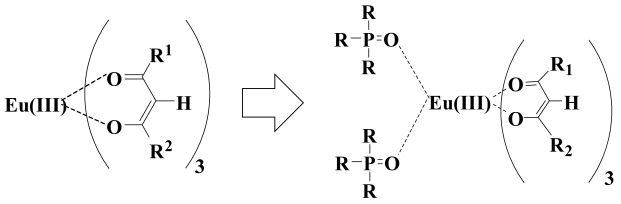


Figure 2: Coordinating phosphine oxides to a Eu(III) ion.

However, there is room for further improvements in emission intensity. At the same time, the solubility of Eu(III) complexes with two identical phosphine oxides are too low to be dissolved in polymers or solvents. For this reason, Eu(III) complexes with low solubility have limited applications.

2. Experimental Section

2.1 Measurement of photoluminescence and excitation spectra

Photoluminescence and excitation spectra were measured at room temperature using a spectrofluorometer (Fluoromax 4, Horiba Jobin Yvon Inc.). Excitation and emission slit widths were set to 0.5 nm for measurement of emission spectra, and to 0.7 and 0.6 nm for measurement of excitation spectra, respectively. Measurement intervals are 1 nm. Scanning rate are 600 nm/min. Dark offset and corrections were applied to both the emission and excitation sites.

2.2 Measurement of emission lifetimes

Measurement of emission lifetimes were performed as follows. Each solution of the Eu(III) complexes was placed in a sealed cell and measured using the spectrofluorometer with the excitation wavelength set to 370 nm. Single exponential functions were used to fit the relative decay curves monitored at the maximum wavelength in order to calculate the emission lifetimes. χ^2 values were in the range of >1.0 and <1.2.

2.3 Measurement of absolute quantum yields

Total absolute quantum yields (Φ_{TOT}) were measured using a photonic multichannel analyzer (PMA-12 C10027-01, Hamamatsu Photonics K.K). An integrating sphere was used for all measurements.

3. Results and Discussion

3-1. Eu(III) complexes with two different phosphine oxides

Figure 3 shows the relationships between molecular structures and photoluminescence spectra of Eu(III) complexes. The photoluminescence intensity of a Eu(III) complex with no phosphine oxide is usually very small, but when two triphenyl phosphine oxides coordinate, it increases to some extent. Furthermore, when two tributyl phosphine oxides coordinate, photoluminescence intensity increases further, and when both triphenyl and tributyl phosphine oxides coordinate, photoluminescence intensity becomes much higher [12]. The important point here is that coordination of two different phosphine oxide ligands is effective for increasing photoluminescence intensity [12–14]. Eu(III) complexes with two different phosphine oxides can be dissolved and are homogeneous at the molecular level in polymers. Polymers containing our Eu(III) complexes are colorless and transparent under room light but emit a pure color when irradiated with UV and 464-nm light.

3.2. Eu(III) complexes with an asymmetric diphosphine dioxide ligand

We detected the ligand exchange of phosphine oxide in Eu(III)-β-diketonates by NMR analysis [13]. However, ligand exchange is expected to have an undesirable effect on durability. To overcome this problem, we developed asymmetric diphosphine dioxide ligands (Figure 4). They have molecular structures consisting of two different phosphine oxide parts and methylene units and suppress ligand exchange via the chelate effect. In addition, the photoluminescence intensity of Eu(III)-β-diketonates with an asymmetric diphosphine dioxide ligand is higher than that with two different phosphine oxides [15, 16].

Tb(III) complexes with two different phosphine oxides or a single asymmetric diphosphine dioxide were also investigated [17]. It was found that solubilities of Tb(III) complexes were increased by coordination of two different phosphine oxide structures. However, photoluminescence intensities are strongly dependent on the substituents of β-diketonates because of the strong influences of back-energy transfer from excited Tb(III) ions to the ligands.

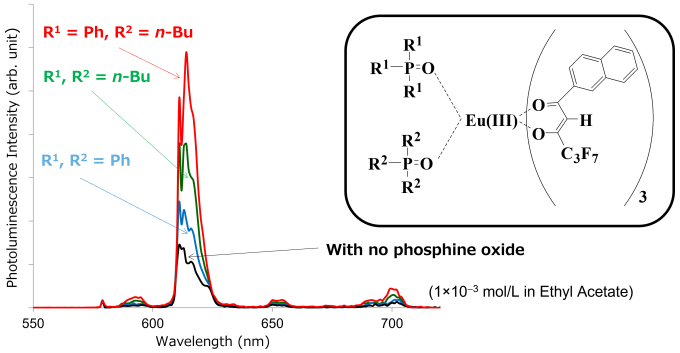


Figure 3: Comparison of the photoluminescence spectra of Eu(III) complexes in ethyl acetate at a concentration of 2×10^{-4} mol/L at room temperature. Showing the effects of phosphine oxides and their combination on photoluminescence intensity [12].

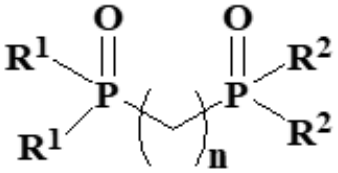


Figure 4: Molecular structures of the asymmetric diphosphine dioxide ligand that increase the photoluminescence intensity of Eu(III) complexes (R^1=aromatic substituent, R^2=aliphatic substituent).

3-3. Solubility of Eu(III) complexes with phosphine oxide ligands

Relationships between the molecular structures of Eu(III) complexes and solubility in solvents were investigated [18]. Eu(III) complexes with two different phosphine oxides are highly soluble in solvents and can be dissolved even in a fluorinated solvent. However, the solubility of Eu(III) complexes with an asymmetric diphosphine dioxide ligand is lower than that of Eu(III) complexes with two different phosphine oxides. We found that meta-substitution of trifluoromethyl groups (CF_3) on the phenyl groups of diphosphine dioxide ligands produces outstanding effects in terms of enhancing the solubility of Eu(III) complexes [19]. Similarly, the solubility of anthraquinone dichroic dyes in fluorinated media are markedly enhanced by the substitution of CF_3 groups [20–22].

3.4. Photoluminescence properties of Eu(III) complexes with an asymmetric diphosphine dioxide ligand

Figure 5 shows the optimal diphosphine dioxide ligand for Eu(III) complexes that increases both quantum yields and solubility [23]. Having CF_3 groups at the meta position of phenyl groups is one of the most important characteristics for achieving both high quantum yield and high solubility. Figures 6 and 7 show the relationships between excitation wavelength and quantum yields of Eu(III) complexes with and without diphosphine dioxide ligands, respectively [23].

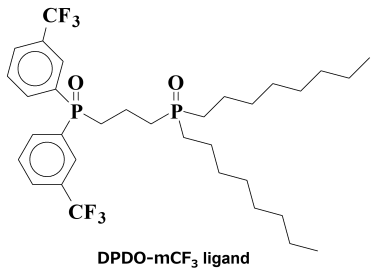


Figure 5: Molecular structure of the asymmetric diphosphine dioxide ligand for Eu(III) complexes that increase the photoluminescence intensity (DPDO-mCF_3 ligand) [23].

The maximum total photoluminescence quantum yield (Φ_{TOT}) of Eu(III)(hfnh)$_3$ is small and the solid-state Φ_{TOT} of Eu(III)(hfnh)$_3$ is smaller than the solution-state Φ_{TOT} caused by concentration quenching (Figure 6). In contrast, Φ_{TOT} of Eu(III)(hfnh)$_3$(DPDO-mCF_3) is much greater than that of Eu(III)(hfnh)$_3$. Furthermore, Φ_{TOT} is greater in the solid state than in the solution state. By coordinating the diphosphine dioxide ligands, quantum yields increase eminently. In the solid state, the maximum quantum yield reaches 0.82.

Diphosphine dioxide ligand functions as a separator, maintaining the distance among Eu(III) ions that prevent concentration quenching. In the solid state, there are no solvent molecules to decrease Φ_{TOT} of the Eu(III) complexes.

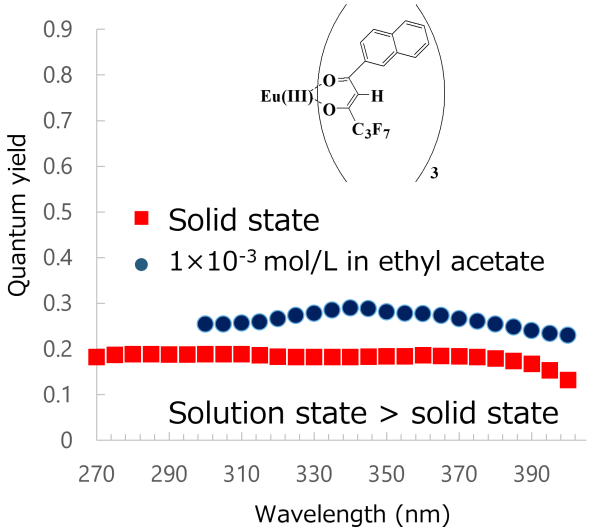


Figure 6: Action spectra (excitation wavelength vs. Φ_{TOT}) of Eu(III)(hfnh)$_3$ in the solid and solution states [23].

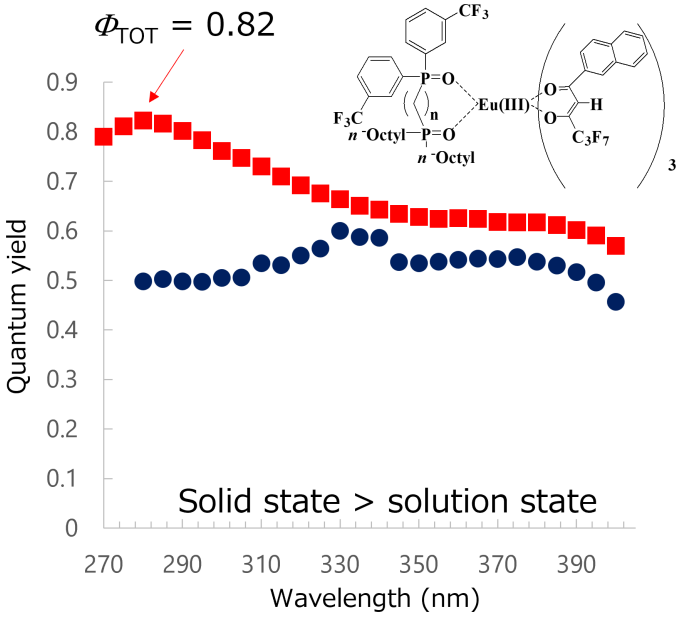


Figure 7: Action spectra (excitation wavelength vs. Φ_{TOT}) of Eu(III)(hfnh)$_3$(DPDO-mCF_3) in the solid and solution states [23].

3.5. Quantum yields of Eu(III) complexes with thienyl substituted diphosphine dioxide

Thienyl groups are electron-donating substituents that are expected to enhance the Lewis basicity of the oxygen atoms in diphosphine dioxide ligands. Dithienyl[3-(dioctylphosphinyl)-propyl] phosphine oxide (DTDOPO) and dithienyl[5-(dibutyl-phosphinyl)pentyl]phosphine oxide (DTDBPO) ligands were developed with the aim of forming stronger coordinate bonds with the Lewis acid Eu(III). A diphenyl[3-(dioctylphosphinyl)propyl]phosphine oxide (DPDO) ligand with

phenyl groups instead of thienyl groups was prepared for comparison (Figure 8) [24].

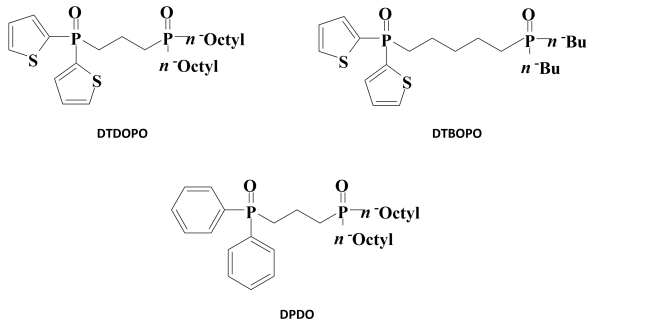


Figure 8: Molecular structures of thienyl-substituted and phenyl-substituted diphosphine dioxides [24].

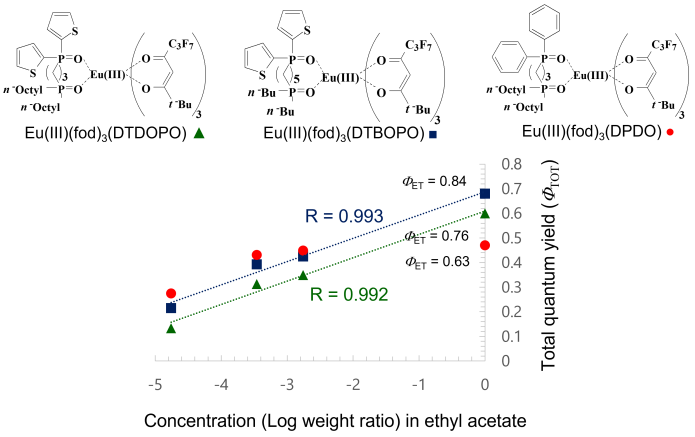


Figure 9: Quantum yields of Eu(III) complexes with thienyl-substituted and phenyl-substituted diphosphine dioxides both in the solid state and in solution (ethyl acetate) [24].

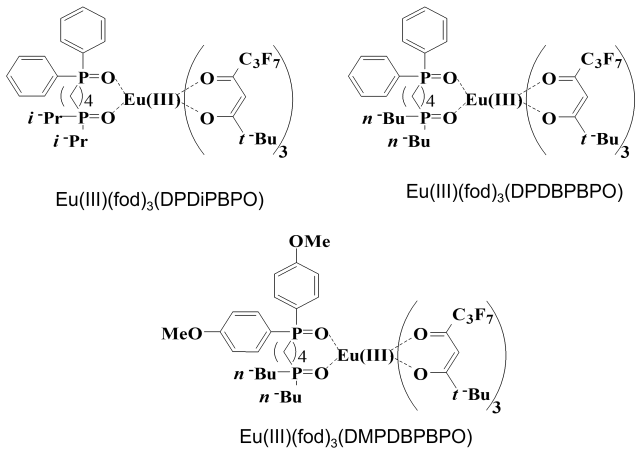


Figure 10: Molecular structures of Eu(III) complexes with a diphosphine dioxide ligand [25]. DPDiPBPO: diphenyl[4-(diisopropylphosphinyl)butyl]phosphine oxide, DPDBPBPO: diphenyl[4-(dibutylphosphinyl)butyl]phosphine oxide, DMPDBPBPO: di(4-methoxyphenyl)[4-(dibutylphosphinyl)butyl]phosphine oxide.

Figure 9 shows the relationships between concentrations in ethyl acetate and quantum yields of Eu(III)(fod)$_3$(DTDOPO), Eu(III)(fod)$_3$(DTBOPO), and Eu(III)(fod)$_3$(DPDO). The concentrations and quantum yields have a strong positive linear correlation and the quantum yields in the solid state (point with concentration [Log weight ratio] 0) are located on the extended line for Eu(III)(fod)$_3$(DTDOPO) and Eu(III)(fod)$_3$(DTBOPO) with thienyl groups. No concentration quenching was observed. As the concentration of Eu(III)(fod)$_3$(DPDO) increases, the differential coefficients become smaller. Research investigating the special feature of Eu(III) complexes with thienyl groups in the solid state is ongoing.

3.6. Effects of alkyl groups in diphosphine dioxide ligands on the photoluminescence properties of Eu(III) complexes

To investigate the effects of the molecular structures of diphosphine dioxide ligands on the photoluminescence properties of Eu(III) complexes, we prepared three Eu(III) complexes with the same molecular structure except for the slight difference in diphosphine dioxide ligands shown in Figure 10 [25].

Eu(III)(fod)$_3$(DPDiPPBPO) has *i*-propyl groups in a diphosphine dioxide ligand, while Eu(III)(fod)$_3$(DPDBPBPO) and Eu(III)(fod)$_3$(DMPDBPBPO) have *n*-butyl groups.

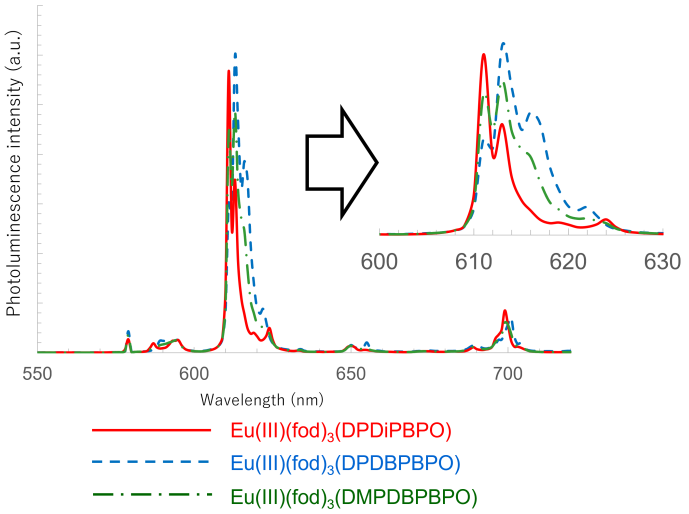


Figure 11: Photoluminescence spectra of the Eu(III) complexes Eu(III)(fod)$_3$(DPDiPBPO), Eu(III)(fod)$_3$(DPDBPBPO), and Eu(III)(fod)$_3$(DMPDBPBPO) in the solid state. They are excited at 370 nm [25].

The photoluminescence spectra of the Eu(III) complexes Eu(III)(fod)$_3$(DPDiPBPO), Eu(III)(fod)$_3$(DPDBPBPO), and Eu(III)(fod)$_3$(DMPDBPBPO) in the solid state are shown in Figure 11. The shapes of the Stalk splitting of the $^5D_0 \rightarrow {}^7F_2$ transition differ among them. Of note, the half-width of the $^5D_0 \rightarrow {}^7F_2$ transition of Eu(III)(fod)$_3$(DPDiPBPO) with *i*-propyl (*i*-Pr) substituted for the diphosphine dioxide ligand was 2 nm and conspicuously smaller compared with Eu(III)(fod)$_3$(DPDBPBPO) and Eu(III)(fod)$_3$(DMPDBPBPO) with *n*-butyl substituted for the diphosphine dioxide ligand. The smaller half-width means the ligand field has a higher symmetry.

Table 2: Photoluminescence properties of the Eu(III) complexes [25]

	Solid state		
Ligand	DPDiPBPO	DPDBPBPO	DMPDBPBPO
τ_{exp} (ms)[a]	0.99	0.84	0.89
k_{exp} (s^{-1})[b]	1009	1193	1129

τ_{rad} (ms)[c]	1.37	0.99	1.09
k_{rad} (s^{-1})[d]	729	1012	992
k_{nrad} (s^{-1})[e]	280	181	207
Φ_{Ln}[f]	0.72	0.85	0.82
Φ_{ET}[g]	0.81	0.83	0.76
I_{MD}/I_{TOT}[h]	0.0678	0.0488	0.0537
Φ_{TOT}[i]	0.58 (345 nm)	0.70 (350 nm)	0.62 (345 nm)
Ratio R[j]	11.1	16.7	14.9

[a]Experimental lifetime measured in solid. χ^2 values were in the range of > 1.0 and < 1.2.

[b]experimental decay rate

[c]Radiative lifetime calculated using the formula

$\tau_{rad} = 1/n^3 A_{MD,0} \times I_{MD}/I_{TOT}$ (n= 1.50).

[d]Radiative decay rate

[e]Non-radiative decay rate

[f]Intrinsic quantum yield calculated using the formula

$\Phi_{Ln} = \tau_{exp} / \tau_{rad}$.

[g]Energy transfer efficiency

[h]Ratio between the integrated intensity of the $^5D_0 \rightarrow {}^7F_1$ transition (I_{MD}) and the total integrated emission intensity $^5D_0 \rightarrow {}^7F_J$(J = 0-6) ($I_{TOT}$)

[i]Total quantum yield measured in solid state. (Peak wavelength of the action spectrum (wavelength vs. quantum yield)).

[j]Calculated from the formula $I(^5D_0 \rightarrow {}^7F_2) / I(^5D_0 \rightarrow {}^7F_1)$

Table 2 shows the photoluminescence properties of the Eu(III) complexes. The Φ_{TOT} of Eu(III)(fod)$_3$(DPDiPBPO) was smaller than that of others because of the smaller intrinsic quantum yield (Φ_{Ln}). The smaller ratio R and larger I_{MD}/I_{TOT} of Eu(III)(fod)$_3$(DPDiPBPO) showed that the Eu(III) complex with *i*-Pr groups in the diphosphine dioxide ligand had a higher symmetry in ligand fields compared with the others in the solid state. These results agree well with the result of the smaller half-width of Eu(III)(fod)$_3$(DPDiPBPO). These noticeable differences in properties are caused by the difference in molecular structures between the *i*-Pr and *n*-Bu groups in diphosphine dioxides.

Based on the above, we propose a hypothesis about Φ_{TOT} and diphosphine dioxide ligand structures: the steric hindrance of diphosphine dioxide ligands with *n*-Bu groups is larger than that of ligands with *i*-Pr groups, and a larger steric hindrance causes diphosphine dioxide ligands to have lower symmetry of the ligand field, thereby inducing a larger Φ_{TOT}. In the next section, we focus on the effects of steric hindrance in diphosphine dioxide ligands.

3.7. Elucidation of the effects of diphosphine dioxide ligands on the quantum yield and photoluminescence intensity of a 6-coordinate Eu(III)-β-diketonate complex

To elucidate the coordination effects of phosphine oxide ligands, the following 6-coordinate Eu(III) complex designed to have low luminescence and a large absorption coefficient was synthesized: (Tris{6,6,7,7,8,8,8-heptafluoro-1-[2-(9,9-dimethylfluorenyl)]-1,3-octanedionate} europium(III) (Eu(III)(hfod)$_3$) (Figure 12) [26].

Dimethylfluorenyl groups are bulky aromatic substituents with large absorption coefficients. Partially fluorinated alkyl groups in β-diketonates are also very bulky. The methylene units in partially fluorinated alkyl groups have the function of decreasing the energy transfer efficiency from the ligands to the Eu(III) ion.

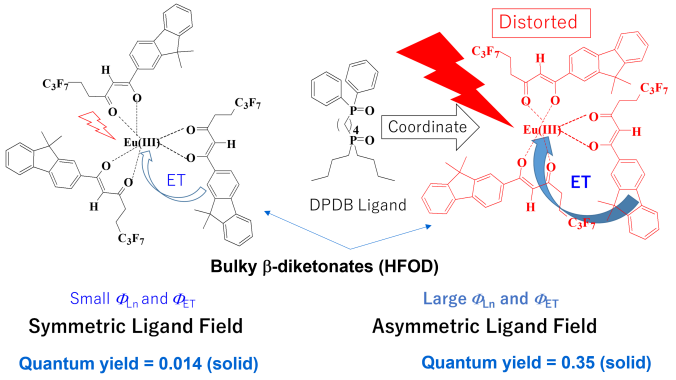


Figure 12: Coordinating effects of DPDB ligand with Eu(III)(hfod)$_3$ with bulky β-diketonates, "hfod" [26].

The photoluminescence intensity of Eu(III)(hfod)$_3$ is dramatically enhanced by coordinating a DPDB ligand (generating Eu(III)(hfod)$_3$(DPDB)) due to the increased Φ_{TOT}, Φ_{Ln}, and Φ_{ET} in both the solution and solid states.

We propose the following hypothesis for the increase in Φ_{TOT} of 6-coordinate Eu(III) complex caused by the effects of the DPDB ligand. When a DPDB ligand coordinates with the Eu(III) ion, the positions of the nearest oxygen atoms around the Eu(III) ion are shifted by steric repulsion, and the relative positions of the nearest oxygen atoms become distorted. Ligand field is asymmetrized by the distorted coordination environment, and that increases Φ_{TOT}.

The norm of the effective dipole moment of the ligand field μ was defined [26]. We demonstrated that the energy transfer efficiencies from the lowest triplet state of the ligands to the $^5D_1$ level of the Eu(III) ion (Φ_{ET}) increases when the ligand fields of the Eu(III) ion become more asymmetric by coordinating the DPDB ligand.

3.8. High-sensitivity method for detecting the pesticide dichlorvos by using Eu(III)-β-diketonate as a quenching probe

Dichlorvos is a general-purpose insecticide with agricultural, household, and animal applications. However, it is harmful to humans, and thus a high-sensitivity method for detecting dichlorvos that provides results in a short time would be desirable.

We found that Eu(III)(hfnh)$_3$ was a highly sensitive luminescent probe for the pesticide dichlorvos. The photoluminescence intensity of Eu(III)(hfnh)$_3$ was drastically and rapidly decreased when a dilute solution of dichlorvos was added to the solutions of Eu(III)(hfnh)$_3$ (Figure 13) [27].

When a solution of dichlorvos was mixed with a solution of Eu(III)(hfnh)$_3$ and shaken, Φ_{ET} drastically decreased (0.64→ 0.15). The photoluminescence quenching of Eu(III)(hfnh)$_3$ by dichlorvos occurred before the energy transfer from β-diketonates to a Eu(III) ion between the dichlorvos molecules and the β-diketonates

The photoluminescence of Eu(III)(hfnh)$_3$ is not quenched by compounds with similar structures and has a favorable selectivity for dichlorvos. These results indicate that Eu(III)(hfnh)$_3$ is a strong candidate for a sensitive, selective, and quick method for detecting dichlorvos. Other organophosphorus pesticides can be selectively detected by other Eu(III)-β-diketonates.

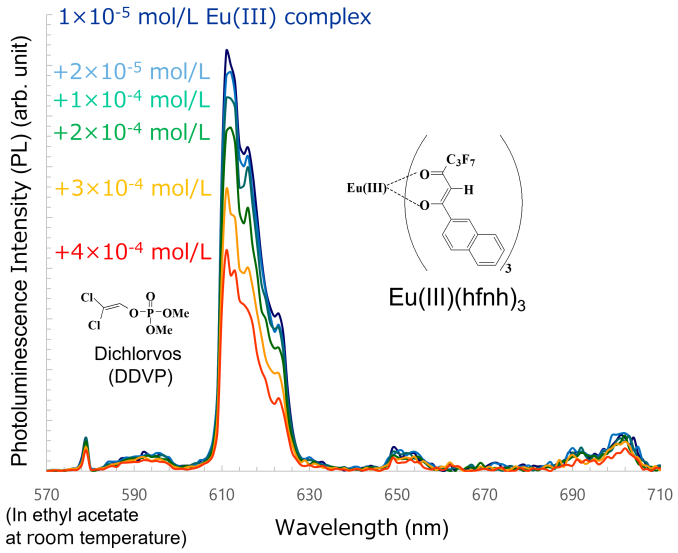


Figure 13: Photoluminescence quenching of Eu(III)(hfod)$_3$ by the pesticide dichlorvos [27].

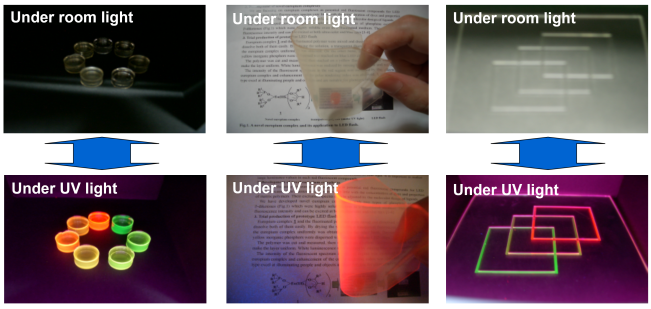


Figure 14: Colorless and transparent photoluminescence materials containing our Eu(III) complexes and/or Tb(III) complexes in a polymer [16].

3.9. Characteristics of colorless and transparent photoluminescence materials involved in our Eu(III) complexes and/or Tb(III) complexes in a polymer

We developed multiple lanthanide complexes having two different phosphine oxides or an asymmetric diphosphine dioxide that improved both photoluminescence intensity and solubility. When Eu(III) or Tb(III) complexes or both are dissolved in a polymer, materials that are colorless and have transparent photoluminescence under room light are produced. These materials emit pure red or green as well as yellow and orange intermediate colors when irradiated with UV or near-UV light (Figure 14) [16].

LED devices comprising a UV-light LED chip and a fluorescent layer consisting of a fluorinated polymer and Eu(III) complexes with two different phosphine oxides were prototyped. The developed devices emit a pure red color. The highest luminous flux obtained under optimum conditions, which to our knowledge is the best result reported as of 2007, was 870 m lumen/20 mA, when excited by a 402-nm LED chip (Figure 15) [28].

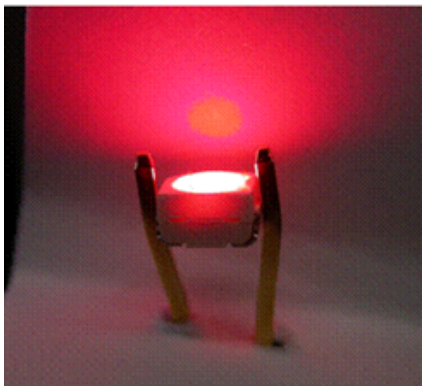


Figure 15: LED devices containing novel Eu(III) complexes in the fluorinated layer [28].

4. Conclusion

We found that coordination of two different phosphine oxide structures to a lanthanide ion is effective for enhancing the photoluminescence intensity and solubility of lanthanide complexes. An asymmetric diphosphine dioxide ligand consisting of two different phosphine oxide parts and methylene units produce further excellent effects in terms of enhancing the quantum yields and photoluminescence intensity of Eu(III) complexes. Asymmetric diphosphine dioxide ligands induce asymmetry in the ligand fields of Eu(III) complexes, thereby improving the quantum yields. Colorless and transparent photoluminescence materials can be obtained by dissolving Eu(III) complexes in polymers or solvents. These materials show great promise for use in LEDs as well as security and sensing devices.

We believe that our Eu(III) complexes have advantages over inorganic phosphors as red phosphors for use in micro-LED displays.

Acknowledgment

The author would like to thank Okuno Atsushi, Takahisa Kobayashi, Junichi Washizuka, Chiari Shimizu, Naruaki Watanabe, Akiko Yuzawa, and Huang Chingchun for fruitful discussions.

References

[1] H. Iwanaga, "Photoluminescence Properties of Eu(III) Complexes with an Asymmetric Diphosphine Dioxide Ligand for Potential Uses in LED, Security, and Sensing Devices," 2022 International Conference on Electronics Packaging (ICEP), 17-18, 2022, doi:10.23919/ICEP55381.2022.9795463.

[2] Q. Xin, W.L. Li, G.B. Che, W.M. Su, X.Y. Sun, B. Chu, B. Li, "Improved

Electroluminescent Performances of Europium-Complex Based Devices by Doping into Electron-Transporting/Hole-Blocking Host," Applied Physics Letters, **89**, 223524, 2006, doi:10.1063/1.2400112.

[3] Q. Xin, W.L. Lia, W.M. Su, T.L. Li, Z.S. Su, B. Chu, B. Li, "Emission Mechanism in Organic Light-Emitting Devices Comprising a Europium Complex as Emitter and an Electron Transporting Material as Host," Journal of Applied Physics, **101**, 044512, 2007, doi:10.1063/1.2655225.

[4] G. Santos, L.G. Paterno, F.J. Fonseca, A.M. Andrade, L.F. Preira, "Enhancement of Light Emission from an Europium(III) Complex Based-OLED by Using Layer-by-Layer Assembled Hole-Transport Films," ECS Transactions, **39**, 307-313, 2011, doi:10.1149/1.3615207.

[5] S.G. Liu, W.Y. Su, R.K. Pan, X.P. Zhou, "Red Emission of Eu(III) Complex Based on 1-(7-(tert-butyl)-9-ethyl-9H-carbazol-2-yl)-4,4,4-trifluorobutane1,3-dione Excited by Blue Light," Chinese Journal of Chemical Physics, **25**, 697-702, 2012, doi:10.1088/1674-0068/25/06/697-702.

[6] A.M. Kaczmarek, Y.Y. Liu, C. Wang, B. Laforce, L. Vincze, P.V.D. Voort, K.V. Hecke, R.V. Deun, "Lanthanide "Chameleon" Multistage Anti-Counterfeit Materials," Advanced Functional Materials, **27**, 1700258, 2017, doi:10.1002/adfm.201700258.

[7] X. Li, J. Gu, Z. Zhou, L. Ma, Y. Tang, J. Gao, Q. Wang, "New Lanthanide Ternary Complex System in Electrospun Nanofibers: Assembly, Physico-Chemical Property and Sensor Application," Chemical Enginerring Journal, **358**(15), 67-73, 2019, doi:10.1016/j.cej.2018.10.003.

[8] G. Lu, X. Kong, C.M. Wang, L.Y. Zhao, D.D. Qi, Y.Y. Jiang, S. Zhao, Y.L. Chen, J.Z. Jiang, "Optimizing the Gas Sensing Properties of Sandwich-Type Phthalocyaninato Europium Complex Through Extending the Conjugated Framework," Dyes and Pigments, **161**, 240-246, 2019, doi:10.1016/j.dyepig.2018.09.062.

[9] S. Sato, M. Wada, T. Seki, "Some properties of europium β-diketone chelates 1. (Synthesis and fluorescent properties)," Japanese Journal of Applied Physics, **7**, 7-13, 1968, doi: 10.1143/JJAP.7.7

[10] S. Sato, M. Wada, "Relation between intramolecular energy transfer efficiencies and triplet state energies in rare earth β-diketone chelate," Bulletin of the Chemical Society of Japan, **43**, 1955-1962, 1970, doi:10.1246/bcsj.43.1955.

[11] J. Yuan, K. Matsumoto, "Fluorescence Enhancement by Electron-Withdrawing Groups on β-Diketones in Eu(III)-β-diketonato-topo Ternary Complexes", Analytical Sciences, **12**, 31–36, 1996, doi:10.2116/analsci.12.31.

[12] H. Iwanaga, "Investigation of strong photoluminescence and highly soluble Eu(III) complexes with phosphine oxides and β-diketonates," Journal of Luminescence, **200**, 233-239, 2018, doi:10.1016/j.jlumin.2018.03.070.

[13] H. Iwanaga, A. Amano, M. Oguchi, "Study of molecular structures and properties of europium(III) complexes with phosphine oxides by NMR analysis," Japanese Journal of Applied Physics, **44**, 3702-3705, 2005, doi:10.1143/JJAP.44.3702.

[14] H. Iwanaga, A. Amano, F. Aiga, K. Harada, M. Oguchi, "Development of Ultraviolet LED Devices Containing Europium (III) Complexes in Fluorinated Layer," Journal of Alloys and Compounds, **408**–**412**, 921-925, 2006, doi:10.1016/j.jallcom.2005.01.138.

[15] H. Iwanaga, F. Aiga and A. Amano, "The molecular structures and properties of novel Eu(III) complexes with asymmetric bis-phosphine oxides," Materials Research Society symposia proceedings, **965**, 211-216, 2006, doi:10.1557/PROC-0965-S03-12.

[16] H. Iwanaga, "Emission properties, solubility, thermodynamic analysis and NMR studies of rare-earth complexes with two different phosphine oxides," Materials, **3**, 4080-4108, 2010, doi:10.3390/ma3084080.

[17] H. Iwanaga, and F. Aiga, "Novel Tb(III) complexes with two different structures of phosphine oxides and their properties", Journal of Luminescence, **130**(5), 812-816, 2010, doi:10.1016/j.jlumin.2009.11.039.

[18] H. Iwanaga, A. Amano, F. Furuya, and Y. Yamasaki, "Solubility in fluorinated medium and thermal properties of Europium(III) complexes with phosphine oxides," Japanese Journal of Applied Physics, **45**, 558-562, 2006, doi:10.1143/JJAP.45.558.

[19] H. Iwanaga, "Relationships between molecular structures of aromatic- and aliphatic-substituted diphosphine dioxide ligands and properties of Eu(III) complexes," Optical Materials, **85**, 418-424, 2018, doi:10.1016/j.optmat.2018.08.071.

[20] H. Iwanaga, K. Naito, and Y. Nakai, "The Molecular structures and properties of anthraquinone-type dichroic dyes," Molecular Crystals and Liquid Crystals, **364**, 211-218, 2001, doi:10.1080/10587250108024989.

[21] H. Iwanaga, K. Naito, and F. Aiga, "Properties of novel yellow anthraquinone dichroic dyes with naphthylthio groups for guest–host liquid crystal displays," Journal of Molecular Structure, **975**, 110-114, 2010, doi:10.1016/j.molstruc.2010.04.003.

[22] H. Iwanaga, and F. Aiga, "Correlations among molecular structures, solubilities in fluorinated media, thermal properties and absorption spectra of anthraquinone dichroic dyes with phenylthio and/or anilino groups," Liquid Crystals, **38**(2), 135-148, 2011, doi:10.1080/02678292.2010.531149.

[23] H. Iwanaga, "A CF3-substituted Diphosphine Dioxide Ligand that Enhances both Photoluminescence Intensity and Solubility of Eu(III) Complexes," Journal of Alloys and Compounds, **790**, 296-304, 2019, doi:10.1016/j.jallcom.2019.03.085.

[24] H. Iwanaga, "Photoluminescence Properties of Eu(III) Complexes with Thienyl-Substituted Diphosphine Dioxide Ligands," Bulletin of the Chemical Society of Japan, **92**(8), 1385-1393, 2019, doi:10.1246/bcsj.20190068.

[25] H. Iwanaga, "Effects of Alkyl Groups in Diphosphine Dioxide-Ligand for Eu(III)-β-Diketonate Complexes on Photoluminescence Properties," Chemical Physics Letters, **736**, 136794, 2019, doi:10.1016/j.cplett.2019.136794.

[26] H. Iwanaga and F. Aiga, "Quantum Yield and Photoluminescence Intensity Enhancement Effects of Diphosphine Dioxide Ligand on a 6-Coordinate Eu(III)-β-Diketonate Complex with Low Luminescence," ACS Omega, **6**(1), 416–424, 2021, doi:10.1021/acsomega.0c04826.

[27] H. Iwanaga and F. Aiga, "A Simple and Sensitive Detection Method for the Pesticide Dichlorvos in Solution using Eu(III)-β-Diketonate as a Luminescent Probe," Japanese Journal of Applied Physics, **59**, SDDF05, 2020, doi:10.7567/1347-4065/ab5c96.

[28] H. Iwanaga, and A. Amano, "Solid-State ^{31}P-Nuclear Magnetic Resonance Analysis of Eu(III) Complexes with Phosphine Oxides in Fluorinated Polymer," Japanese Journal of Applied Physics, **46**, L495-L497, 2007, doi:10.1143/JJAP.46.L495.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 161-169 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Design and Implementation of an Automated Medicinal-Pill Dispenser with Wireless and Cellular Connectivity

Chanuka Bandara[1], Yehan Kodithuwakku[1], Ashan Sandanayake[1], R. A. R. Wijesinghe[2], Velmanickam Logeeshan[*,3]

[1]*Department of Electrical, Electronic and Telecommunication Engineering, Faculty of Engineering, General Sir John Kotelawala Defence University, Sri Lanka*

[2]*Department of Mechanical Engineering, Faculty of Engineering, General Sir John Kotelawala Defence University, Sri Lanka*

[3]*Department of Electrical Engineering, University of Moratuwa, Sri Lanka*

ARTICLE INFO

Article history:
Received: 01 January, 2023
Accepted: 23 February, 2023
Online: 12 June, 2023

Keywords:
Adherence
IoT
Dispenser
Web Interface
Pills
Regimen

ABSTRACT

Medical adherence is a major concern globally and is increasing with improved access to medication. Unfortunately, patients taking multiple medications often struggle with confusion about when and how to take each medication. To address this issue, an inexpensive domestic device has been proposed to improve medication adherence. This device uses Wi-Fi and cellular Internet of Things (IoT) integration to dispense medication at the prescribed times, making it suitable for use in both home and long-term care settings. The device also includes a web interface that allows users to control the device and make changes to dosage and other related information. Additionally, the device features an intricate system for sorting pills to ensure accurate and efficient medication delivery. Automating medication taking through this device can improve patient adherence and overall health outcomes, which could significantly impact public health and quality of life for patients struggling with medication adherence.

1. Introduction

The purpose of this paper is to present a revised and expanded version of the smart medicinal pill dispenser originally presented at the World AI IoT Congress in 2022 [1]. Additional research and testing have been conducted to further improve the device and to provide more in-depth explanations of its functionality and effectiveness. This paper presents the updated findings and conclusions from the continued research, development and testing of the smart medicinal pill dispenser.

Medicine has evolved alongside humanity, with treatments and technologies constantly improving and advancing. While there are still a handful of currently incurable diseases, the vast majority of illnesses can be treated with some form of medication. Despite numerous medical advances, oral medication remains the most convenient and widely available form of treatment [2]. This is due, in part, to the fact that oral medication is easy to administer and does not require specialized equipment or training. As a result, oral medication continues to be a crucial aspect of modern healthcare.

The increasing use of oral medication highlights the importance of effective systems for managing and administering these drugs. In the United States, prescription drug usage in 2020 was 6.324 billion, significantly increasing from the 3.953 billion doses used in 2009 [3]. Globally, the IMS Institute for Healthcare Informatics estimates that 4.5 trillion doses of oral medication were used in 2020, a 24% increase over 2015 [4]. These statistics underscore the need for efficient and convenient systems for administering oral medication, particularly for patients taking multiple drugs or those in long-term care facilities. Effective management of oral medication can help improve patient adherence and overall health outcomes, making it a crucial aspect of modern healthcare.

Managing multiple medications can be challenging for patients, especially when multiple types of drugs are prescribed in a single prescription, as is common practice among doctors. This can be especially difficult for patients who are taking multiple medications or who have complex treatment regimens. A study

*Corresponding Author: Dr. Velmanickam Logeeshan, Department of Electrical Engineering, University of Moratuwa, Sri Lanka. Email: logeeshanv@uom.lk

conducted by [5] found that a significant percentage of patients struggle with complex medication regimens. In fact, only about 15% of the target group in the study organized their dose times to create a more manageable routine. This can be particularly problematic for patients taking number of medications. In [6], the authors showed that the target group in their study used an average of 8 medications in a single prescription.

Adherence to medications is crucial in healthcare, particularly for those taking oral medications. The World Health Organization (WHO) has determined that if a patient's adherence to medication is generally considered satisfactory if the proportion of prescribed medication taken as directed is greater than 80%. This is determined by calculating the number of pills absent in each time period and dividing it by the number of pills prescribed by the physician in that same time period [7].

Unfortunately, many patients struggle to consistently follow their prescribed treatment regimens, leading to significant problems with non-adherence. The WHO recognized this as a major health issue in 2003 [8], and subsequent studies have further highlighted the prevalence and consequences of non-adherence. For example, a study by Eindhoven et al. [9] found that only 40% of the general population consistently followed their prescribed medical routines. Another study conducted in Sri Lanka [10] revealed that a large majority (84.5%) of 303 patients with high blood pressure at the Teaching Hospital of Jaffna were non-adherent to their medication regimens. This was primarily due to forgetfulness and disruptions in daily routines, but other factors, such as managing multiple medications or feeling that they do not receive sufficient attention (particularly among elderly patients) can also contribute to non-adherence. The consequences of medical non-adherence can be severe, including increased risk of antibiotic resistance, worsening of existing conditions, and falling outside of the therapeutic range [11].

According to [12], over 15% of patients ignore the recommended dosages for over-the-counter (OTC) drugs. OTC medications are widely available and do not require a prescription, making them convenient for treating common ailments. Some common OTC medications include acetaminophen, antihistamines, and antacids [13]. However, the ease of access and lack of direct supervision by a healthcare provider can lead to issues with non-adherence to dosage recommendations. This can have negative consequences for patient health and well-being, as taking too much or too little of a medication can have dire consequences. As such, it is important for patients to follow dosage recommendations for OTC drugs carefully and to seek guidance from a healthcare provider if they have any questions or concerns.

In [14], the authors found that accidental overdose on acetaminophen (also known as paracetamol) is a common problem with OTC medications. The same study states that over 23% of participants accidentally overdosed on acetaminophen products. This is often due to a lack of knowledge about proper dosage and dosing intervals for OTC medications and the fact that different medications have different recommended dosages and dosing intervals. For example, the recommended adult dosage for acetaminophen is two pills every 6 hours [15], while the recommended adult dosage for ibuprofen is 2 pills every 8 hours [16]. This suggests that while many people are aware of common OTC medications and their intended uses, they may not clearly understand proper dosage and dosing intervals. This lack of knowledge can increase the risk of accidental overdose and other adverse health consequences.

Inadequate adherence to medication regimens is a significant issue that contributes to negative health outcomes and increased healthcare costs. One of the major challenges in ensuring proper medication adherence is the complexity of some regimens, which can involve taking multiple medications at different times of day or administering drugs in unconventional ways. These factors can be confusing for patients and may lead to doses being missed or taken at the wrong times [17]. According to the Centers for Disease Control and Prevention (CDC), the United States experiences over 2.8 million antibiotic-resistant infections annually, and antibiotic resistance is responsible for over 35,000 deaths [18]. Improper medication use, including self-medication with antibiotics and not following dosage instructions, contributes to the growing problem of antibiotic resistance.

2. Analysis of Survey and Interview Data on Medicine Usage

2.1. Survey analysis

A survey was conducted among a sample of Sri Lankan adults over the age of 40 years to examine patterns of medicine used in this population in 2022. The survey sample was obtained through distribution to university students, resulting in a total of approximately 320 participants.

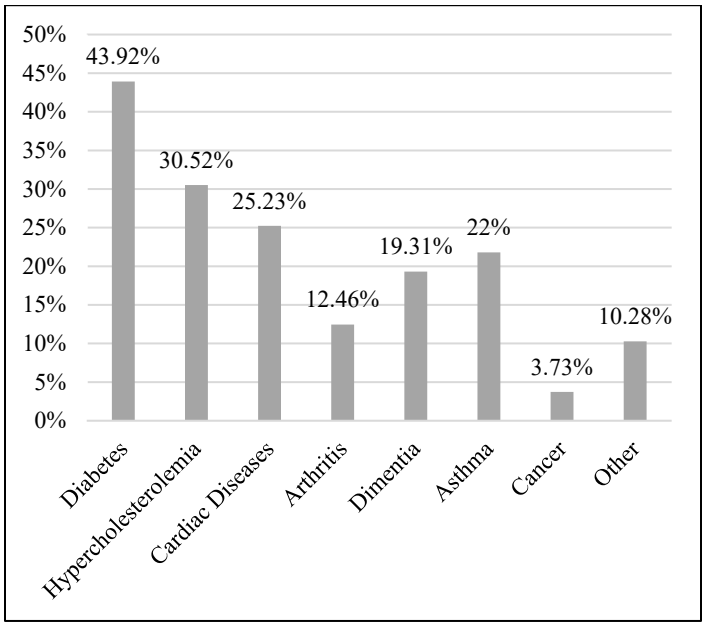


Figure 1: Proportion of long-term illnesses in the survey sample population

The survey results revealed a high prevalence of chronic diseases among Sri Lankan adults (Figure 1). In particular, it was found that over 95% of the target group had at least one chronic condition, with diabetes being the most common at over 40%. Additionally, the survey revealed that over 30% of Sri Lankan adults had high blood cholesterol or hypercholesterolemia. In addition to the high prevalence of diabetes and hypercholesterolemia among Sri Lankan adults, the survey also

identified other common chronic conditions. Cardiac diseases, such as hypertension, were prevalent among a significant percentage of the target group. Additionally, a significant number of participants reported suffering from arthritis. Cancer was also reported among a small percentage of the sample. The findings of the 2022 survey are consistent with those obtained in a previous survey on the same topic in 2021 by Bandara C. et al. [1].

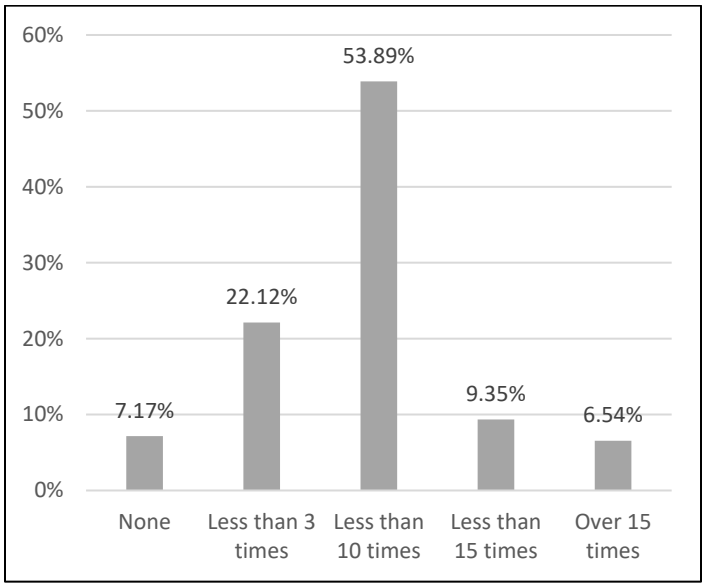


Figure 2: Proportion of sample population reporting medication non-adherence within a month

In addition to examining the prevalence of chronic diseases among the survey sample, the survey also assessed participants' adherence to their medication regimens (Figure 2). The results showed that under 30% of the population consistently took their medication without missing doses or rarely lacking. All other participants reported frequently forgetting to take their medication. These findings suggest that medication adherence is a significant issue among the Sri Lankan adult population, with the majority of participants struggling to consistently follow their prescribed treatment regimens.

2.2. Interview findings

A series of interviews were conducted with professionals in the medical sector to explore the issue of medication adherence. The interviews confirmed that patients are often prescribed multiple medications to treat their underlying diseases, manage symptoms, and control pain. It was noted that similar drugs are often included in the same regimen, which can lead to confusion for some patients. The professionals interviewed emphasized the importance of clear communication between healthcare providers and patients to ensure that patients fully understand their prescribed treatment regimens and the potential risks and benefits of each medication.

One potential source of confusion and non-adherence among patients taking multiple medications is the similarity in the appearance of certain drugs. For example, diabetes medications such as Metformin and Gliclazide may be easily mistaken for one another, particularly if they are taken at different frequencies (e.g., Metformin once daily and Gliclazide twice daily). Similarly, hypertension medications like Hydrochlorothiazide and Losartan Potassium, which may be taken at different frequencies (e.g., Hydrochlorothiazide twice daily and Losartan Potassium once daily), may also be confusing for some patients.

The interviews also revealed that patients may neglect their medication dosages due to busy schedules and may simply ignore their medication altogether. Additionally, older patients may not receive sufficient care to remind and administer their medication at home. These findings suggest that time constraints and lack of support may contribute to non-adherence among certain patient populations.

Although medical professionals are actively working to minimize the issue of non-adherence to medication regimens, the lack of resources and lack of patient cooperation are significant challenges. Despite efforts to educate patients about the importance of adherence and to provide support to help patients follow their prescribed treatment regimens, many patients continue to struggle with non-adherence.

3. Existing Approaches

3.1. The traditional approach to managing multiple medications.

The traditional method of accessing and administering prescribed medications (Figure 3) typically involves manually checking the prescription and individually taking each type of medication in the prescribed dosage. This process can be tedious and time-consuming, and it can also lead to errors. For example, if the patient becomes distracted or fatigued while following their medication regimen, they may accidentally skip a dose or take an incorrect dosage. Additionally, if the patient has multiple medications with different dosing instructions, it can be easy to confuse or mix up the instructions, leading to further errors in medication adherence. This can have serious consequences for the patient's health, as incorrect dosages or missed doses can lead to ineffective treatment or negative side effects.

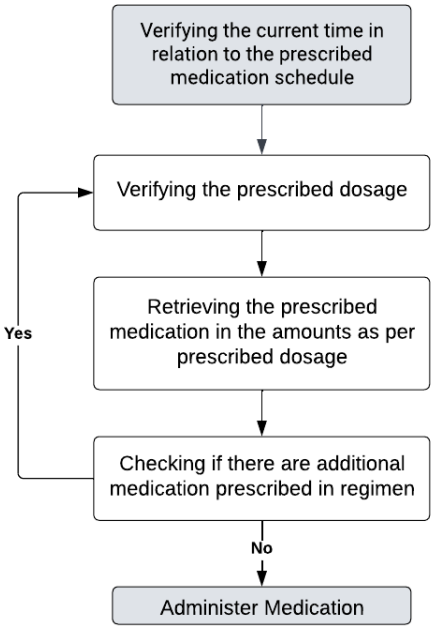


Figure 3: Flowchart of managing multiple medication by hand.

3.2. Prevailing products

To address these challenges, a number of products have been developed and are currently available on the market where each of these products has its own limitations and challenges. Some major examples include, Pointells Automatic Pill Dispenser, Hero Medication Manager, e-Pill Voice Pro, MedaCube, and RxPense.

The Pointells Automatic Pill Dispenser requires users to manually sort and store each dose in a separate compartment, which can be time-consuming and inconvenient [19]. The Hero Medication Manager offers a user-friendly interface but is only available in certain regions and requires a subscription fee in addition to its high cost [20]. The e-Pill Voice Pro has many features, but also shares the limitations of the Pointells Automatic Pill Dispenser and includes an alarm feature that may disrupt users' daily routines [21]. The MedaCube is a highly advanced medication dispenser but is expensive and not widely available [22]. The RxPense is a more compact alternative to the MedaCube, but also has a subscription-based payment model and limited global availability [23].

3.3. Comparable prototypes

The automated pill dispensing device by Ramkumar et al. [24] proposes a solution to the problem of medication non-adherence in patients using an automated pill dispenser that is connected to the Internet of Things (IoT). The proposed solution leverages the IoT to provide patients with medication reminders, dispense the right pills at the right time, and enable healthcare professionals to monitor their patients' medication adherence in real-time. However, the device by Ramkumar et al. lack the ability to provide connectivity when the Wi-Fi connectivity is absent.

In [25], the researchers proposes an autonomous bot to administer medication to elderly patients. The bot uses line-following mechanism to track the patient's location and dispenses the required medicines based on user-programmed instructions. The system store information in the cloud for future reference. However, the device does not present a mechanism for the sorting and dispensing of pills in accordance with a predetermined schedule. Rather, its function lies in the delivery of medication to the patient and dispensing the pre-sorted pills as required.

The smart automated pill dispenser by Kumar [26] proposes a Wi-Fi enabled device with a smart app and a pill dispensing mechanism to address medication non-adherence. The device uses four separate cartridges for different sized pills. Which could limit the usability of the device.

To overcome the limitations of the implementations, it is suggested that a Global System for Mobile Communications (GSM) module be integrated into the system, a mechanism for automated sorting of medication without manual pre-sorting should be proposed, and the design of cartridges must also be optimized to accommodate pills of varying sizes and function seamlessly with the proposed sorting mechanism.

4. Proposed Device

The proposed device is designed to dispense medication at the prescribed times and remind the patient or guardian to administer the medication. This device conveniently stores the medications and manages the dosage, eliminating the need for the patient or guardian to worry about finding and correctly dosing the medications. With this device, users simply need to take the medication at the prescribed times, with the device managing all other aspects of the medication regimen. This feature allows for increased convenience and ease of use for the patient or guardian.

The proposed device is designed with several key components, including an ESP32 microcontroller, infrared sensors, a GSM module, a real-time clock module, and motors and motor drivers. These components work together to ensure that the device can dispense medication accurately at the prescribed times and provide reminders to patients or guardians as needed. Advanced technologies such as the ESP32 microcontroller and GSM module allow for integration with the IoT, enabling control of the device through a web interface. The inclusion of infrared sensors and a real-time clock module allow for precise timing and accuracy in the dispensing process. The motors and motor drivers ensure that the device is able to dispense the required medication doses efficiently. Overall, these components are essential for the effective functioning and performance of the proposed device.

4.1. Operation

The proposed device is designed to ensure that patients are able to take their medication on schedule and avoid missing doses (Figure 4). To accomplish this, the device uses a real-time clock module to track the current time and a GSM module to send reminders to the patient if they are behind schedule for their medication. Additionally, the device is equipped with infrared sensors and motors to dispense the appropriate dosage of medication based on the prescribed schedule. If the patient has not taken their previous dose, the device will not dispense the next one, but will instead send a reminder and reset its status so that the next dosage can be issued at the appropriate time.

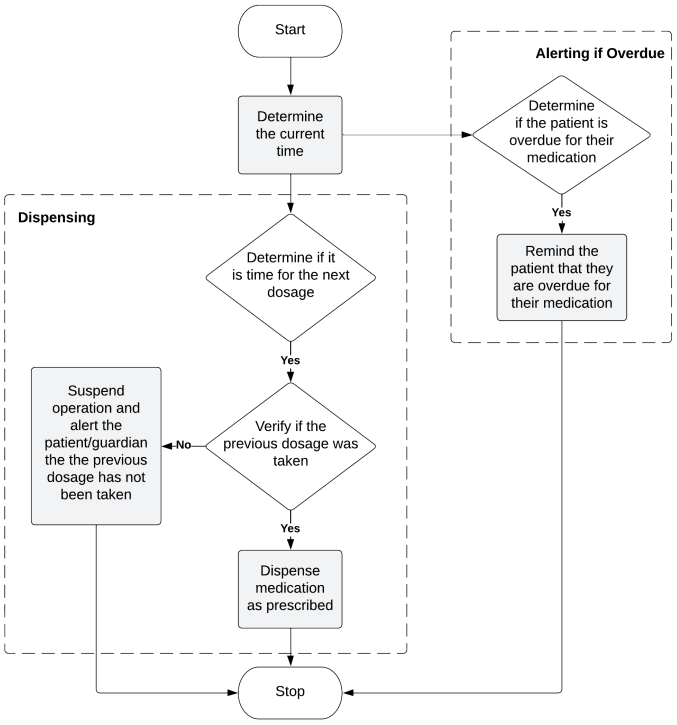


Figure 4: Operation diagram of the proposed device

4.2. Physical design

The proposed device includes several compartments designed to dispense a specific medication. The design of these compartments includes a handle, an arm, a funnel, an infrared sensor space, a pill exit, and a disk (Figure 5).

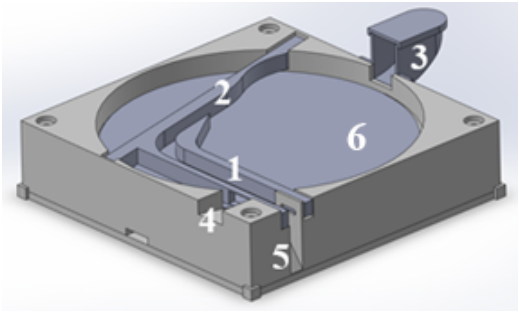


Figure 5: 3D model of a compartment. 1. Handle 2. Arm 3. Funnel 4. IR sensor space 5. Pill exit 6. Disk

The proposed device is designed to dispense medication using a rotating disk and an arm mechanism. Medication is inserted into the device through a funnel, and then the disk rotates to separate a single pill. The pill is then passed through a gap between the arm and handle, and an infrared sensor detects its movement and signals the microcontroller to stop the rotation and dispense the pill. The motor used to rotate the disk is located within the compartment and attached to the disk. An additional infrared sensor is included to detect when the cup, which holds the dispensed medication, has been moved by the user, indicating that the medication has been taken.

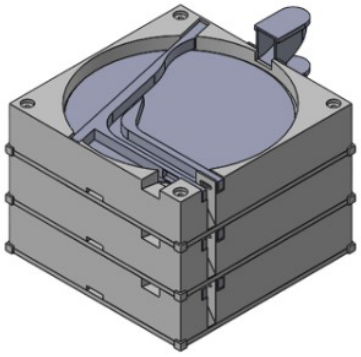


Figure 6: Compartments stacked

The proposed device includes individual compartments for each type of medication, with each compartment being approximately 15 cm by 15 cm in size. This allows for proper storage and organization of medications, ensuring that patients or caregivers can easily access the correct medication at the appropriate time. Multiple compartments can be stacked on top of each other for efficient storage and organization of multiple medications (Figure 6). The device also includes motors and infrared sensors to facilitate the dispensing process and track medication intake.

4.3. Working mechanism

The proposed device utilizes a rotating disc mechanism to dispense individual pills. The rotational speed of the disc is calculated to ensure that only a single pill is dispensed at a time. This is achieved by considering the relationship between angular velocity and linear velocity. The linear velocity of a rotating disc increases as it moves further away from the center, and by carefully controlling the angular velocity, the device is able to accurately dispense single pills. This mechanism allows for the device to effectively dispense a wide range of pill sizes and shapes.

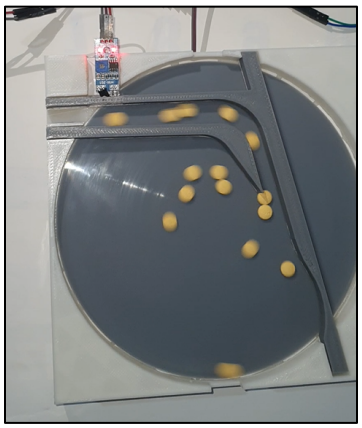


Figure 7: Extracting a single pill

Figure 7 illustrates a momentary snapshot of the rotation of the disk in an open compartment unit. Pills closer to the center of the disc experience a lower linear velocity, while those further from the center experience a higher linear velocity. This can be observed in the differential blur of pills in the center versus those at the edges of the disc in Figure 7. This method allows for efficient and accurate dispensing of medication.

The pill separation is achieved by first rotating the disk in a backward direction (clockwise in Figure 7) at varying speeds to disperse the pills across the surface of the disk. The disk is also rapidly oscillated back and forth to further spread out the pills. After this initial shaking process, the disk is then rotated in the forward direction through the path between the arm and handle of the device. As the pills are guided through this path, the linear velocity helps to create a gap between each pill. This method allows for the efficient separation of individual pills for accurate and reliable dispensing.

To optimize the dispensing process, the proposed device includes a dynamic adjustment feature for the rotation time and speed of the disk. This allows for the proper isolation of individual pills and ensures that they are dispensed accurately. The user can also manually reset and calibrate the rotation time using the provided controls. If a pill is not dispensed after two attempts, the rotation time and speed is adjusted, and the process is repeated. The updated rotation time and speed are then recorded in the microcontroller for future use.

It is important to consider the potential for damaging medication when spinning pills at high speeds. While high speeds may be effective in isolating individual pills, they may also cause pills to break or crumble. This can result in reduced efficacy of the medication or even potential harm to the patient if they ingest broken or damaged pills. It is essential to carefully consider the appropriate speed range for spinning pills in order to minimize the risk of damage while still effectively isolating individual doses. Additionally, the material and construction of the spinning

mechanism should be carefully considered to ensure that it is strong enough to handle the forces involved without causing damage to the pills.

4.4. IoT implementation

A cellular IoT module was implemented to provide reminders to users via short message service (SMS). This allows users to be reminded of their medication schedule even when they are not at home, simply by using their mobile device. For example, if a user's medication time is set for 8 pm, they will receive an SMS at 8.30 pm reminding them that they are half an hour past their medication time and encouraging them to take their medication as soon as possible. The delay in sending reminders for medication is intended to provide patients with a grace period to self-administer medication, if they recall it, without being prompted unnecessarily. This approach is aimed at minimizing the potential inconvenience. This implementation in overall adds convenience for users without the need for additional equipment.

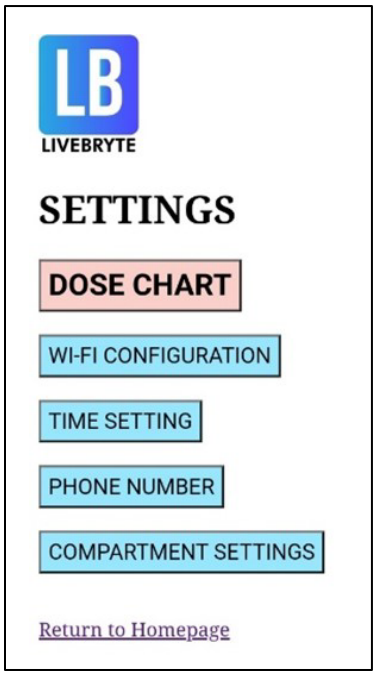


Figure 8: Settings page of the web-interface

The inclusion of the ESP32 microcontroller offers numerous benefits, including its lightweight design and 4 MB of SPI flash memory for permanent data storage [27]. The ESP32 microcontroller and Wi-Fi IoT capabilities to enable access to device configurations through an HTML and CSS based web interface in Figure 8. This interface can be accessed through a local Wi-Fi network or through the microcontroller's own soft access point.

The web interface for the proposed device allows users to easily enter and modify their medication prescription information, including dosage and timing. This interface also allows users to customize device settings such as the device time, phone number for reminder messages, and Wi-Fi network to which the device is connected. By leveraging the IoT through the use of the ESP32 microcontroller, this interface can be accessed remotely, providing users with greater convenience and flexibility. It is important to note that utilizing an internet-connected network is recommended to enhance the user experience. Additionally, this web interface can be easily updated by linking additional information through URLs, allowing developers to easily make changes to the device without the need for hardcoding data directly into the device.

The proposed device includes a microprocessor with the capability to record and notify the user of any errors that may occur during device operation. This feature is implemented through the ESP32 microcontroller and its ability to store data in flash memory, even during a power interruption. This ensures that the device remains functional and able to provide necessary medication to the user [28].

4.5. User operation

To initiate the operation of the device, To use the device for the first time, the user must access the device's own soft access point (AP) address using the provided instructions. Then, they should enter the Service Set Identifier (SSID) and password for their current local network into the device. This will allow the device to connect to the local network and be accessible for further configuration and use.

Afterwards, the patient or guardian must first the user must load the appropriate medications into compartments using the provided funnel and input relevant information, such as the type of medication pills being stored in each container and their prescribed dosages using the “Dose Chart” page in the web-interface settings page shown in Figure 8. The "calibrate" feature should then be run using the “Compartment Settings” page, during which the device will dispense a single pill from each container to determine the optimal settings, including time and rotation speed, for the specific type of pill. The user may also need to adjust the timing and provide a phone number for alert notifications.

In the event of an error, whether due to missed medication or a technical malfunction of the device, the device ceases operation and alerts the user via SMS. The guardian or nurse must then confirm that the issue has been resolved and instruct the device to resume operation. This helps to ensure that any errors or disruptions in the medication regimen are promptly addressed and corrected.

Each dose must be double-checked before administration due to the sensitivity of the device's operation.

5. Final Device Design

The completed product, named "LIVEBRYTE," was made primarily of wood, with 3D printed parts made of polylactic acid (PLA).

5.1. Accuracy

There are several limitations to the current design of the proposed device. The device is currently limited to pills. Another limitation is the reliance on infrared sensors to detect pills as they are dispensed. This can be problematic as some pills may not be properly detected, as pills are passing the sensor at high velocities leading to inaccurate dispensing. In addition, the device is not able to handle pills that are particularly large or oddly shaped, as they may not pass through the dispensing path correctly. The device is also limited in its ability to handle many medications, as it can only

store a limited number of pills in each compartment. Finally, the device relies on the user to accurately input information about the medications and their prescribed dosages, which can lead to errors if the user is not careful or if there are changes to the prescription. Despite these limitations, the device has shown significant improvement in accuracy during the troubleshooting phase, increasing from 40% to its current level through algorithmic improvements.

Figure 9: The final product

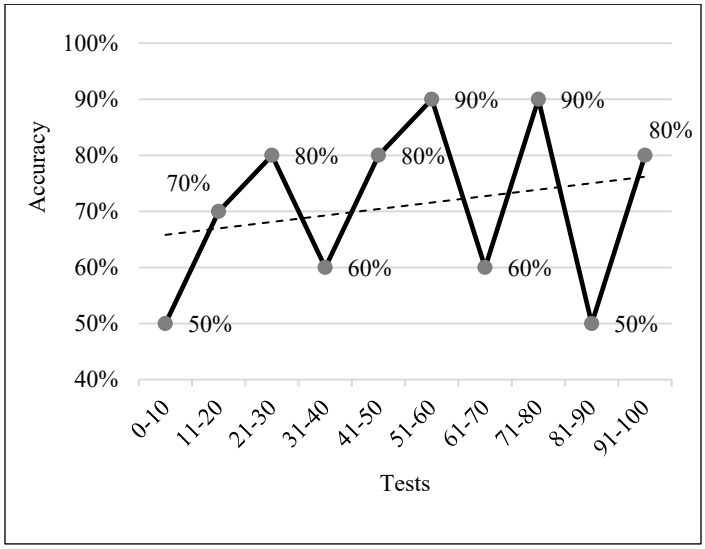


Figure 10: Accuracy for 100 tests

The accuracy of the LIVEBRYTE device was evaluated through a series of tests in which the device was instructed to dispense a single pill. First, the number of successful dispenses, defined as instances in which only a single pill was released, was recorded, and compared to the total number of dispense attempts. Next, unsuccessful dispenses, defined as instances in which no pills or more than one pill were released, were also recorded. The device's accuracy was then calculated as the percentage of successful dispenses out of the total number of dispense attempts. This indicated the reliability and precision of the device in accurately fulfilling prescribed dosage regimens.

The results of these tests, conducted 100 times with an average accuracy of 71%, are depicted in Figure 10. The trendline illustrates that the accuracy of the device improved dynamically with dynamic adjustments.

The use of an infrared sensor as a proximity sensor ensures that the accuracy of the device is not affected by the color of the medication. However, it has been observed that pills with a glossy outer coating tend to be less responsive to the infrared sensor. Despite this, it can be ensured that there will not be any confusion between different medications, as each medication is stored in a separate compartment.

5.2. User feedback

The final product was provided to 10 long-term patients with varying levels of proficiency in using technology for a period of 10 weeks, or approximately 2.5 months. The patients were instructed to carefully review their medication doses and prescriptions before taking their medications, and the device was used to dispense and track their medication adherence. A survey was conducted at the end of the trial to gather feedback on the patients' experiences with the device.

The results of the survey indicated that the majority of the participants were satisfied with the LIVEBRYTE device and believed that it aided in their medication adherence. On a scale of 1 to 5, the majority of participants rated the product as 4 out of 5. They reported that the device improved their medical adherence and that they experienced few errors during the week-long trial period. On average, participants experienced errors 6.09 times per week. This is consistent with the expectations for the device, given that the device was expected to produce 7.1 errors in average.

6. Conclusion

The proposed device is a smart pill dispenser designed to help patients and their caregivers manage and adhere to their prescribed medication regimens. The device stores multiple medications and dispenses them at the prescribed times, providing a convenient and reliable solution for those who struggle with medication management. In addition to dispensing medication, the device also sends reminders to the patient or caregiver to take their medication as prescribed. This can be especially helpful for those who take multiple medications and may have difficulty remembering which medications to take and when. The device is also equipped with a web interface that allows users to easily access and control the device, making it suitable for use in both home and long-term care settings. Overall, the automation of medication taking through this device has the potential to improve patient adherence and overall health outcomes, which could have a significant impact on public health and quality of life for patients struggling with medication adherence.

To further assess the usability and effectiveness of the device, it was given to 10 long-term patients for a period of 10 weeks (about 2 and a half months). These participants were instructed to always double-check their medication doses and their prescription before taking their medications. The results of a survey conducted after this trial period showed that most participants were satisfied

with the product, with most rating it 4 out of 5. They reported that the device had helped with their medication adherence and that they had experienced few errors, with an average expected error rate of 6 per participant.

Despite the efforts made to improve accuracy, it is important to note that the proposed device is not intended to replace the role of a nurse or guardian in ensuring that patients take their medication as prescribed. The device is simply intended to assist in the process of medication adherence and should be used in conjunction with supervision to ensure the safe and effective administration of medication. It is also important to note that the device is not foolproof and may still be subject to errors or malfunctions. Therefore, it is essential that patients and guardians continue to carefully monitor their medication use and seek medical advice if any concerns arise.

It is important to consider the ethical implications of using this device. One potential concern is the potential for patients to rely too heavily on the device and neglect their own responsibility to manage their medication. It is essential that patients are educated on the importance of self-management and encouraged to take an active role in their own healthcare. Additionally, it is important to ensure that the device is used in accordance with the patient's prescription and that any changes to the dosage or medication schedule are made in consultation with a healthcare provider.

Overall, the proposed device has the potential to significantly improve medication adherence and overall health outcomes for long-term patients. As outlined above, the device has successfully addressed the limitations of the compared devices and has incorporated additional features, thus enhancing its overall functionality.

There are several potential future improvements that could be made to the device to make it even more effective at improving medication adherence and overall patient health outcomes. One potential improvement could be to integrate more advanced sensors, such as weight sensors, to track medication usage and detect any potential issues with pill dispensing more accurately. Moreover, improvements for the physical compartment designs can be done to elevate the device efficiency and to accommodate a wider range of medications. The device could be developed to accommodate medication for multiple patients, providing the device much more capable in long-term care facility administrators. Additionally, incorporating machine learning algorithms into the device's software could allow it to better adapt to individual patient needs and preferences, as well as optimize the timing and frequency of medication reminders based on user data. Other potential improvements could include integrating more advanced communication capabilities, such as voice recognition and virtual assistant functionality, to make the device more user-friendly and accessible for patients with limited technology skills. Finally, incorporating more robust data storage and analysis capabilities could allow the device to track patient medication usage and provide insights to healthcare providers on how to optimize treatment plans more effectively. Furthermore, implementing automatic refill reminders could help to ensure that patients never run out of their medication, while the ability to track and record medication intake could provide valuable information for healthcare professionals to use in managing their patients' treatment. Additionally, integrating the device with healthcare systems could facilitate better communication and collaboration between patients and healthcare professionals. To ensure that patients are able to effectively use and benefit from the device, developing user-friendly interfaces and providing instructional materials could be helpful. Finally, it is important to conduct further testing and evaluation of the device to understand its potential impact and ensure that it is being used safely and appropriately.

References

[1] C. Bandara, A.D. Sandanayake, Y. Kodithuwakku, V. Logeeshan, "Automated Medicinal-Pill Dispenser with Cellular and Wi-Fi IoT Integration," in 2022 IEEE World AI IoT Congress (AIIoT), IEEE: 692–698, 2022, doi:10.1109/AIIoT54504.2022.9817226.

[2] M.S. Alqahtani, M. Kazi, M.A. Alsenaidy, M.Z. Ahmad, "Advances in Oral Drug Delivery," Frontiers in Pharmacology, **12**, 62, 2021, doi:10.3389/FPHAR.2021.618411/BIBTEX.

[3] Medicines Use and Spending in the U.S, 2017.

[4] Global Medicines Use in 2020, Mar. 2022.

[5] M.S. Wolf, L.M. Curtis, K. Waite, S.C. Bailey, L.A. Hedlund, T.C. Davis, W.H. Shrank, R.M. Parker, A.J.J. Wood, "Helping Patients Simplify and Safely Use Complex Prescription Regimens," Archives of Internal Medicine, **171**(4), 300–305, 2011, doi:10.1001/ARCHINTERNMED.2011.39.

[6] D. Garfinkel, D. Mangin, "Feasibility Study of a Systematic Approach for Discontinuation of Multiple Medications in Older Adults," Archives of Internal Medicine, **170**(18), 2010, doi:10.1001/archinternmed.2010.355.

[7] M.T. Brown, J.K. Bussell, "Medication Adherence: WHO Cares?," Mayo Clinic Proceedings, **86**(4), 304–314, 2011, doi:10.4065/mcp.2010.0575.

[8] S. de Geest, E. Sabaté, "Adherence to long-term therapies: evidence for action," European Journal of Cardiovascular Nursing : Journal of the Working Group on Cardiovascular Nursing of the European Society of Cardiology, **2**(4), 323, 2003, doi:10.1016/S1474-5151(03)00091-4.

[9] D. lle C. Eindhoven, A.D. Hilt, T.C. Zwaan, M.J. Schalij, C.J.W. Borleffs, "Age and gender differences in medical adherence after myocardial infarction: Women do not receive optimal treatment The Netherlands claims database," European Journal of Preventive Cardiology, **25**(2), 181–189, 2017, doi:10.1177/2047487317744363.

[10] S. Pirasath, T. Kumanan, M. Guruparan, "A Study on Knowledge, Awareness, and Medication Adherence in Patients with Hypertension from a Tertiary Care Centre from Northern Sri Lanka," International Journal of Hypertension, **2017**, 2017, doi:10.1155/2017/9656450.

[11] A.F. Yap, T. Thirumoorthy, Y.H. Kwan, "Medication adherence in the elderly," Journal of Clinical Gerontology and Geriatrics, **7**(2), 64–67, 2016, doi:10.1016/J.JCGG.2015.05.001.

[12] N. Kheir, M.S. el Hajj, K. Wilbur, R.M.L. Kaissi, A. Yousif, "An exploratory study on medications in Qatar homes," Drug, Healthcare and Patient Safety, **3**(1), 99, 2011, doi:10.2147/DHPS.S25372.

[13] M.S. Rsfa, J.A.C. Perera, P.P.R. Perera, The usage of over the counter (OTC) medicines and traditional medicines (TMs) for common ailments in selected urban and rural areas in Sri Lanka, 2015.

[14] M.S. Wolf, J. King, K. Jacobson, L. di Francesco, S.C. Bailey, R. Mullen, D. McCarthy, M. Serper, T.C. Davis, R.M. Parker, "Risk of Unintentional Overdose with Non-Prescription Acetaminophen Products," Journal of General Internal Medicine, **27**(12), 1587, 2012, doi:10.1007/S11606-012-2096-3.

[15] Adult Acetaminophen Dosage Chart | GET RELIEF RESPONSIBLY®, Apr. 2022.

[16] Ibuprofen Dosing Table for Fever and Pain - HealthyChildren.org, Apr. 2022.

[17] B. Jimmy, J. Jose, "Patient Medication Adherence: Measures in Daily Practice," Oman Medical Journal, **26**(3), 155–159, 2011, doi:10.5001/omj.2011.38.

[18] Centers for Disease Control and Prevention, Antibiotic resistance threats in the United States, 2019, Atlanta, Georgia, 2019, doi:10.15620/cdc:82532.

[19] Amazon.com: Pointells Automatic Pill Dispenser – 28-Day Portable Medication Planner and Organizer – Dispense Vitamins and Tablets Up 6 Times Per Day – Includes Flashing Light, Alarm and Safety Lock : Health & Household, Apr. 2022.

[20] Hero Health - A dose of calm for the whole family, Apr. 2022.

[21] Voice Pro, Apr. 2022.

[22] PharmAdva MedaCube™, Apr. 2022.

[23] Medipense » RxPense the best pill dispenser for seniors, chronic care, +, Apr. 2022.
[24] J. Ramkumar, C. Karthikeyan, E. Vamsidhar, K.N. Dattatraya, Automated Pill Dispenser Application Based on IoT for Patient Medication, 231–253, 2020, doi:10.1007/978-3-030-42934-8_13.
[25] K.R. Karthikeyan, E. Dharan Babu, S. Ranjith, S. Arunkumar, "Smart Pill Dispenser for Aged Patients," in 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), IEEE: 1–5, 2021, doi:10.1109/ICAECA52838.2021.9675784.
[26] R.H. Kumar, Design and Prototype of Smart Automated Pill Dispenser, VISVESVARAYA TECHNOLOGICAL UNIVERSITY, 2021.
[27] ESP32WROOM32, 2022.
[28] N. Koumaris, Using ESP32's Flash Memory for data storage - Electronics-Lab.com, Apr. 2022.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 170-177 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Smart Healthcare Kit for Domestic Purposes

Yehan Kodithuwakku[1], Chanuka Bandara[1], Ashan Sandanayake[1], R.A.R Wijesinghe[2], Velmanickam Logeeshan[*,3]

[1]*Department of Electrical, Electronic and Telecommunication Engineering, Faculty of Engineering, General Sir John Kotelawala Defence University, Sri Lanka*

[2]*Department of Mechanical Engineering, Faculty of Engineering, General Sir John Kotelawala Defence University, Sri Lanka*

[3]*Department of Electrical Engineering, Faculty of Engineering, University of Moratuwa, Sri Lanka*

ARTICLE INFO

Article history:
Received: 01 January, 2023
Accepted: 20 March, 2023
Online: 12 June, 2023

Keywords:
COVID-19
Beats per Minute
SpO2
Electrocardiogram
Telemedicine
Telehealth
Vital signs
Healthcare Kit

ABSTRACT

The COVID-19 pandemic has caused a substantial death toll throughout the world. The pandemic has created a threat to public health, the economy, food systems, and the workplace. An increased reprioritization of health expenditure towards COVID-19 vaccines will impact on allocations to other medical facilities. In developing countries, hospitals shortage the infrastructure to facilitate patients. Therefore, traditional checkups and clinics are not practical. According to research done in this article, 95 percent would prefer telemedicine and telehealth rather than conventional inspections. Even though smart healthcare technology has been implemented, it does not show adequate effectiveness when commercialized. Therefore, in this paper, a microcontroller-based, low-cost, automated, real-time system has been proposed to give a convenient solution for measuring the vital signs of the body. In this project, Multiple sensors with a microcontroller were intended to measure heartbeats per minute, Oxygen Saturation, body temperature and electrocardiogram of a patient at home without going to a hospital. The developed system indicated very less percentage error in temperature measurement and was able to maintain high accuracy on Beats per Minute, Oxygen Saturation and Electrocardiogram. This approach provides a feasible solution for both patients and medical professionals.

1. Introduction

The goal of this article is to propose a revised and expanded version of the IoT-based Healthcare Kit for Domestic Use, which was first presented at the World AI(Artificial Intelligence) IoT(Internet of Things) Congress in 2022 [1]. Additional research and testing have been carried out in order to improve the device and provide more detailed explanations of its performance and efficiency. This document summarizes the most recent findings and conclusions from the smart healthcare kit's ongoing research, development, and testing.

Since the start of the Coronavirus Disease 2019 (COVID-19) outbreak in early Wuhan in December 2019, it has spread to every country, including those with a low income [2]. It is possible that all nations will be vulnerable to this catastrophe as the world becomes a smaller, more integrated "global village." As a result, the entire world is at risk if a pandemic cannot be stopped in one nation. Not only has a pandemic had a harmful impact on the health sector, but it has also caused difficult economic, social, and political crises that, if they are not resolved quickly, will leave lasting scars [3].

Recent analysis shows that patients with COVID-19 frequently have an increase in severity or death due to related illnesses like hypertension, diabetes, obesity, cardiac diseases etc[2]. The prevalence of chronic diseases is higher in low and middle-income countries due to several risk factors, including dietary habits, physical inactivity, and alcohol consumption[4]. According to a World Health Organization study, 4.9 million people pass away from lung cancer as a result of snuff usage, 2.6 million people die from being overweight, 4.4 million people are dying from excessive cholesterol, and 7.1 million deaths occur from high blood pressure [4].

*Corresponding Author: Dr. Velmanickam Logeeshan, Department of Electrical Engineering, University of Moratuwa, Sri Lanka. Email: logeeshanv@uom.lk

https://dx.doi.org/10.25046/aj080319

Traditional examinations in specialized medical facilities were the standard for many years when it came to keeping an eye on heart rhythm, blood pressure, and glucose levels. Because of the rising global population and the increased demand for healthcare resources, there is an increasing burden on medical services. As a result, Healthcare availability in developing countries is already in crisis, and they are seeking to establish a new strategy to distribute medical resources.

In the management of COVID-19 over the world, the terms "Telehealth" and "Telemedicine" are frequently used in public and scholarly research to refer to techniques for providing healthcare remotely [5]. The provision of healthcare remotely via telecommunications technologies is known as telehealth. Telemedicine is the technique of transferring medical information from one location to another using electronic connections in order to enhance a patient's clinical health state [5].

Telehealth services can help in managing COVID-19 by screening high-risk populations and identifying probable cases, assisting with hospitalized patient care and remotely monitoring people on self-quarantine. In addition, Systems for telehealth can be used to avoid congestion in hospitals and other healthcare facilities. Telehealth can also be utilized to reduce the number of physical clinic visits for patients with chronic illnesses like diabetes or cardiac conditions who are at high risk for COVID-19-related problems [5].

By virtually tying patients and healthcare professionals together around-the-clock, seven days a week, the internet of things provides a new approach. The most crucial factors in a disease diagnosis are vital signs [4]. That is common for Chronic illnesses and Covid-19. Therefore, an inexpensive Healthcare kit on hand would save both time and money, especially who are vulnerable to above-mentioned diseases. Additionally, contactless monitoring and telemedicine capabilities ensure the safety of the both medical staff and patients against Covid-19. Therefore, home monitoring of those indicators will offer a better method for resolving those problems. The smart healthcare kit described in this paper allows both the patient and the doctor to view real-time data at any moment.

2. Literature Review

Vital signs are clinical assessments of a person's basic body functions that indicate the healthiness of their vital physiological systems. Biological aspects related to this article are mentioned below.

2.1. Vital Signs

Body temperature, Pulse Rate, respiration rate and blood pressure are the vital signs of a human. Medical officials express that vital signs such as blood pressure, temperature, heart rate, oxygen saturation(SpO_2), and respiration rate are key indicators of a patient's present health and must be routinely and precisely recorded [6]. Often, the first evidence of aberrant physiological body changes is seen in the vital signs [6].

2.2 Beats per Minute (BPM)

The number of heart contractions per minute, or heart rate, is a measure of how frequently the heart beats. A pulse is the tactile arterial palpation of the cardiac cycle in medicine (heartbeat). A suitable area to feel the pulse is anywhere on the body where an artery may be squeezed near the skin, such as the neck, wrist, groin, behind the knee, close to the ankle joint, and on the foot [7].

2.3 Oxygen Saturation

The ratio of hemoglobin currently bound to oxygen still unbound is known as oxygen saturation. There are countless numbers of these tiny air sacs in the lungs. They are crucial in bringing carbon dioxide and oxygen molecules into and out of the bloodstream. The body's tissues receive oxygen when hemoglobin binds to it during circulation.

2.4 Body Temperature

The body uses its temperature as a measure to produce and release heat. The blood vessels on the skin enlarge when someone is overheated to transfer extra heat to the skin [8]. Conversely, a person's blood vessels narrow when they become too cold, reducing blood flow to the skin and increasing heat production. A person's average body temperature may be determined via the mouth, ear, underarm, and scrotum.

2.5 Electrocardiogram (ECG)

The electrical activity of the heart during a specific time period is recorded on a graph called an electrocardiogram (ECG) using an electrograph. An ECG is a diagnostic test that examines the cardiac conduction system and provides the doctor with information about the patient's potential illness processes as well as the health of the patient's heart.

The heart, a two-stage electrical pump, can have electrodes placed on its surface to assess its electrical activity. The ECG can track the rate and rhythm of the heartbeat as well as give an imprecise indication of blood flow to the cardiac muscle [7]. The PR and QT intervals are the two fundamental intervals of an ECG. The fundamental ECG segments are shown in Figure 1.

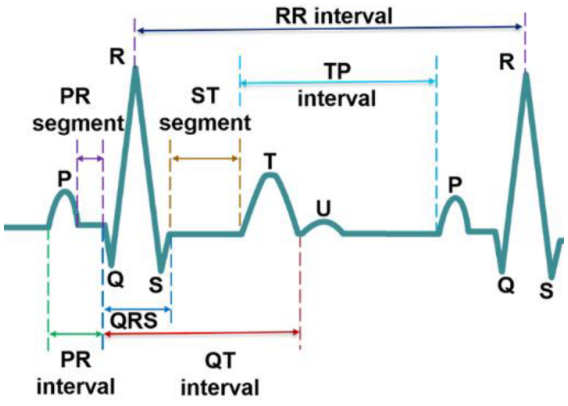


Figure 1: Standard ECG Pulse. Reprinted with permission from [9]

Currently, hospitals use appropriate devices to keep track of their patient's health. However, doctors must be present in such situations. Researchers came up with various IoT models for

healthcare and forecasted various diseases to address this issue. The following are some of the research works done by the authors.

2.6 Wireless Sensor Network-based Smart Healthcare

In the proposed system, a Body Area Network is created by attaching a variety of wireless sensors to the patient's body to measure important body metrics like temperature, blood pressure, and heart rate. The data is wirelessly gathered and shown on the Patient Bed Monitor (PBM). In addition, the room Server (RS) is connected to it for data archiving and analysis [10].

Three tiers make up the suggested system architecture. Tier-1 comprises wireless sensor nodes fastened to the patient's temperature and pulse rate to monitor essential body functions. Tier 2 is the intermediate receiving unit that receives the data that has been transmitted. Alert systems and data transmission over larger distances using the right internet connections are the focus of Tier 3 [10].

2.7 Raspberry-pi based smart healthcare

Sensors have been connected to the proposed system in appropriate ways. The unit integrates the data from the sensors with the board after receiving it from the sensors [11].

This device connects with the temperature, heart rate, ECG, acceleration, and pressure sensor. In addition, the user's and doctor's devices are connected to the internet, and the generated results are shown on a Liquid Crystal Display(LCD) monitor at regular intervals [11].

There are two ways to connect and operate the raspberry device: the first is directly attaching peripherals and the second is connecting the computer after installing the putty program with a IP address, subnet mask, and gateway to that system [11]. If any irregularities in the patient's health are detected, they will be promptly reported to the authorized or guardian via the Global system for mobile communication (GSM) via the network.

2.8 Arduino-based smart healthcare

The design considered detecting temperature, humidity, and blood pressure. Remote patient monitoring is possible for the aforementioned metrics by medical personnel. The goal of this project was to build a patient health monitoring system that could measure ECG, blood pressure, pulse rate, and temperature[4]. The “ThingView” application displays the ECG, temperature, and pulse rate characteristics visually, and the readings are conveyed to a phone through SMS(Short Message Service).

2.9 GSM-based smart healthcare

Sensors are used to measure vital signs, and Zigbee and GSM are also used to send the parameters wirelessly. An SMS is sent to the doctor in an emergency. In an emergency, a SMS is delivered to the doctor. The patient's family and the medical community are the key users of the GSM-based health monitoring system [12]. On the “Thingspeak” website, the results are displayed. The output is produced using the sensor data. An LCD shows the computed heartbeat rate.

2.10 Shortcomings of the available products

The majority of the aforementioned versions do not monitor BPM, SpO2, body temperature, and ECG at the same time. Currently, available items require a significant amount of power to operate. The various models' precision was insufficient for accurate measurements [13]. Despite the plans being implemented, they were unable to distribute the final products because of the absence of internal resources and a late market launch. These architectural designs may have some privacy issues. People might not want to be completely watched over. They can be concerned that crucial information, including their whereabouts or health status, could be disclosed to other parties with the proper authorization [14].

This project takes an innovative approach to health monitoring by creating a small and affordable system that can measure ECG, blood oxygen, BPM, and body temperature. While similar tools are available, they are frequently pricey and lack the versatility required for daily use. Existing products for measuring ECG, blood oxygen, BPM, and body temperature have limitations that make them unsuitable for usage in the home. For instance, hospital equipment is large and requires qualified personnel to operate. Similarly, most home-based monitoring systems are inaccurate and have a limited feature set. The kit produced in this study, on the other hand, solves these disadvantages by adding modern sensors and microcontrollers which have better accuracy on the body parameters. The kit is intended for home usage, which is both affordable and portable, making it ideal for remote patient monitoring.

3. Product Survey

Several data-gathering approaches were used to get information from the general public and doctors on the smart healthcare kit for home use.

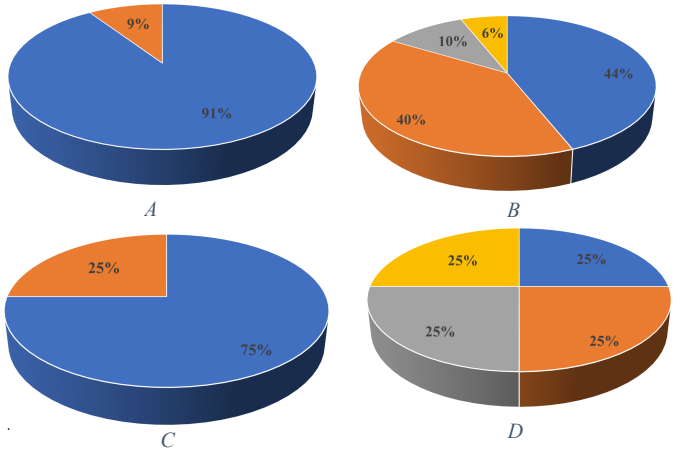


Figure 2: A,B for general people C,D for doctors

3.1 Study for the general public

The authors conducted the research using a Google form. The study enlisted the participation of 120 persons. The experiment lasted for three weeks. Questions that had been asked for the survey are attached below.

A. Do you prefer to test your vital signs at home rather than visiting a hospital, and consult a doctor online when you are sick?
B. What are your thoughts on the 'smart-Medicare kit for home usage'?
C. Are you willing to monitor patients' vital signs online, especially in the midst of a pandemic?
D. Are you likely to be able to diagnose an illness by taking a patient's temperature, pulse, SpO2, and ECG?

Figure 2 chart A illustrates that when suffering from a disease, 91% of individuals would prefer to evaluate their vital signs at home than go to a hospital and visit a doctor online. However, 9% would consider going to a hospital physically.

Figure 2 B displays the general public's perception of the Smart Medicare kit for household use. Almost 45 percent of individuals thought it was a great idea. Approximately 40% voted as 'Good'. Nevertheless, around 6% think that the idea of a healthcare kit is poor.

3.2 Survey for doctors

If the patient is in critical condition, the system sends an emergency message to a healthcare professional. To calculate vital sign threshold values, a Google form is given to 50 family doctors. Numerous interviews were also done in order to acquire critical information, such as selecting the optimal places of the body to achieve the most exact results. Additionally, the user handbook includes all of the best locations for sensor placement. According to Figure 2, C, almost 75% of doctors are ready to monitor vital signs remotely.

Figure 2, D shows that about a quarter of doctors would be able to detect a basic illness by evaluating the kit's measurement. Half of those polled believe they can diagnose a sickness to some amount by reading values.

3.3 Survey on Hospitals

Telehealth has been available in the country for a decade, but adoption among the general public has been modest [5]. The Information and Communication Technology Agency of Sri Lanka inaugurated the first IoT project in Sri Lanka in 2015, with 25 state hospitals participating. Following it, various numerous were carried out over the nation [15]. The Ministry of Health has also created mobile applications to help increase Telehealth services in Sri Lanka [5]

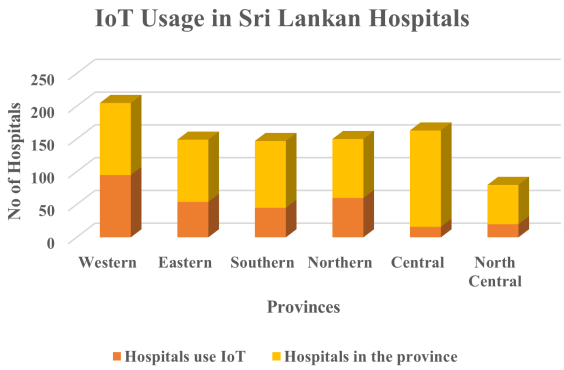


Figure 3: IoT Usage in Sri Lankan Hospitals

An analysis was conducted to determine IoT usage in Sri Lankan hospitals. Personal communications and statistical records were used to obtain data from the Ministry of Health and the Health Informatics Society of Sri Lanka. Figure 3 shows that IoT technology is utilized in practically every province. Medical data management systems, Mhealth apps, and personal health numbers are some examples of IoT applications used in Sri Lankan hospitals [15]. As a result, hospitals will find it easier to integrate IoT.

According to studies, hospitals may leverage Internet of Things technology to improve the efficiency of healthcare services. IoT technology is transforming the healthcare industry; by incorporating IoT, hospitals may enhance professional efficiency while also offering higher-quality medical treatment. Remote patient monitoring technologies allow medical personnel to minimize fatigue while saving patients the trouble of traveling. This is a huge advantage, particularly for persons with restricted mobility.

4. Flowchart

First and foremost, the Wi-Fi name and Passcode should be supplied in the code. The NodeMCU will attempt to link up to Wi-Fi after connecting the micro-USB to the PC. The Message Queuing Telemetry Transport protocol will be attempted if the device is connected.

The system will then check the MAX30102 sensor's availability. If the sensor is present, the kit begins reading data from the MAX-30102, AD-8232, and Temperature sensors. It will then sort out and upload the records to the Serial Monitor and the "Ubidots" cloud platform.

Lastly, any deviations from previously defined threshold values will be detected by the microprocessor. If any of the abnormalities are found, the system will send a message to a doctor based on previously given data, as shown in figure 4.

5. Methodology

The project's purpose is to create a smart Medicare system that can alert a doctor when a patient is in a severe condition. The project's aims are to assess a person's BPM, SpO2, and temperature from home if the person is unwell, to plot an ECG and have it evaluated by a doctor if the person is in a critical

condition, and to safeguard both patients and medical staff against Coronavirus.

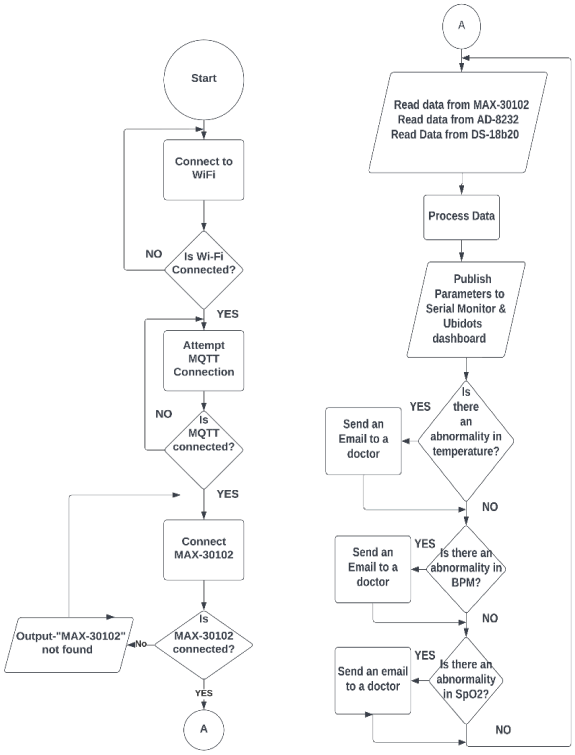


Figure 4: Flowchart of the proposed model

The esp-8266 offers various functionalities that are extremely useful when working with IoT. The Wi-Fi antenna on the module allows embedded devices to connect to networks and send data. Figure 5 depicts the integration of three sensors. Furthermore, the processer handles fundamental inputs from analog and digital sensors in order to perform significantly more sophisticated calculations. As the “Ubidots” cloud platform is accessed here, the esp-8266 can browse websites written in Hyper Text Markup Language(HTML) or any other development language.

The process began with the purchase of required circuits from shops. Individual codes were then written for each circuit. Following that, all circuitries were rigorously tested for flaws and defects to ensure precise and reliable performance.

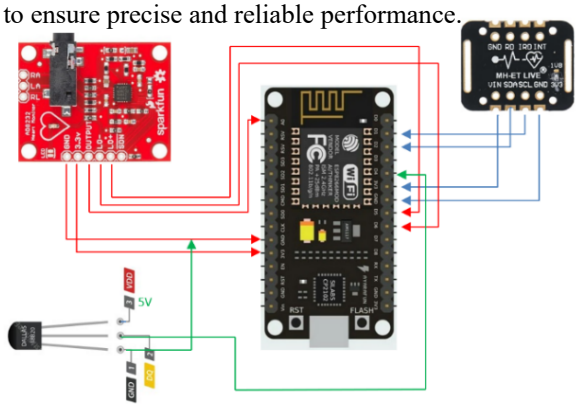


Figure 5: Wiring Diagram

The different circuits were tested and validated before being merged onto a single breadboard, as shown in Figure 5. The codes created for each circuit were concatenated to form a single code. This code was then transferred to the esp-8266 module and thoroughly tested for accuracy and faults. To achieve precise readings, special steps were taken, such as configuring the max30102 sensor to take an average of four data points and setting the AD8232 sensor to record 400 samples per second.

"Ubidots" was chosen as the cloud service for the application. To begin, each sensor's data was transmitted to the cloud service and reviewed for errors. Second, the sensors were combined, and data was transferred to the cloud. Real-time data was used to discover faults.

A poll of doctors was conducted to determine the temperature, pulse rate, and blood oxygen level thresholds. If the measured values deviate from the threshold value, an email is automatically sent to the relevant family doctor. This was done through the “Ubidots” website. This cloud platform housed medical records. The ECG was slightly off due to network latencies. Hence, the readings from an ECG were exported to Microsoft Excel utilizing the data streamer option. A caretaker then plotted the graph and sent it to a doctor.

The printed circuit board was designed using Autodesk-Eagle. Separate footprints were created for AD-8232 and MAX-3012. The schematic diagram was then created. Third, all of the components were carefully arranged to reduce the length of the copper cables. After the placement, the routing was completed. The trace width was calculated using ‘EE web PCB trace width calculator online tool. Finally, the Gerber files were created and delivered to the manufacturer.

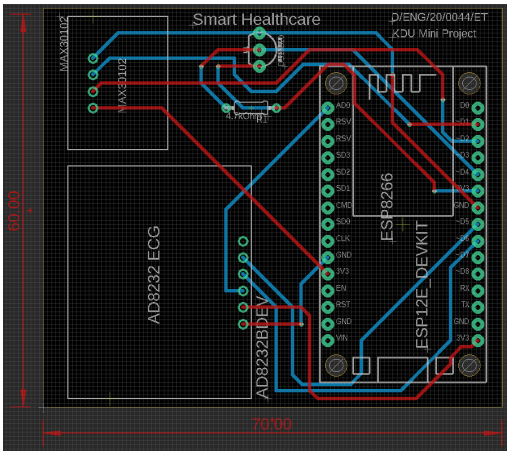


Figure 6: PCB (Printed Circuit Board) Design

The PCB was received and tested for quality after a three-week wait. On the PCB, female header pins were attached, and all sensors were connected to these pins. The sensors' operation was checked and verified when they were put into the board. A second Vero board was required to fit the Max-30102 sensor. To provide reliable readings, the sensor was positioned on top of the plastic stage and isolated from the other sensors to reduce interference. Since the photoplethysmography (PPG) technique is used in

MAX-30102, the sensor was positioned in a dark area to reduce noise and improve measurement accuracy. The configuration was double-checked to confirm that all sensors were working properly.

Figure 7: End Product

Finally, the 3D model was created with AutoCAD. The 3D model was then fed into a laser cutter, which sliced the plastic into the desired shape. Finally, everything was fitted together to make the final product seen in figure 7.

6. Validation of Experimental Results

The complete results of the model implementation are shown below. The model's output was validated using a temperature and an oximeter. The experiment was carried out on a healthy middle-aged man.

6.1 Body Temperature

Below equation was used to calculate the percentage error.

$$Percentage\ error = \frac{|actual\ value - expermintal\ value|}{expermintal\ value} \times 100\%$$

Equation 1 percentage error

$$Percentage\ error = \frac{|35.7-35.005|}{35.005} \times 100\% = 1.9\%$$

The percentage of body temperature was calculated, as shown in the preceding equation. The experimental value was determined as the average of seven consecutive 30-second observations. The exact value was determined by averaging seven successive "aeg FT4904" thermometer readings taken within 30 seconds.

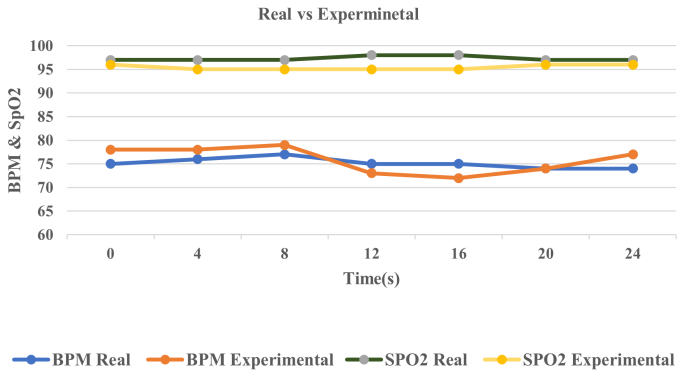


Figure 8: Experimental values from the kit vs real values from the oximeter

6.2 Bpm & SpO2

The real and experimental data were displayed for a time of 24 seconds, as illustrated in figure 8. Real measurements were taken with a "ROHS ABH23" fingertip pulse oximeter. The X-axis depicts the time with four consecutive periods, while the Y-axis reflects BPM and oxygen saturation. There is a slight difference between the real and experimental numbers.

6.3 Electrocardiogram

Figure 9 depicts a tested graph taken from the system. When compared to a standard graph (figure 1), the key segments (PR,ST) and intervals (PR,QT) can be seen. Likewise, it is observed that the experimented data has high accuracy.

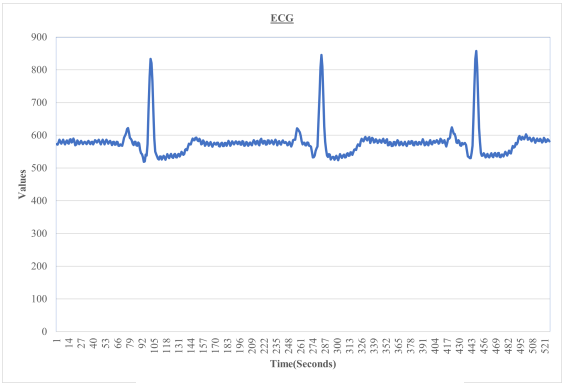


Figure 9: Obtained ECG from the kit

7. Discussion

7.1. Strengths

The model has a number of strengths. The device can assess body temperature, pulse rate, oxygen level, and ECG. (Four in one). As a result, there is no need to acquire a pricey oximeter, thermometer, and ECG machine individually (Cost-effectiveness). If the values differ, it automatically sends a notification to a doctor. Furthermore, it consumes less electricity and can be carried anywhere (Easily equipped). Finally, the model saves time by storing past data.

7.2 Accuracy Improvements

Accuracy enhancement approaches were used to decrease the constraints. For SpO2 and BPM, an average of four data points were collected, while 400 samples were collected every second for ECG and the application of filtering algorithms.

7.3 Issues experienced

The first issue encountered was with the MAX-30102. Pullup resistors were the source of the problem. To make it pull up, they were connected to 1.8v. The microcontrollers, on the other hand, ran on 3.3V. As a result, the three pull-up resistors were taken out with a soldering iron.

The second issue was cloud-based monitoring of the ECG. The real-time graph varied from its conventional graph due to network latency issues. As a response, the values were exported to Microsoft Excel, where the graph was created.

7.4 Price Assessment

The smart healthcare kit was expected to cost 8500.00 LKR in total. In contrast, purchasing the medical equipment individually would cost roughly 50,000 LKR. As a result, there is a large amount of difference that is financially favorable.

7.5 IoT Security Design

Because of the fast development and integration of the Internet of Things in various sectors of our everyday lives, Internet of vulnerabilities has emerged. Cyber-attacks risk the security and privacy of IoT device users. IoT devices don't have the adequate processing power, energy, storage capacity, and memory[16]. The best method for protecting communications between IoT devices is still being developed.

Security issues are a continual threat to IoT. Because of a large number of Internet-connected IoT devices, the Internet of Vulnerabilities has emerged. "Symantec" reports stated that IoT threats increased 600% in 2017[16]. The criteria for safeguarding IoT systems are authentication, authorization, communication, and device identification. The most secure authentication techniques concentrate on three characteristics: network assumptions, communication sessions, and users.

The "Ubidots" login information is unique to each individual. The password is only known to one person. This is how authentication is protected. The account token, according to the proposed technique, should be sent to someone's family doctor so that they may verify their live data. As a result, only those people have access to particular information.

Table 1: The comparison of vital signs to the existing systems

Vital signs and systems	Body Temp. (°C)	BMP (BPM)	SpO2 (%)	ECG
Proposed kit	35.005	76	96	Similar to fig1
Available system-1 [17]	28	86	76	Vertically shifted
Available system-2 [18]	29.81	108	63	unclear
Available system-3 [19]	30.45	69	-	Dissimilar to fig1

Table 2: Percentage errors of the existing systems

	Percentage error of Body Temp(%)	Percentage error of BPM(%)	Percentage error of SpO2(%)
Proposed Kit	5.6	5.2	1.0
system-1	32.21	16.27	27.6
system-2	24.11	33.33	53.9
system-3	21.5	4.3	-

To assess the efficacy of our IoT-based healthcare equipment, the accuracy of industry benchmarks was compared. As presented in table 1 BPM, SpO2, ECG, and body temperature results were compared to the previous similar studies.

According to the findings in table 2, the proposed healthcare kit had fewer percentage errors. The percentage error was calculated according to equation 1. Actual values were considered as 37°C, 72 and 97% for body temperature, BPM and SpO_2 respectively which are general truths. When compared to currently commercialized products, our healthcare kit outperformed the above systems in terms of body temperature and SpO_2 accuracy. The ECG of the available system 1 was shifted vertically. Therefore, the vertical measurements were affected. ECG of the available system 2 was unclear and the diagnosis procedure is quite hard. By mitigating the drawback of each existing product mentioned in 2.10, these findings indicate that our healthcare kit has the potential to deliver precise and accurate vital sign measurements for home users.

8. Future Directions and Concluding Remarks

Future enhancements include adding blood pressure equipment to the smart healthcare unit, developing a model utilizing Machine Learning Algorithms to assess ECG data, and sending an automated message to an ambulance service. Furthermore, security enhancements should be implemented to tighten privacy and reduce cyber-attacks.

Artificial Intelligence (AI) in healthcare combines computation with the viewpoints of doctors to build better healthcare technologies. Using NLP, knowledge representation, automated reasoning, and machine learning, AI attempts to make computers more practical and intelligent. Understanding patient health status by a pre-trained model from symptoms could be a future advancement.

Some of the product's primary features include keeping a patient's medical records, alerting a doctor in an urgent emergency, four in one, and many more. The combination of these features makes it a great tool for both individuals looking to monitor their health from the comfort of their own homes and healthcare professionals looking for an accurate solution for remote patient monitoring.

The goal of this research is to develop an IoT-based smart healthcare kit for people suffering from Covid and Chronic diseases. The project was completed and tested with people successfully. We'd like to thank everyone who took the time to fill out the survey. The information obtained was useful in determining a continuing problem and a long-term remedy.

References

[1] Y. Kodithuwakku, A.D. Sandanayake, C. Bandara, V. Logeeshan, "IoT Based Healthcare Kit for Domestic Usage," 2022 IEEE World AI IoT Congress, AIIoT 2022, 760–765, 2022, doi:10.1109/AIIoT54504.2022.9817235.

[2] A.K. Singh, A. Misra, "Impact of COVID-19 and comorbidities on health and economics: Focus on developing countries and India," Diabetes and

Metabolic Syndrome: Clinical Research and Reviews, **14**(6), 1625–1630, 2020, doi:10.1016/j.dsx.2020.08.032.

[3] E. Workie, J. Mackolil, J. Nyika, S. Ramadas, "Deciphering the impact of COVID-19 pandemic on food security, agriculture, and livelihoods: A review of the evidence from developing countries," Current Research in Environmental Sustainability, **2**, 100014, 2020, doi:10.1016/j.crsust.2020.100014.

[4] K. Latha, G. Sravanth, N.V.P. Kumar, ..., "An IoT based patient monitoring system using arduino uno," … Research Journal of …, 2020.

[5] G.G. Kulatunga, R. Hewapathirana, R.B. Marasinghe, V.H.W. Dissanayake, "A review of Telehealth practices in Sri Lanka in the context of the COVID-19 pandemic," Sri Lanka Journal of Bio-Medical Informatics, **11**(1), 8, 2020, doi:10.4038/sljbmi.v11i1.8090.

[6] W.Q. Mok, W. Wang, S.Y. Liaw, "Vital signs monitoring to detect patient deterioration: An integrative literature review," International Journal of Nursing Practice, **21**(S2), 91–98, 2015, doi:10.1111/ijn.12329.

[7] R. Mahajan, D. Bansal, "Identification of heart beat abnormality using heart rate and power spectral analysis of ECG," International Conference on Soft Computing Techniques and Implementations, ICSCTI 2015, 131–135, 2016, doi:10.1109/ICSCTI.2015.7489555.

[8] M.U.H. Al Rasyid, B.H. Lee, A. Sudarsono, "Wireless body area network for monitoring body temperature, heart beat and oxygen in blood," 2015 International Seminar on Intelligent Technology and Its Applications, ISITIA 2015 - Proceeding, 95–98, 2015, doi:10.1109/ISITIA.2015.7219960.

[9] L. Xie, Z. Li, Y. Zhou, Y. He, J. Zhu, "Computational diagnostic techniques for electrocardiogram signal analysis," Sensors (Switzerland), **20**(21), 1–32, 2020, doi:10.3390/s20216318.

[10] U. Gogate, J.W. Bakal, "Smart Healthcare Monitoring System based on Wireless Sensor Networks," International Conference on Computing, Analytics and Security Trends, CAST 2016, 594–599, 2017, doi:10.1109/CAST.2016.7915037.

[11] N. Arunpradeep, G. Niranjana, G. Suseela, "Smart healthcare monitoring system using iot," International Journal of Advanced Science and Technology, **29**(6), 2788–2796, 2020, doi:10.22214/ijraset.2020.5101.

[12] K. Aziz, S. Tarapiah, S.H. Ismail, S. Atalla, "Smart real-time healthcare monitoring and tracking system using GSM/GPS technologies," 2016 3rd MEC International Conference on Big Data and Smart City, ICBDSC 2016, 357–363, 2016, doi:10.1109/ICBDSC.2016.7460394.

[13] A. Rahaman, M.M. Islam, M.R. Islam, M.S. Sadi, S. Nooruddin, "Developing iot based smart health monitoring systems: A review," Revue d'Intelligence Artificielle, **33**(6), 435–440, 2019, doi:10.18280/ria.330605.

[14] D. Ding, M. Conti, A. Solanas, "A smart health application and its related privacy issues," Proceedings of the 2016 Smart City Security and Privacy Workshop, SCSP-W 2016, 11–15, 2016, doi:10.1109/SCSPW.2016.7509558.

[15] R. Rajmohan, P.D.D.M.G.M. Johar, "Adoption of the Internet of Things in the Healthcare Services of Sri Lanka," International Journal of Recent Technology and Engineering (IJRTE), **9**(1), 1095–1104, 2020, doi:10.35940/ijrte.a2260.059120.

[16] M.A. Obaidat, S. Obeidat, J. Holst, A. Al Hayajneh, J. Brown, "A comprehensive and systematic survey on the internet of things: Security and privacy challenges, security frameworks, enabling technologies, threats, vulnerabilities and countermeasures," Computers, **9**(2), 2020, doi:10.3390/computers9020044.

[17] A.D. Acharya, S.N. Patil, "IoT based Health Care Monitoring Kit," Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020, (Iccmc), 363–368, 2020, doi:10.1109/ICCMC48092.2020.ICCMC-00068.

[18] M.R.R. Akash, Yousuf, K. Shikder, "IoT Based Real Time Health Monitoring System," Proceedings of International Conference on Research, Innovation, Knowledge Management and Technology Application for Business Sustainability, INBUSH 2020, 167–171, 2020, doi:10.1109/INBUSH46973.2020.9392163.

[19] M.R. Ruman, A. Barua, W. Rahman, K.R. Jahan, M. Jamil Roni, M.F. Rahman, "IoT Based Emergency Health Monitoring System," 2020 International Conference on Industry 4.0 Technology, I4Tech 2020, 159–162, 2020, doi:10.1109/I4Tech48345.2020.9102647.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 6, 189-194 (2022)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

A Review of the Role of Information Technology in Brazilian Higher Educational Institutions during Covid-19 Pandemic

Luís Cláudio Dallier Saldanha*

Universidade Estácio de Sá, Department of Education, Rio de Janeiro, 20071-001, Brazil

ARTICLE INFO

Article history:
Received: 28 October, 2022
Accepted: 30 November, 2022
Online: 20 December, 2022

Keywords:
Remote teaching
Educational technology
Technological mediation
Documentary research

ABSTRACT

This paper presents the results of a documentary research on the use of information technology in emergency remote teaching in 66 higher educational institutions in Brazil. The theoretical background of this study is based on the works of Feenberg, Bagglaey, Veloso & Mill, Castañeda & Selwyn and Hodges. The methodological approach consisted of analyzing reports published by YDUQS, an educational holding responsible for managing all the 66 institutions examined in this research. Such analysis aimed at identifying data concerning investments in information technology and its use throughout the Covid-19 pandemic. Results have revealed that investment in information systems as well as technological mediation of academic routines and pedagogical practices paved the way for a rapid response to the crisis triggered by the pandemic and the maintenance of student satisfaction. Nevertheless, the data available within the reports was not enough to draw conclusions on learning management neither on other pedagogical aspects of emergency remote teaching.

1. Introduction

Remote teaching was the educational alternative when in-person pedagogical activities came to a halt with the social distancing imposed by the COVID-19 pandemic during the years of 2020 and 2021.

In Higher Education, the prevalent trend was the adoption of technological solutions based on videoconference platforms and virtual learning environments, in which synchronous communication in online pedagogical activities was predominant with the shift from the four walls of the conventional classroom to digital settings.

Some institutions that already possessed experience with remote or hybrid learning were able to rely on existing digital content, intensive use of technology, technological infrastructure, and more appropriate methodology to promote both real-time classes as well as opportunities for the development of pedagogical activities in virtual environments.

In light of this context, it is fitting to put under scrutiny the implications of these institutions' total dependence on technological mediation so that classes and pedagogical activities could be preserved during the Covid-19 pandemic.

The present research summarizes the analysis of data on the technological mediation of emergency remote teaching in these 66 Brazilian institutions of Higher Education, managed by the same corporate group.

Thus, the results stem from documentary research aimed at examining the educational consequences of information technology within the context of remote learning.

This work was originally presented in the 17th Iberian Conference on Information Systems and Technologies [1]. This introduction is followed by a theoretical background on e-learning and information technology in the field of education, a section on methodological aspects, and, at last, the results of the present research on the investment and on the use of technologies within the higher educational institutions of the YDUQS group.

2. Remote Teaching and Technological Mediation

2.1. Remote Teaching

Expressions such as *remote teaching* (RT), *emergency remote teaching* (ERT) and emergency remote learning and teaching (ERLT) have become recurrent in contemporary literature to describe online pedagogical activities during the Covid-19 pandemic.

*Corresponding Author: Luís Cláudio Dallier Saldanha, luis.dallier@ensineme.com.br

Initially, in [2] the authors argued that emergency remote teaching (ERT) should not be taken as a byword for online learning or e-learning, while other authors [3] understand remote teaching as equivalent to online learning or e-learning.

The perspective that underscores differences between remote learning and e-learning have prevailed in recent literature on this topic, with only few studies still endorsing their interchangeability. To some extent, the need to set these two concepts apart is, indeed, justified so that the reputation of e-learning is not tainted by the chaos, the improvisation, the frail theoretical background, and the lack of suitable methodologies to cater for students' profiles due to the abrupt adoption of remote learning during the Covid-19 outburst [4].

In addition, their distinctiveness is also be sustained by the fact that private higher educational institutions charge cheaper fees for e-learning, whereas remote learning implicates more costs, with teachers offering synchronous lessons.

In a nutshell, the dissimilarities between remote teaching and e-learning can be described in terms of three major traits of remote teaching: a) the urgent and temporary status of remote learning; b) the transposition of in-person classes into virtual environments; and c) the prevalence of synchronous communication via real time transmission of lectures and video classes.

In turn, e-learning relies on five fundamental prerogatives: a) it must count on didactic and pedagogical frameworks of its own; b) educational contents and activities must be adequately designed; c) pedagogical model and methodologies must cater for students' needs and take their profiles into account; d) students must be familiar with the methodology as well as its technological resources; e) efficient tutoring must supervise and support students' academic performances [5].

In [6] the authors prefer not to set apart the concepts of remote teaching and e-learning based on the assumption that there are more similarities than differences between these two educational modalities. In addition to intrinsic dependence on technology, remote teaching shares other characteristics with e-learning, since in both cases teachers and learners are physically separated and the learning process is mediated by technology.

Furthermore, both modalities may alternate between synchronous and asynchronous communication, virtual environments and videoconference functionalities or application programs. The prevalence of synchronous communication and videoconference platforms in remote teaching would, therefore, consolidate a type of e-learning, instead of an independent educational modality. Due to prolonged social distancing, educational institutions planned remote teaching or a transition to hybrid teaching, developed specific methodologies and contents for these scenarios, in addition to acquiring suitable didactic materials.

Either standing as an independent educational modality or constituting a type of e-learning, remote teaching during the Covid-19 pandemic must be analyzed in light of the many particularities that marked the uniqueness of this period.

2.2. Information Technology and Education

The use of technology for educational purposes is not limited to information systems nor to the era of digital technologies. In a broad sense, technology has been omnipresent in education ever since resources such as chalk, blackboards, books, pens, and pencils were used in the classroom. In other words, technology in the field of education includes far more than computers and mobile digital resources [7]. That is not to deny the huge impact of digital technologies on educational processes, but to acknowledge the longstanding role technology has played in pedagogical mediation over the course of history.

Thus, the use of technology in remote teaching during the Covid-19 pandemic is necessarily intertwined with the broader history of technology in education – and there are many lessons to be learned from such previous experiences.

Many innovations advertised by the EdTechs date back to projects and experiences developed between the 1920s and the 1950s [8].

The infrastructure in information technology (IT) and the use of diverse technological resources have favored not only teaching in situations in which teachers and students are separated by time and space, but also brought forward alternative (and often more interactive) pedagogical practices from which any educational modality can profit.

New digital technologies have become more interactive and user-centered, offering new possibilities within educational settings [9].

In [10] the authors had already identified technological trends focused on ubiquitous and networked experiences with the massification of mobile devices and the further developments of the Web with novel and diversified forms of representation, stimulating environments, and a global IT infrastructure combining decentralization and interoperability.

In this scenario, information technology has become both integrating and pervasive, which explains its ubiquity in the lives of teachers and students as well as the emergence of demands and possibilities regarding its appropriation and use in formal educational contexts.

Information and communication technologies are not solely devoted to the production and the availability of digital contents in different media and languages. In addition to granting access to texts and video classes, these technologies have allowed users to come up with new ways to represent data, concepts, processes and phenomena. Simulations, animations, and games have propelled more meaningful learning experiences. Virtual and augmented realities have also proved efficient in the educational realm.

Moreover, due to recent advances in terms of artificial intelligence, technology has given rise to new forms of pedagogical planning, student-tailored learning experiences, learning management, as well as innovating tools and procedures for assessment.

Therefore, technology has gone far beyond its primary function of guaranteeing remote interaction among teachers and students, surpassing the mere offer of online education to produce new methodological approaches – useful for both in-person as well as remote learning experiences.

Technology has made feasible the mediation or the rise of different forms of communication, interaction and relationship within academic and educational processes.

Thus, it is possible to promote technological mediation within any sort of pedagogical modality, whether in-person or online, even though technologically mediated education has been mostly associated to e-learning. However, this setting has rapidly changed because of the Covid-19 pandemic.

3. Dependence and Resistance to Technology

3.1. Technological dependence

Social distancing during the Covid-19 pandemic imposed technological dependence as the only means of pedagogical mediation for previously in-person Higher Educational courses. Technological dependence was not restricted to the use of digital platforms nor to the videoconferences that suddenly replaced conventional classrooms. In addition to substituting face to face classes with virtual encounters, it was essential to sustain the entire academic routine digitally, making room for other pedagogical activities in the online environment.

At first, part of Brazilian higher educational institutions halted their academic activities for longer periods – some for a few months – which tended to be a more recurrent choice amongst public institutions [11]. Other institutions, mostly private, resumed their academic activities in virtual environments sooner, some improvising more than others. Learning Management Systems (LMS), once confined to e-learning, were adopted as well as applications for videoconference.

Initially, their technological response was most frequently based on platforms or applications responsible for transmitting live-expositive classes, often lacking previous planning and a more careful design process [12]. Services related to academic activities were preserved and especially enhanced by institutions that already counted on online customer service channels, which allowed students to remain in close contact through the use of applications, in addition to providing support and guidance throughout the implementation of remote classes.

So, institutions which had previously invested on a more intensive use of technology in their pedagogical activities and educational services were able to respond more rapidly and more efficiently to the sanitary crisis.

These institutions managed to increase their investments on technology, paving the way for necessary improvements and for the development of information systems capable of handling an increased demand.

Those that lacked prior experiences with technology in face-to-face classes or that did not have online customer services tended to improvise in their attempts to adjust to the crisis, and often resorted to the suspension of pedagogical activities for a few months or even for the entire term.

Such evidence revealed the technological dependence of these higher educational institutions and their need to invest in IT within the contexts of e-learning as well as in-person classes – the latter adapted to hybrid formats due to pandemic restrictions.

3.2. Between resistance and adoption: the technological dilemma

The strong necessity to resort to technology in the field of education during the Covid-19 pandemic heightened tensions: on one hand, it intensified resistance to technology; on the other, it endorsed the unescapable need to adopt it. This dilemma can be described as a conflict that opposed those who saw online education as a natural (and necessary) development of human communication and those who criticized it for automating and mechanizing the learning process.

To fully understand this debate, it is worth analyzing the controversies that marked the rise of distance education mediated digital technologies, especially the pioneering contributions of Andrew Feenberg to philosophy of technology in the United States.

In the early 1980, distance education chiefly relied on printed materials sent to students and on one-way communication via radio, television, and satellite transmission. Back then, Internet was not an option for the general public and electronic mail was still mainly restricted to computing companies and universities developing research on the new technology.

The first program for online education was created when computers were still regarded as devices for data organization and mathematical calculation. Nevertheless, the use of computers in the realm of education helped pave the way for their reinvention as means of communication [13].

While narrating his experiment with online education, in [13] the author describes that the invention of e-learning was aimed at providing a human interface for distance education, which basically consisted of mailing printed didactic materials to learners.

Feenberg's pioneering experiment lasted for about ten years and was initially characterized by difficulties posed by technological limitations of the time: students, for example, had to flawlessly carry out an entire page of commands just to log in the system. It was also necessary to create a new software just to guarantee asynchronous interactions, such as the simple exchange of messages.

In [13] the author distinguished his demonstrations from the increased interests in large-scale distance education in the 1990s, when the financing crisis that hit universities in the United States motivated the adoption of digital technologies and the choice of the Internet to offer and organize online courses. These attempts in the field of online education sought automating learning through the use of the Internet and completely eliminated classroom interactions.

In [13] the author remembers that David Noble, the Marxist historian that denounced the loss of skilled workers to industrial automation, had become the main critic of online education, and joined him in several debates on the vices and virtues of the new system.

The consolidation of online education had to overcome at least two major challenges: the first stemmed from humanist criticism, which basically rejected any sort of electronic mediation in education; the second came from technocrats, looking forward to

completely eliminating the roles of teachers from the educational scenario. What humanist criticism and technocrat approaches to education brought forward was a deterministic understanding of e-learning as either a dehumanizing modality or a profitable business opportunity [13].

Contrary to such deterministic understandings of technology in the educational realm, the instrumentalist approach conceives technology as a neutral tool – either good or bad depending on its use.

Based on this instrumentalist perspective, adopting technology may enhance interaction and the learning experiences itself, mainly in the field of e-learning. The use of technology is seen as unavoidable, after all technology represents innovation and the new forms of communication.

Beyond the pessimistic resistance or the optimistic and unrestrained adoption of technology, it is possible to implement a critical approach that recognizes both the possibilities of expanding and enriching the learning experience as well as the risks of undermining and reducing the educational practices through technological mediation.

Raising critical awareness on the possibilities and shortcomings of technology is crucial within an educational approach mediated by technological resources. Technology should not be understood as an end itself and its uses must be molded in accordance with the objectives of each pedagogical project.

On the other hand, technology should not be seen as the mere means through which an educational objective can be accomplished: technology itself is embedded in its own social, cultural and economic aspects. In other words, technological appropriation demands acknowledging and addressing sociocultural tensions between the educational and the technological realms.

It is essential to recognize the vital influence of the digital technology industry and the big corporations on the molding of educational policies and on the craft of higher educational thought. In [14] the authors argue that this industry has kept on pushing higher educational reforms partially under the guise of necessary help in times of crisis.

Therefore, the data and the indicators related to technology and education should always be read from this critical viewpoint.

4. Method and Analysis of the Results

4.1. Procedures and data obtained

So as to investigate the impact and the results of technological mediation in remote teaching, documentary research was carried out based on data collected by one of the largest higher educational groups in Brazil.

The reports and documents analyzed were: Results Reported 1 T20; Results Reported 3T20; Results Reported 1T21; Results Reported 2T21 & 1S21; YDUQS Ecosystem; and Business Units.

This data was obtained through reports and other documents listing information, indicators as well as operational, academic, and financial analyses of the educational group between 2020 and 2021.

The group identified as YDUQS Participações S.A., a technological holding in the field of education, pulling together a board of trustees comprising 66 higher educational institutions, distributed throughout 52 cities in every state of Brazil. All of them are private institutions characterized according to the Brazilian legislation as non-profit organizations.

The group is listed in the Novo Mercado da B3 (the Brazilian Stock Market) as YDUQ3 and, has also gotten its ADRs (American Depositary Receipts) traded in the North American market under the name YDUQY.

In the beginning of the pandemic, in March 2020, the group comprised 319,000 students in face-to-face courses and 314,000 students in distance education [15].

Since 2018, YDUQS has increased its investments in technology. Nearly half of it was devoted to digital transformation and to enabling technologies.

More than 80% of all procedures related to academic routine and other transactions are carried out digitally aided by applications designed specifically for students and teachers. The application programs are available in any application store and have been rated very positively by their users.

In its digital ecosystem, YDUQS has made remarkable progress in its digital transformation throughout the pandemic and offered services to cater for the demands of the academic community aided by an online model that optimizes students' time and experiences. Besides, YDUQS has also implemented a new virtual classroom (SAVA – Sala de Aula Virtual) for students from online as well as in-person courses.

As part of its digital transformation, two months before the eruption of the pandemic, EnsineMe, the edtech of the group, had already started producing digital content based on pioneering formats and methodologies. Just before the pandemic, YDUQS had started developing and implementing the *Aura* teaching model in its face-to-face graduation courses. Owned by YDUQS itself, this model relies on digital platform, active methodologies and digital contents to support teachers and students in the classroom. Just one year after its implementation, in May 2021, students' approval ratings surpassed 90% according to data collected by an internal survey [16].

In the outset of the pandemic, in the first term of 2020, about 300,000 students from in-person courses migrated to remote teaching with digital classes on the Microsoft Teams platform. Before the Covid-19 outburst, most of these students had already been given access to virtual learning environments with the help of the WebAula platform, where available digital contents and a virtual library complemented their face-to-face learning experiences. The same virtual environment also hosted the online disciplines of in-person graduation courses.

With the pandemic and the implementation of remote teaching, in-person courses were split into two virtual environments: the Teams Platform was used for the transmission of synchronous classes, while the WebAula platform granted access to asynchronous digital contents, such as recorded video classes and other resources.

It took only fifteen days between the suspension of in-person classes and the beginning of remote teaching for graduation courses.

After the first term of 2020, with students from in-person classes studying in entirely virtual environments with remote encounters on the Teams platform, indicators signaled 83% of the students had remained within the institution – a percentage that represented a dropout rate below the expected levels. In the second semester of 2020, in-person courses witnessed a 5.9% decrease in the number of new enrollments, while distance education had a 58% increase [17].

In 2021, in the first term, in-person enrollments recovered and the number of students reached 299,000.

In addition to platforms for remote classes and digital content, students from in-person courses could also rely on the BdQ platform (Banco de Questões), a question database so they could get ready for assessments during the pandemic. The platform had been used previously in face-to-face as well as in remote courses as a source of quizzes and exercises. During 2020, over 4.7 million tests and exercises were done via the BdQ platform.

Furthermore, in the first term of 2021, 43% of the alumni from YDUQS face-to-face courses had access to contents produced in 2020 by the group's edtech, EnsineMe. About 600,000 students from YDUQS from in-person and distance course modalities already had the application program offered by the institution in 2021 [18].

In 2021, YDUQS acquired Qconcursos, an edtech focusing on preparing candidates for admission processes, so that it could progress in providing customized digital learning through adaptative evaluations. Thus, the institutions from the YDUQS group were able to count on various technological resources, as Table I describes.

Table 1: Students Enrolled in Face-to-Face Courses and Digital Platforms

	2020 First term	**2021 First term**
Students enrolled in face-to-face courses	300,000	299,000
Platforms	Teams WebAula BdQ	Teams WebAula SAVA BdQ
Students' application programs	Minha Estácio	Minha Estácio Meu Ibmec
Teachers' application programs	Estácio docente	Estácio docente Wyden docente Ibmec docente
Edtechs' internal Hub	EnsineMe	EnsineMe QConcursos
Pedagogical activities in face-to-face courses	Remote classes	Remote classes. In-person practical classes.

Students' evaluation of educational services throughout the pandemic was, indeed, very positive.

Data from an internal survey – carried out by the institution, updated on April 30th 2021, and published in its financial report – reveals a 17-percentage point improvement in terms of the Net Promoter Score (NPS). When compared to results obtained between 2020 and 2021, the NPS presented a record-breaking 21 percentage point increase.

4.2. Analysis of the results

This data reveals that the educational group targeted by this research quickly responded to the challenge of resuming pedagogical activities of in-person courses due to a favorable internal context.

A greater investment in digital transformation had already been set in motion two years before the Covid-19 pandemic, propelling the process of partial digitalization in face-to-face courses, by offering some online and hybrid disciplines.

Such rapid response to the sanitary crisis is also connected to the prior existence of separate application programs for teachers and students. These applications went through several improvements during the pandemic, which helped the institution resume its typical academic routines.

For some students enrolled in face-to-face, communication with the institution, the supervision of the academic programme, and the possibility of studying while connected to mobile devices granted more flexibility to their learning experiences – an advantage priorly restricted to distance education.

For some teachers, submitting information such as students' grades, attendance records, and the content covered within each class became easier once digitalized. Nevertheless, they had to overcome the challenge of transforming remote teaching into a meaningful virtual encounter, not a depleted replica of a classroom.

In the case of YDUQS, the fact teachers from in-person courses could resort to previously acquired knowledge and experiences with digital resources and online environments. Furthermore, teachers' continuing education helped downplay the need to improvise in the transposition of in-person to online classes

Also, the implementation of a new teaching model (the *Aura* model) in face-to-face courses coincided with the beginning of the pandemic, which helped make the replacement of in-person classes smoother.

The support offered by internal EdTechs provided the institution with digital content produced in the pandemic context, furnishing teachers with the tools and resources they needed to assess students' performances in virtual environments.

Most technological solutions were created and managed internally, which explains the development of technological responses that suited the institutions' pedagogical needs and criteria.

The success of this response to the sanitary crisis is clearly reflected in the internal surveys promoted by YDUQS with the students and in the high levels of retention.

The increase of the satisfaction level of students enrolled in face-to-face courses indicates that the institution adopted adequate technological, didactic and pedagogical solutions.

However, the rapid migration from in-person courses to virtual environments in addition to retention indicators and the high levels of students' satisfaction relates to educational management. But the data presented in this report does not deal, for example, with learning management nor with students' academic performance over the course of the pandemic.

Also, this analysis does not cover the effects of large-scale digitalized learning on the relationships between teachers and students, nor amongst students themselves. These other aspects are extremely relevant; after all, the commercial design of educational systems and softwares has increasingly modelled teaching and learning experiences within universities [14]. Therefore, regardless of teachers' pedagogical intentions, softwares used in the educational process can either limit or expand what can be done within the classroom.

5. Conclusion

Based on the theoretical framework presented and on data regarding the implementation of remote teaching in the institutions of the YDUQS group, it is possible to infer that technology operates in both ways, curtailing or favoring communication, interaction and relationships in educational contexts.

A sound technological infrastructure and the use of different digital resources in the field of education before the pandemic created favorable conditions for a more rapid and less improvised response in the process of transposing conventional face-to-face classes to virtual environments.

Technological mediation by itself does not guarantee neither explains the success of the learning experience, since other aspects also play an important role in the educational process.

Despite being both necessary and relevant, this data is not enough to account for all the different aspects of technological mediation of the pedagogical processes during the Covid-19 pandemic in the higher educational institutions of the YDUQS group.

It is necessary to provide further details for this research so that the analysis can go beyond the macro level and look into variables regarding the implementation of remote learning from a micro perspective.

The analysis of data on academic management must be complemented by specific indicators to account for educational challenges of learning through technological mediation during the pandemic, which is certainly a necessary possibility for further investigation on this theme.

Conflict of Interest

The author declares no conflict of interest.

References

[1] L. C. D. Saldanha, "Technological Mediation in Remote Teaching During the Covid-19 Pandemic," 17th Iberian Conference on Information Systems and Technologies (CISTI), 1-7, 2022, doi: 10.23919/CISTI54924.2022.9820052.

[2] C. Hodges, S. Moore, B. Loockee, T. Trust, A. Bond, "The difference between emergency remote teaching and online learning," Educause Review, 27, 2020, Available: https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning.

[3] E. Davis, "What is remote teaching," Top Hat, Glossary, 2020.

[4] J. Baggaley, "Educational distancing," Distance Education, **41**(4), 582-588, 2020, doi: 10.1080/01587919.2020.1821609.

[5] A. Behar, "O ensino remoto emergencial e a educação a distância," Coronavírus, UFRGS, 2020, Available: https://www.ufrgs.br/coronavirus/base/artigo-o-ensino-remoto-emergencial-e-a-educacao-a-distancia/#:~:text=O%20Ensino%20Remoto%20Emergencial%20e%20a%20Educa%C3%A7%C3%A3o%20a%20Dist%C3%A2ncia%20n%C3%A3o,refere%20a%20um%20distanciamento%20geogr%C3%A1fico.

[6] B. Veloso, D. Mill, "Distance Education and Remote Teaching: opposition by the vertex," SciELO Print, 2022, doi: 10.1590/SciELOPreprints.3506.

[7] V. Dusek, Philosophy of Technology: An Introduccion, Blackwell Publishing, 2006.

[8] A. Watters, "Teaching Machines: The History of Peronalized Learning Cambridge, The MIT Press, 2021.

[9] G. Canole, "Bridging the gap between policy and practice: A framework for technological intervention," Journal of e-Learning and Knowledge Society, **6**(1), 13-27, 2010, doi: 10.20368/1971-8829/384.

[10] S. Freitas, G. Canole, "Leaners experiences: How pervasive and integrative tolos influence expectations of study," In R. Sharpe, H, Beetham, S. Freitas (Eds.). Rethinking learning for a digital age: How learners are shaping their own experiences, Routledge, 2010.

[11] L. C. D. Saldanha, "The Discourse of Remote Teaching During the COVID-19 Pandemic," Journal of Higher Education Theory and Practice, [S. l.], **21**(4), 53-63, 2021, doi: 10.33423/jhetp.v21i4.4207.

[12] J. Mattar, A. Loureiro, E. Rodrigues, "Educação online em tempos de pandemia: desafios e oportunidades para professores e alunos," Interacções, **16**(55), 1-5, 2020, doi: 10.25755/int.22001.

[13] A. Feenberg, Tecnologia, modernidade e democracia, MIT Portugal, Inovatec, 2015.

[14] L. Castañeda, N. Selwyn, Reiniciando la universidad**:** buscando un modelo de universidad en tiempos digitales. Editorial UOC, 2019.

[15] YDUQS, "Divulgação de resultados 1T20," YDUQS Participações S.A., 2020.

[16] YDUQS, "Unidades de negócio," YDUQS Participações S.A., 2022.

[17] YDQUS, "Divulgação de resultados 3T20," YDUQS Participações S.A, 2020.

[18] YDUQS, "Divulgação de resultados 2T21 & 1S21," YDUQS Participações S.A, 2021.

Advances in Science, Technology and Engineering Systems Journal Vol. 7, No. 5, 193-198 (2022)

www.astesj.com

Special Issue on Computer Engineering and Multi. Sciences

ASTESJ
ISSN: 2415-6698

Hybrid Neural Network Method for Predicting the SOH and RUL of Lithium-Ion Batteries

Brahim Zraibi*,[1], Mohamed Mansouri[1], Salah Eddine Loukili[2], Said Ben Alla[2]

[1]National School of Applied Sciences of Berrechid, Laboratory LAMSAD, Hassan First University of Settat, Morocco

[2]Faculty of Science and Technology, Laboratory VTE, Hassan First University of Settat, Morocco

ARTICLE INFO

Article history:
Received: 28 August, 2022
Accepted: 25 October, 2022
Online: 31 October, 2022

Keywords:
Lithium-ion Batteries
Remaining Useful Life
Deep Neural Network
State-of-Health

ABSTRACT

The use of a battery to power an electrical or electronic system is accompanied by battery management, i.e. a set of measures intended to preserve it for preventative maintenance, thus the cost reduction. This management is generally based on two key parameters, the (remaining useful life) RUL and the (State-of-health) SOH, which relate respectively to the charge output and the aging of the Lithium-ion battery. The issue will be resolved and advances in production, battery utilization, and optimization will be made possible by accurate SOH determination and dependable RUL prediction. The CNN-BGRU-DNN hybrid strategy, which we suggest in this study, integrates Convolutional Neural Networks (CNN), Bidirectional Gated Recurrent Units (BGRU), and Deep Neural Networks (DNN) to increase the precision of SOH and RUL estimates for Lithium-ion batteries. To that purpose, the performance of the prediction findings is assessed using the MAE, RMSE, AE, and RE as well as the NASA datasets of lithium-ion batteries for experimental validation. The verification tests' findings show that, in comparison to existing approaches in the literature, the suggested method may greatly reduce prediction error and achieve high estimation accuracy of the battery's state of health.

1. Introduction

Electric vehicles are a promising technology for reducing the increasing air pollution such as decreasing CO2 emission from worldwide transportation. They are operated by battery packs [1]. The accelerated electrification of vehicles is significantly facilitated by batteries [2]. There are five key considerations for EV batteries: longevity, specific energy, specific power, cost, and safety. Over the past ten years, the first four factors have greatly aided in the optimization of electrode and electrolyte materials. Many researchers worldwide, however, have not fully addressed the question of safety [3]. The repeated operation of batteries leads to loss of capacity and increase the resistance, which allow some catastrophes to happen like explosion and combustion resulted on the excessive usage. The solution will enable advancements in battery production, use, and optimization through accurate state of health (SOH) determination and reliable remaining useful life (RUL) prediction. For instance, end users can make an estimation of the predicted battery life to ensure that batteries are used to their greatest capability before being replaced or discarded. To expedite the testing, validation, and production processes, manufacturers might group new cells according to their anticipated lifetime [4]. As a result, the complete electrification system requires an intelligent BMS capable of forecasting and monitoring battery behavior, which are very important for the safety and reliability of EVs and ESS [5]. Among different batteries, Li-Bs are widely regarded as potential options for a variety of applications, owing to their high energy density, power density, low self-discharge rate, and extended lifespan. Recently, many researches have started to focus on parameters of the BMS battery to estimate each of them. Many factors, including the state of charge (SOC), SOH, RUL, the charge capacity, and the internal resistance, must be monitored to ensure that Li-ion batteries are used efficiently and safely [6] [7]. Throughout the life cycle of lithium batteries in electrified vehicles, SOH is an essential parameter for problem diagnostics and safety early warnings in addition to its capacity to precisely predict the remaining mileage of EVs [8]. The RUL prediction of Li-Bs considers a significant choice for reliability, safety, and efficient battery operations, which is the number of cycles (charge/discharge) left before the battery fails, which is between 70 and 80% of its maximum capacity [7], [9], [10]. The equivalent circuit model [11], electrochemical model, data-driven model, and hybrid method

*Corresponding Author: Brahim Zraibi, b.zraibi@uhp.ac.ma

model are the four primary models that have been used in recent decades to perform substantial research on RUL estimate and SOH prediction of lithium batteries. The approach of the data model is receiving increasing amounts of attention as a result of the growth in lithium battery data [12], [13].

An accurate state of health (SOH) estimate helps ensure dependability and safety while the battery is operating. There are several ways to estimate it, including hybrid techniques based on neural networks [14], [15]. In 2017 [16], the author proposed the OS-ELM method and they utilized the discharge time of equal voltage interval as the HI. In 2019 [17], the author focused on their paper on the SOH estimation of lithium-ion battery using PKNN and Markov Chain. For verification, they compared the PKNN with other methods, which it obtained a high prediction accuracy. Besides, in [18], the author propose a method that combines the partial incremental capacity and ANN. Additionally, in [19], the author combine the ANN method with the PF algorithm for estimating the SOH, where they obtained an accurate estimation. In 2020 [20], the author integrated the deep Boltzmann machines and LSTM for obtaining the health prediction of a medical Li-ion battery. The empirical results obtain a good of SOH prediction. In [21], the author proposed a combination between GRU and CNN. While, in [8], the author combined the WNN with UPF. The performance results demonstrate their capability in improving the accuracy of SOH prediction.

The goal of this paper is to estimate the SOH and RUL of a lithium-ion battery using a hybrid method named CNN-BGRU-DNN. The comparison is performed between the proposed hybrid method and various prediction methods. The experiment obtained good results for the proposed method that achieved high predictive accuracy for the SOH and estimation compared to the other results.

The remaining parts of this essay are written as follows: The tools and methods for forecasting the RUL and SOH of battery lithium-ion batteries using the suggested method are presented in Section 2. A comparison of the SOH estimate accuracy is shown in Section 3. A conclusion is then offered.

2. RUL and SOH Prediction

2.1. CNN-BGRU-DNN architecture

The SOH and RUL of Li-ion batteries have been predicted using the CNN, BGRU, and DNN methods in prior literary works, where they performed well. By merging CNN, BGRU, and DNN, our study aims to enhance and attain high accuracy of SOH and RUL estimate.

In terms of feature extraction, CNN is proficient and benefits from both scale invariance and local dependence. Its feature extraction process is organized hierarchically. Through the use of many feature planes and neurons, the first layer of convolution extracts various input characteristics. In order to acquire continuous spatial features, the second layer, known as secondary feature extraction, decreases the feature surface dimension and its resolution. The outputs of the convolution layer are the inputs of the pooling layer, and the two layers are mapped one to one and each to the other. The data from the first two levels can be combined in the third layer. Full connection outputs are delivered to the last layer. [22].

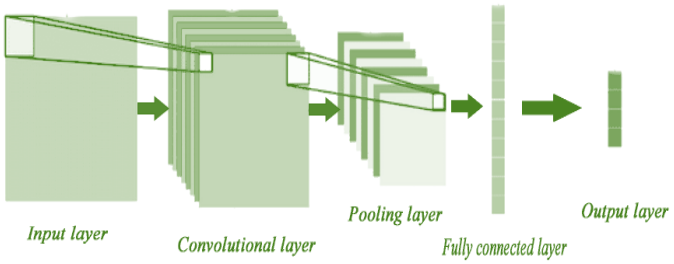


Figure 1: CNN structure

The RNN is one of the most well-liked deep learning (DL) algorithms since it makes use of the temporal correlations between neurons, although it suffers from the gradient vanishing issue [23]. Two RNN variations, LSTM and GRU, are utilized to regulate the propagation of gradient information and remember the parameters as successive inputs during the long-term sequence in order to solve this problem. [13].

GRU is classified as one of the RNN's variants. Its ability to regulate the propagation of gradient information and retain the parameters as future inputs over the long-term sequence is a core element. GRU consists of two gates: update gate z, which regulates the updating of the hidden state, and reset gate r, which determines whether or not to ignore the prior hidden state.

GRU's equations can be defined as follow:

$$z_t = \sigma(W_z[h_{t-1}, x_t])$$
$$r_t = \sigma(W_r[h_{t-1}, x_t]) \quad (1)$$
$$\tilde{h}_t = \tanh(W_h[r_t * h_{t-1}, x_t])$$
$$h_t = ((1 - z_t) * h_{t-1}) + (z_t * \tilde{h}_t)$$

where, $\tilde{h}_t$ is the candidate gate and h_t is output activation, the unit output as (h), W is the weight matrices, and σ is the sigmoid function represented [24].

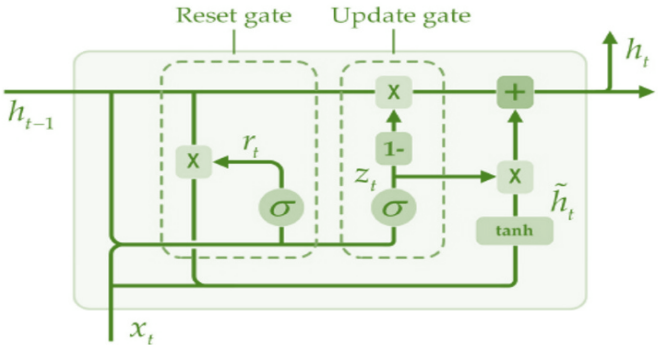


Figure 2: GRU structure

A GRU neural network with a two-layer structure is known as a Bidirectional Gate Recurrent Unit (BGRU) neural network. They have the ability to process the data inputs in both directions,

i.e. the forward and backward temporal sequences, with the outputs of both connected in the same output layer, allowing these bidirectional algorithms to be more efficient with defining the relationship between the sequences and its model. BGRU can save cost and time by reducing the amount of calculations required.

To benefit from the advantages of each algorithm and enhance the performance, they were combined with each other. The results of the integration of CNN, bidirectional of GRU, and DNN into one framework obtained good performance. The DL technology uses multiple layers to extract higher-level features from the raw input progressively.

Data processing is the initial step. We chose the discharge data from datasets that we extract from specific batteries that comprise charge, discharge, and impedance features. For each cycle of our experiment, where the input is the prior capacity and the output is the current capacity, we only choose one feature from this data, the capacity. We then used a window size of eight values to organize this data for the training step, which predicts data sequences. Finally, we split the data into test and training sets using the same split ratios for each battery. To predict the RUL of the Li-ion battery, we used the CNN-BGRU-DNN technique based on univariate time series.

We try to profit from their advantages where CNN is applied to extract local features, capture the spatial relationship, and use shared weights structure to reduce the amount of the weights and try to find the shared information from the measurement of data. Where we use one convolutional layer with 64 filters inclusive of the kernel size of 4, also we employ in this structure one default stride, causal padding, and Relu activation function. Then, the BGRU is applied to understand the temporal relationships in the feature sequence and it uses their internal state (memory) to learn features and time dependencies from the sequential data, and capture temporal features. Where we utilize two layers from each of them, which consist of 160 nodes then a flatten layer comes next. While DNN maps the features by choosing 3 dense layers, containing the Relu activation functions of each layer, with 128 nodes. Then we use one dense layer with one node to employ as a regression layer for getting the final SOH output and contribute to accurate prediction. Thus, the architecture of the proposed method shown in figure 3 is chosen after numerous experiments.

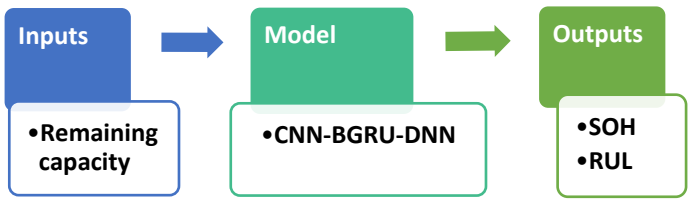


Figure 3 : The framework of proposed method

2.2. Experiment description

This research validates its findings using experimental data from the NASA Prognostics Center of Excellence [25], which includes of aging information 18650 Li-ion batteries. Where Table 1 provides the following information regarding these batteries:

Table 1: NASA Lithium-Ion Batteries description

Batteries	NASA
Temperature (C)	24
Constant charge current	1.5 (A)
Cut-off voltage of Charge/Discharge	4.2 / 2.5 (V)
Minimum charge current	20 (mA)
Rated capacity (Ah)	2
Cycles	168 (B5,B6,B7)

The proposed approach, CNN-BGRU-DNN, was implemented using the hyper-parameters presented in table 2 and using the following environment and tools:

- Google Colaboratory notebook
- 1 CPU Core: Intel(R) Xeon(R) CPU @ 2.20GHz
- Physical memory: 12G
- GPU: Tesla K80 - 11441MiB memory
- CUDA Version: 11.2
- TensorFlow version: 2.7.0
- Python version: 3.7.12.

Table 2: Hyper-parameters values

Hyper parameters	values
Window size	8
Batch size	32
Shuffle buffer size	1000
Epochs	1400
Learning rate	8e-4
Regularization	without
Activation function	ReLU
Optimizer	Adam
Loss function	Huber

We utilize MAE and RMSE [24] to assess how well the algorithms execute SOH estimation, while AE and RE are used to assess RUL prediction accuracy. These are their definitions [26] :

$$MAE = \frac{1}{K}\sum_{k=1}^{k}|y_k - \widehat{y_k}| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_k - \widehat{y_k})^2} \quad (3)$$

$$AE=|RULreal - RULpredicted| \quad (4)$$

$$RE= |RULreal-RULpredicted| / RULreal \times 100\%. | \quad (5)$$

Where $\widehat{y_k}$ is the predicted value and y_k is the actual value. The accuracy of the SOH forecast is greater when the MAE and RMSE are near to zero.

2.3. RUL and SOH estimation

The major of this section is to present the ability of the proposed hybrid method CNN-BGRU-DNN to estimate the SOH and RUL of different Li-ion batteries; it is also for confirming our method's prediction accuracy.

The battery's capacity, performance, and state of health are shown by SOH indicator. It is the ratio of a battery's actual capacity (Ca) to its rated capacity (Cr), where actual capacity refers to how much of the battery's capacity is actually used when it is fully charged. The rated capacity of a totally charged battery is 100%, whereas the capacity of a totally failed battery is 0%. The battery's SOH is defined as follows [27]:

$$\text{SOH} = \frac{\text{Ca}}{\text{Cr}} \tag{6}$$

The remaining number of cycles of battery capacity to reach at its failure threshold that means the time between now and the end-of-life "EOL" is defined as RUL, showing as follows [7]:

$$\text{RUL} = C_{EOL} - C_{cc} \tag{7}$$

C_{cc} is the number of cycle at the actual capacity and C_{EOL} is the cycle number when the capacity of battery arrives at the EOL.

The experiment were terminated only when battery attained their EOL, as seen in figure 4, where the line of EOL represented by a red color, which considered as the time when the capacity reaches 70% in rated capacity for the NASA batteries. The EOL is calculated as:

$$\text{EOL} = \text{Cr} * 0.7 = 1.4 \text{ Ah} \tag{8}$$

In this study, we separated the datasets into training and prediction data with the identical beginning prediction point of each dataset, which is 80 cycles. We utilized three batteries, B0005, B0006, and B0007, to establish the degradation sample of the battery's capacity.

Figure 4 above displays the outcomes of the SOH and RUL predictions, where Real values are displayed in blue and predicted values are shown in orange. The SOH and RUL predictions for NASA batteries demonstrate how the suggested hybrid technique, CNN-BGRU-DNN, practically always results in almost identical actual and predicted curves for all batteries. As a result, the hybrid method's SOH estimation accuracy is good. The point of failure at the end of life (EOL) for all batteries is when both curves almost exactly meet. As a result, CNN-BGRU-DNN achieves the maximum level of RUL prediction accuracy.

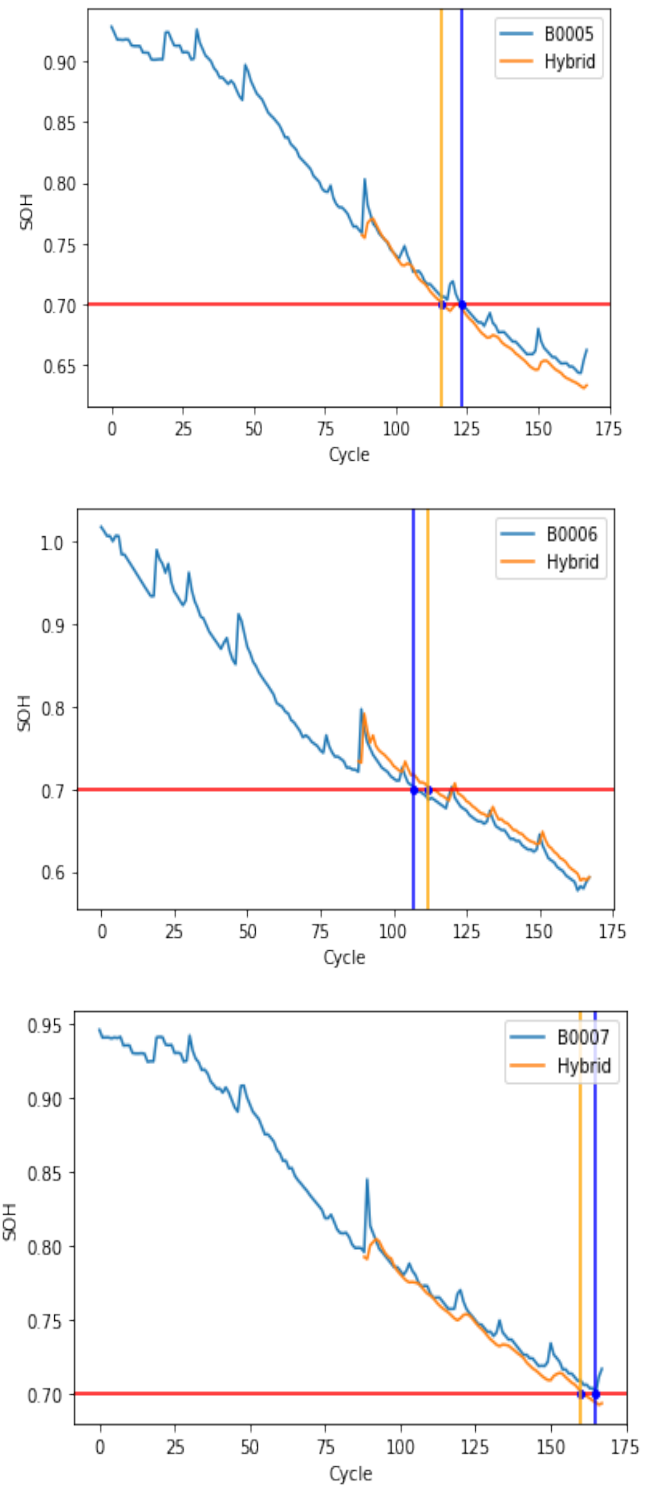


Figure 4: The SOH and RULprediction results using CNN-BGRU-DNN.

Table 3: SOH estimation results

Batteries	Methods	RMSE	MAE
B0005	CNN-BGRU-DNN	0.01165	0.00884
B0006	CNN-BGRU-DNN	0.00884	0.01334

B0007	CNN-BGRU-DNN	0.00990	0.00667

Table 4: RUL estimation results

CNN-BGRU-DNN				
Batteries	RUL_{real}	$RUL_{predicted}$	**AE**	**RE %**
B0005	123	116	7	5.69
B0006	107	112	5	4.67
B0007	165	160	5	3.03

The values for MAE, AE, and RMSE are extremely low, as seen in Tables 3 and 4. The CNN-BGRU-DNN approach helps minimize error during SOH deterioration, as this experiment shows. Furthermore, the Li-ion battery RUL estimation using the CNN-BGRU-DNN method is accurate. This demonstrates that CNN-BGRU-DNN approach is the best at predicting battery SOH, and the experiment illustrates perfectly the hybrid method's capacity for having greater forecast accuracy.

3. Comparison between proposed method and other methods in literature

This subsection provides a comparison of the SOH prediction accuracy of different other studies' predictions. We use performance method findings from earlier papers to compare more widely with other forms of prediction, since these approaches use the same NASA datasets and performance measures.

Table 5: SOH estimation results of some papers

Batteries	**Methods**	**RMSE**
B0005	UPF Elman NN WNN-UPF [8]	1.088 0.210 0.027
	GPR- LSTM [28]	0.012
	CNN-BGRU-DNN	0.011
B0006	CGTSSA_Cat_Boost [29] SSA_Cat_Boost PSO_Cat_Boost Cat_Boost CGTSSA-SVM CGTSSA-ELM	0.0268 0.0317 0.0531 0.0708 0.0362 0.0579
	UPF Elman NN WNN-UPF [8]	1.115 0.223 0.050
	GPR- LSTM [28]	0.013
	CNN-BGRU-DNN	0.008
B0007	UPF Elman NN WNN-UPF [8]	1.161 0.145 0.024
	CGTSSA_Cat_Boost [29] SSA_Cat_Boost PSO_Cat_Boost Cat_Boost CGTSSA-SVM CGTSSA-ELM	0.0118 0.0147 0.0263 0.0465 0.0178 0.0447
	GPR- LSTM [28]	0.009
	CNN-BGRU-DNN	0.009

From the results of table 5, we can clearly see that the RMSE value of our proposed method is smaller than the values reported by the studies, the RMSE metric is widely used in regression problems where we predict continues values, which is the case is the prediction of the SOH of Ion-Lithium batteries.

We can conclude based on the results of table 5 that the proposed suggested named CNN-BGRU-DNN is a good estimator with its high accuracy for predicting the RUL and SOH.

Conclusion

This study proposes a hybrid approach known as CNN-BGRU-DNN to predict Li-ion battery SOH and RUL. A dataset received from NASA is used to experimentally validate the suggested strategy. The results of the proposed hybrid method demonstrate that we achieved a big performance improvement and satisfying results evaluated by the performance indicators called MAE, RE %, AE and RMSE, where error rates are reduced and accuracy increased. In comparison to the outcomes of previous publications, four prediction performance indices show that CNN-BGRU-DNN has the greatest accuracy.

Nomenclature

AE	absolute error
ANN	artificial neural network
BMS	battery management system
BGRU	bidirectional gated recurrent units
CNN	convolutional neural network
DNN	deep neural networks
DL	deep Learning
ELM	extreme learning machine
LSTM	long short-term memory
Li-B	lithium-ion battery
MAE	mean absolute error
ML	machine learning
NASA	national aeronautics and space administration
PF	particle filter
RE	relative error
RNN	recurrent neural network
RMSE	root mean square error
RUL	remaining useful life
ReLU	rectified linear unit
UKF	unscented Kalman filter
WNN	wavelet neural network
UPF	unscented Kalman particle filter

References

[1] M.U. Ali, A. Zafar, S.H. Nengroo, S. Hussain, G.S. Park, H.J. Kim, 'Online remaining useful life prediction for lithium-ion batteries using partial discharge data features', Energies, **12**(22), 2019, doi:10.3390/en12224366.

[2] R. Xiong, Y. Zhang, J. Wang, H. He, S. Peng, M. Pecht, 'Lithium-Ion Battery Health Prognosis Based on a Real Battery Management System Used in Electric Vehicles', IEEE Transactions on Vehicular Technology, **68**(5), 4110–4121, 2019, doi:10.1109/TVT.2018.2864688.

[3] H. Chaoui, C.C. Ibe-Ekeocha, 'State of Charge and State of Health Estimation for Lithium Batteries Using Recurrent Neural Networks', IEEE Transactions on Vehicular Technology, **66**(10), 8773–8783, 2017, doi:10.1109/TVT.2017.2715333.

[4] W. Luo, C. Lv, L. Wang, C. Liu, 'Study on impedance model of Li-ion battery', Proceedings of the 2011 6th IEEE Conference on Industrial Electronics and Applications, ICIEA 2011, 1943–1947, 2011, doi:10.1109/ICIEA.2011.5975910.

[5] J. Fan, J. Fan, F. Liu, J. Qu, R. Li, 'A Novel Machine Learning Method Based Approach for Li-Ion Battery Prognostic and Health Management', IEEE Access, **7**(1), 160043–160061, 2019, doi:10.1109/ACCESS.2019.2947843.

[6] B. Zraibi, M. Mansouri, C. Okar, 'Comparing Single and Hybrid methods of Deep Learning for Remaining Useful Life Prediction of Lithium-ion Batteries', E3S Web of Conferences, **297**, 01043, 2021, doi:10.1051/e3sconf/202129701043.

[7] B. Zraibi, C. Okar, H. Chaoui, M. Mansouri, 'Remaining Useful Life Assessment for Lithium-ion Batteries using CNN-LSTM-DNN Hybrid Method', IEEE Transactions on Vehicular Technology, 2021, doi:10.1109/TVT.2021.3071622.

[8] J. Jianfang, W. Keke, P. Xiaoqiong, S. Yuanhao, W. Jie, Z. Jianchao, 'Multi-Scale Prediction of RUL and SOH for Lithium-Ion Batteries Based on WNN-UPF Combined Model', Chinese Journal of Electronics, **30**(1), 26–35, 2021, doi:10.1049/cje.2020.10.012.

[9] Y. Toughzaoui, S. Bamati, H. Chaoui, H. Louahlia, 'State of health estimation and remaining useful life assessment of lithium-ion batteries : A comparative study', **51**(March), 2022, doi:10.1016/j.est.2022.104520.

[10] J. Wei, G. Dong, Z. Chen, 'Remaining Useful Life Prediction and State of Health Diagnosis for Lithium-Ion Batteries Using Particle Filter and Support Vector Regression', IEEE Transactions on Industrial Electronics, **65**(7), 5634–5643, 2018, doi:10.1109/TIE.2017.2782224.

[11] C. Chang, Q. Wang, J. Jiang, T. Wu, 'Lithium-ion battery state of health estimation using the incremental capacity and wavelet neural networks with genetic algorithm', Journal of Energy Storage, **38**(September 2020), 102570, 2021, doi:10.1016/j.est.2021.102570.

[12] L. Yao, S. Xu, A. Tang, F. Zhou, J. Hou, Y. Xiao, Z. Fu, 'A Review of Lithium-Ion Battery State of Health Estimation and Prediction Methods', 2021.

[13] B. Zraibi, M. Mansouri, S.E. Loukili, 'Comparing deep learning methods to predict the remaining useful life of lithium-ion batteries', Materials Today: Proceedings, (xxxx), 2022, doi:10.1016/j.matpr.2022.04.082.

[14] S. Yang, C. Zhang, J. Jiang, W. Zhang, L. Zhang, Y. Wang, 'Review on state-of-health of lithium-ion batteries : Characterizations , estimations and applications', Journal of Cleaner Production, **314**(May), 128015, 2021, doi:10.1016/j.jclepro.2021.128015.

[15] A. Basia, Z. Simeu-abazi, E. Gascard, P. Zwolinski, 'Review on State of Health estimation methodologies for lithium-ion batteries in the context of circular economy', CIRP Journal of Manufacturing Science and Technology, **32**, 517–528, 2021, doi:10.1016/j.cirpj.2021.02.004.

[16] Y. Zhu, F. Yan, J. Kang, C. Du, 'State of health estimation based on OS-ELM for lithium-ion batteries', International Journal of Electrochemical Science, **12**(7), 6895–6907, 2017, doi:10.20964/2017.07.35.

[17] H. Dai, G. Zhao, M. Lin, J. Wu, G. Zheng, 'A novel estimation method for the state of health of lithium-ion battery using prior knowledge-based neural network and markov chain', IEEE Transactions on Industrial Electronics, **66**(10), 7706–7716, 2019, doi:10.1109/TIE.2018.2880703.

[18] S. Zhang, B. Zhai, X. Guo, K. Wang, N. Peng, X. Zhang, 'Synchronous estimation of state of health and remaining useful lifetime for lithium-ion battery using the incremental capacity and artificial neural networks', Journal of Energy Storage, **26**(July), 100951, 2019, doi:10.1016/j.est.2019.100951.

[19] W. Qin, H. Lv, C. Liu, D. Nirmalya, P. Jahanshahi, 'Remaining useful life prediction for lithium-ion batteries using particle filter and artificial neural network', Industrial Management and Data Systems, **120**(2), 312–328, 2019, doi:10.1108/IMDS-03-2019-0195.

[20] C.C. Liu, T. Wu, C. He, 'State of health prediction of medical lithium batteries based on multi-scale decomposition and deep learning', Advances in Mechanical Engineering, **12**(5), 2020, doi:10.1177/1687814020923202.

[21] Y. Fan, F. Xiao, C. Li, G. Yang, X. Tang, 'A novel deep learning framework for state of health estimation of lithium-ion battery', Journal of Energy Storage, **32**(August), 101741, 2020, doi:10.1016/j.est.2020.101741.

[22] X. Song, F. Yang, D. Wang, K.L. Tsui, 'Combined CNN-LSTM Network for State-of-Charge Estimation of Lithium-Ion Batteries', IEEE Access, **7**, 88894–88902, 2019, doi:10.1109/ACCESS.2019.2926517.

[23] Y. Zhang, R. Xiong, H. He, Z. Liu, 'A LSTM-RNN method for the lithuim-ion battery remaining useful life prediction', 2017 Prognostics and System Health Management Conference, PHM-Harbin 2017 - Proceedings, (51507012), 2017, doi:10.1109/PHM.2017.8079316.

[24] M. Sajjad, Z.A. Khan, A. Ullah, T. Hussain, W. Ullah, M.Y. Lee, S.W. Baik, 'A Novel CNN-GRU-Based Hybrid Approach for Short-Term Residential Load Forecasting', IEEE Access, **8**, 143759–143768, 2020, doi:10.1109/ACCESS.2020.3009537.

[25] K.G. Saha, 'Battery data set', NASA AMES Prognostics Data Repository,.

[26] D. Liu, Y. Luo, J. Liu, Y. Peng, L. Guo, M. Pecht, 'Lithium-ion battery remaining useful life estimation based on fusion nonlinear degradation AR model and RPF algorithm', Neural Computing and Applications, **25**(3–4), 557–572, 2014, doi:10.1007/s00521-013-1520-x.

[27] X. Bian, Z. Wei, J. He, F. Yan, 'A Novel Model-based Voltage Construction Method for Robust State-of-health Estimation of Lithium-ion Batteries', (December), 2020, doi:10.1109/TIE.2020.3044779.

[28] J. Zhao, Y. Zhu, B. Zhang, M. Liu, J. Wang, C. Liu, Y. Zhang, 'Method of Predicting SOH and RUL of Lithium-Ion Battery Based on the Combination of LSTM and GPR', 2022.

[29] M. Zhang, W. Chen, J. Yin, T. Feng, 'Health Factor Extraction of Lithium-Ion Batteries Based on Discrete Wavelet Transform and SOH Prediction Based on CatBoost', 2022.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 193-201 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

Investigation of Swimming Behavior and Performance of the Soft Milli-Robots Embedded with Different Aspects of Magnetic Moments

Xiuzhen Tang, Laliphat Manamanchaiyaporn*

Center of Excellence in Creative Engineering Design and Development, Department of Mechanical Engineering, and Research Unit of Multi-Scale Robotics, Thammasat School of Engineering, Faculty of Engineering, Thammasat University, Thailand

ARTICLE INFO

Article history:
Received: 28 February, 2023
Accepted: 05 May, 2023
Online: 25 June, 2023

Keywords:
Magneto-elastic actuator
Soft-robotics
Magnetic manipulation

ABSTRACT

Among the development of technology, a large number of medical devices have been implemented in various forms for therapy and treatment. Remote controllability, real-time response, small size, and non-toxicity of the devices are the critically requirement to be operated in the blind, unstructured and fluidic environments of biomedical regions. Untethered soft swimming milli-robots have been developed to fulfill the remote operation in such that region under magnetic navigation. A motor-less mechanism of the soft robots utilizes a high degree of freedom provided by magnetic compliance of the deformable structure with a minimal control of the oscillating magnetic field. Theoretically, magnetic property of the soft robots is defined by magnetic moments consisting of orientation and strength. Orientation of magnetic moments can be defined by magnetizing technique, and strength of magnetic moments is obtained by their quantity in the magnetic structure. Herein, this work investigates how magnetic moments through the details of magnetic orientation and quantity affects swimming behavior and performance. The soft robots are fabricated with elastomer embedded with NdFeB microparticles to obtain three types of distinguish magnetic property in the deformable structure; the I robot has non-uniform magnetic orientation and uniform magnetic strength, the II robot has uniform magnetic orientation and non-uniform magnetic strength, and the III robot has non-uniform magnetic orientation and non-uniform magnetic strength. The results interestingly report that each type of robot's property functions mechanism and benefits swimming performance differently under the same magnetically control parameters. The I robot does not have any exceptional potential, but the II robot can be operated at the higher control frequency even reaching the step-out point. The III robot shows the greatest performance in swimming and maneuverability. These results would be useful to design a swimming soft-robot capable of applying for various purposes, especially when the demand concerns non-harm, small-scale size, soft interface and remote controllability.

1. Introduction

This paper is an extension of work originally presented in the eighth (8th) edition in the series of the International Conference on Control, Decision and Information Technologies, CoDIT'22 [1], in order to clarify how orientation and quantity of magnetic moments affect the in-fluid swimming potential of the soft robots.

Our body inside is a fantastic and complex system consisting of the circulation of diverse biological fluids and unstructured environments. Untethered miniature robots with millimeter scale or less have been promising to access the hard-to-reach biomedical regions for minimally invasive treatments (e.g., targeted drug delivery, biopsy) [2]. Due to the small size of the robot, the battery and motion mechanism were unable to set up inside the robot's structure. In order to function locomotion of the robots, active elements to respond with the external power sources were embedded in their body instead, during the fabrication process. Some types of the robots received the light emission pulse to transform its structure for mobility. Another type of robots employed the chemical reaction with the surrounding

* Corresponding Author: Laliphat Manamanchaiyaporn, mlalipha@engr.tu.ac.th

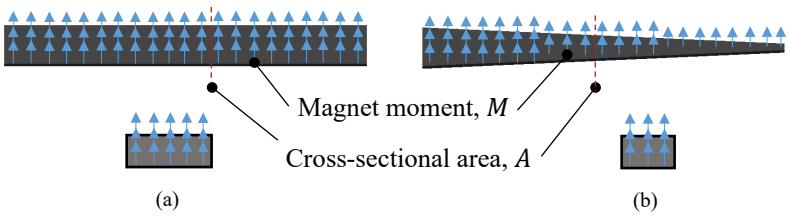


Figure 1: The different magnetic property of magnetic-soft structure between previous works and this work (a) uniform distribution of magnetic moments, (b) non-uniform distribution of magnetic moments

environment for propulsion. Other types of robots holding magnetic property were active after experiencing the actuating magnetic field. Each of robots has own mechanism functioned by particular active elements to create locomotion in environments.

The use of magnetic field was proved no harm to human tissue, especially affording the clinical imaging (e.g., MRI). It was one of the most selective sources to remotely manipulate robots. In order to generate magnetic energy to manipulate the robots, magnetic actuation systems were specifically designed in diverse configurations with a variety of control techniques (e.g. magnetic navigation in a large workspace) [3]. Once the actuation system generated magnetic field, the magnetic property of the robots as called magnetization responded with such that actuation through magnetic alignment. This concept allowed the robot to have effective locomotion via motor-less mechanism [4].

In life, environments are mostly viscous, which is hard for matters to perform motion, but microorganism-inspired robots adopt the asymmetric body movement with solid components as a key of success in fluidic maneuverability to swim effectively (e.g. beating or waving of flagella or cilia, helical propulsion) [5, 6], but having a solid structure limits safe interaction. The transition from a hard structure to a soft structure is wide-opening to fabricate swimming tiny robots by using materials embedded with functional elements. The integration of magnetic particles into the deformable structure enables controllability, leading to continuous and controllable movement under magnetic actuation due to magnetic alignment of magnetic moments within specimens of the structure. This aspect allows medical tools and medical robots to become more deformable and dexterous in various types of biomedical application (e.g. compliant-soft medical tools, flexible wearable devices) [7]-[10]. Moreover, having a deformable structure benefits a soft interface to greatly deal with uneven terrains and unstructured geometry without harmfulness [11, 12]. Such that soft structure with additional matters (e.g., drug molecule, chemical nanoparticles) still fulfil the remote applications in medicine, such as drug delivery, biopsy, detoxification [13]-[16].

In previous research, there were soft robots employing only anisotropic magnetization or non-uniform magnetic orientation to generate the body movement based on the continuous alignment of magnetic moments with the dynamic magnetic field. The deformation degree of each specimen of the robot's structure is equal and uniform because the distribution of magnetic moments is uniform across the whole soft structure, resulting in uniform magnetic strength. However, property of magnetic moments embedded in the soft structure still remains a challenge in the core detail. Magnetic moment typically comprises of orientation and strength, and what if they are not uniform across the whole deformable structure. In this paper, effect of orientation and strength of magnetic moments in the deformable structure of the soft robot is investigated in term of swimming behavior and performance based on lateral undulation. Three types of soft milli-robots are fabricated; the I robot: uniform magnetic strength and non-uniform magnetic orientation, II: non-uniform magnetic strength and uniform magnetic orientation, and III: non-uniform magnetic strength and orientation.

Those robots having the magnetically deformable structure utilizes high-degree of freedom provided by magnetic compliance as if motor-less mechanism set up inside, and they become more dexterous, especially in the applications of biomedicines (e.g.

Table 1: Raw materials and tools

Material	Property	Function
Liquid silicone rubber: LSR (SIMTEC)	Density: 1.13 g/cm^3, Young modulus: 300 kPa, temperature resistance: -50 to 250 ºC, tensile strength: 1.5 MPa, elongation at break: 700%	Base material to form a soft structure
NdFeB (Neodymium) magnetic microparticles	Particle size: 4 µm to 40 µm, density: 7.57 g/cm^3, Remanence: 720-760 mT, coercivity: 360-480 A/m, magnetic energy: 80 to 98 kJ/m^3	Active elements; base material to respond to magnetic actuation
A stainless-iron plate	30 mm x 100 mm with 80 µm±10 µm cavity	A mold doe the mixture to fabricate a magnetic-soft sheet
A shape guider	PLA (Polylactic Acid) 3D-printed parts with sinusoidal profile* (sin ($\frac{2\pi l}{\lambda}$)) at the inner surface	To constraint a magnetic-soft sheet before applying a magnetizing magnetic field to specific magnetic orientation of magnetic moments

* Profile to specify orientation of magnetic moments can be adjusted to any form.

compliant robotics, flexible medical tools). In particular, in order to operate the robot for treatment and therapy, potential of the robot to swim in fluidic and wet environments is a critical requirement.

Methods and materials are detailed in the next section. Next, experiments are conducted to test performance in swimming of the robots in fluid under the change in controlled parameters. Finally, the conclusion is issued.

2. Material and Method

2.1. Conceptual design

In the case that the size of a matter moving in media is millimeter or less, inertia term is dominated over viscous term, resulting in low Reynold number condition ($Re < 1$). Purcell stated that one of feasible motion patterns for small objects to effectively move in fluid was the use of the asymmetric body deformation under time-reversal [17]. Consequently, liquid silicone rubber (LSR) as a base material is used to obtain a deformable property, and magnetic particles is filled in order to respond with magnetic actuation. The combination of them leads to a controllable-deformable structure under magnetic field. This aspect results a motor-less mechanism in a soft small-scaled robot remotely controlled by the dynamic magnetic field to serve as a medical device.

Theoretically, magnetic moment or magnetic dipole is a vector that consists of magnitude and direction. It can be played in the detailed to define a specifically magnetic property. The soft milli-robot is designed to have three types of distinguish magnetic property; the I robot is with uniform magnetic strength and non-uniform magnetic orientation, the II robot is with non-uniform magnetic strength and uniform magnetic orientation, and the III robot is with non-uniform magnetic strength and orientation. Based on the modeling of magnetism [18], magnetic moment in a structure can be expressed by

$$\vec{\boldsymbol{m}} = \iiint M\, dV = \iiint M\, dAdl \tag{1}$$

where A, l, M are respectively the cross-sectional area, length of the structure, and magnetization which is a quantity of magnetized magnetic moment in a concerned volume, V. Thus, the existence of magnetic moment, according to Figure 1a and the definition of the eq. (1), can expressed that each specimen of the soft rectangular structure has the uniform magnetic orientation which points upward, and the uniform magnetic strength due to having equal number of the moments.

Otherwise, in Figure 1b, the existence of magnetic moment expresses that each specimen of the soft triangular structure has the uniform magnetic orientation which directs upward, but magnetic strength is non-uniform due to unequal number of magnetic moments along the length of the structure. The specimen where is the biggest cross-sectional area contains the strongest magnetic strength, but at the smallest area, magnetic strength is the weakest. Thus, the eq. (1) is rewritten to

$$\vec{\boldsymbol{m}}(x) = \iiint M(x)\, dV(x) = \iiint M\,(x) dA(x) dl \tag{2}$$

Eq. (2) expresses that at a position, x, on the length, l, if cross-sectional area across the length of the structure is unequal, magnetic moments at each position is different along the length, resulting in non-uniform magnetic strength.

2.2. Fabrication process

In this work, three-type property of the soft robot is modified with the existence of magnetic moments in the soft structure. List

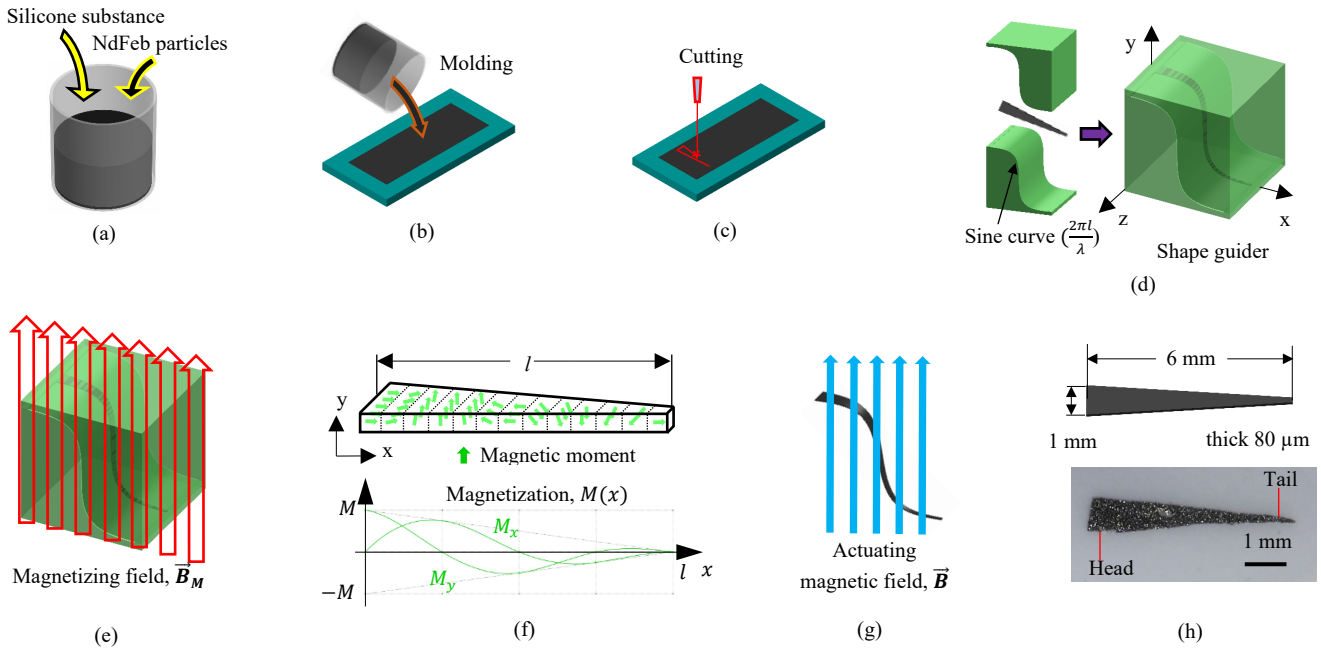
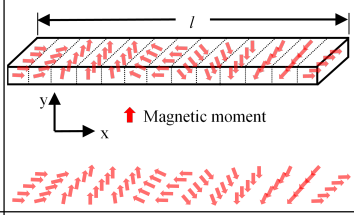
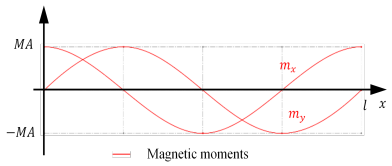
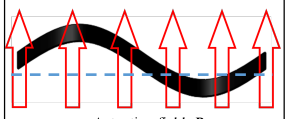
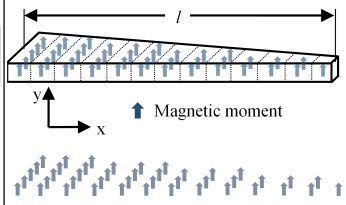
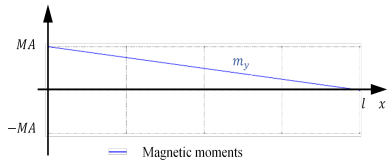
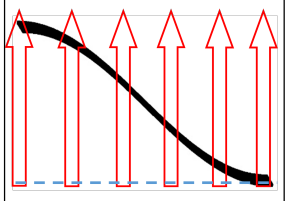
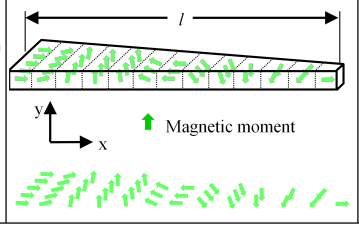
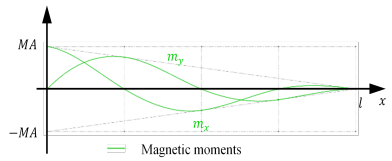
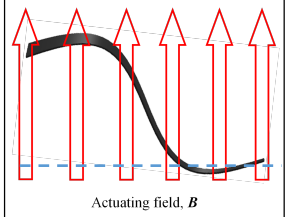


Figure 2: Example of fabricating the III robot. (a) Liquid silicone substances and NdFeb particles with 1:1 of mixing ratio. (b) molding on the stainless-steel plate, then subjects it to heat curing. (c) cutting into a triangular shape. (d) inserting into the sinusoidal-surface shape guider. (e) 700 mT of uniform magnetizing magnetic field. (f) magnetization profile of the film robot: (upper) magnetic orientation. (lower) magnetic strength decreases by increasing the body length. (g) the robot responding to the actuating magnetic field. (h) the triangular body: 1 mm × 6 mm × 80 µm.

Table 2: Three types of soft swimming milli-robots

Type	Width (w) Length (l) Thick (h)	Shape	Profile of magnetic direction*	Magnetic strength**	Magnetic response***
I	1 mm 6 mm 80 µm	Triangular	Sine 	Uniform 	
II	1 mm 6 mm 80 µm	Rectangular	Transverse 	Non-uniform 	
III	1 mm 6 mm 80 µm	Triangular	Sine 	Non-uniform 	

* A set of arrows depicts the direction of magnetic moments, and its quantity refers to magnetic strength in the robot's structure.
** A graph shows the relation of magnetic strength and magnetic orientation of magnetic moments as a function of body length.
*** Blue hidden line represents the original shape of teach robot before magnetic deformation due to alignment of magnetic moments with respect to the direction of the actuating magnetic field.

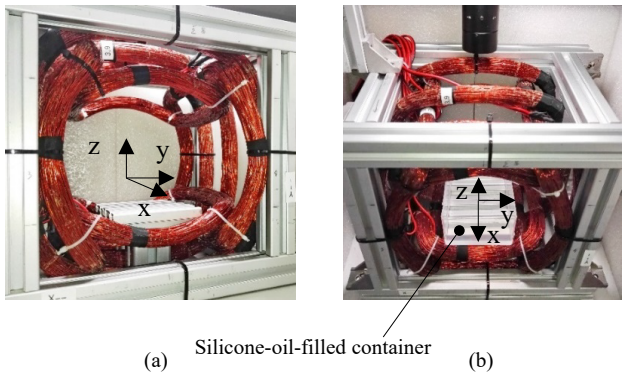


Figure 3: Magnetic actuation system. (a) It generates uniform-based field across the large bore. (b) a silicone-filled container as a workspace inserted into the bore.

of the main materials is reported on Table 1. In Figure 2, the robot fabrication starts from the mixing of liquid silicone rubber and NdFeB magnetic microparticles with the 1:1 mass ratio. Later, pouring the mixture into a stainless-steel mold with 80 µm cavity to form a magnetic soft sheet with 80 µm thickness, and putting the mold to cure at 250 ºC about 8-10 minutes. Finally, getting a magnetic soft sheet to prepare for making a magnetic-soft robot.

At this step, the finishing is the sheet embedded with magnetic elements across the whole volume, but its magnetic property is still not defined. In order to specify the direction of all magnetic moments, the sheet must be subjected to the 700-mT magnetizing magnetic field, $\vec{\boldsymbol{B}}_{\boldsymbol{M}}$. For example, if the direction of magnetic moments is similar to a wave, the sheet must be confined into the wave pattern, and then placing it into the $\vec{\boldsymbol{B}}_{\boldsymbol{M}}$ for permanent direction. Once removing it from the $\vec{\boldsymbol{B}}_{\boldsymbol{M}}$, the direction of magnetic moments follows as the wave across the whole volume permanently.

According to the eq. (2), uniform or non-uniform magnetic strength across the whole volume can be defined by the shape of the structure. For example, if the structure is symmetric as a rectangle, quantity of magnetic moments at each specimen will be equal across the whole length, resulting in uniform magnetic strength of magnetic moments. Otherwise, if being triangular, the cross-sectional area is not equal across the length, resulting in non-uniform magnetic strength of magnetic moments. As depicted in Figure 1, at the same position on the length, quantity of magnetic moments is different between two shapes. In Figure 3, after getting a desire shape, the sheet is cut into a triangular shape. Then, it is placed into a shape guider to form a sine-based

curve, and subjected into the $\vec{B}_M$ to profile the magnetic orientation of magnetic moments embedded in the triangular sheet. In this case, if it is actuated by magnetic field, the triangular sheet self transforms into a wave shape or a body wave deformation. Finally, the magnetic-soft sheet becomes a magnetic-soft robot. In this work, the robot is fabricated into three types to have different property of magnetic moments according to the Table 2.

2.3. Actuation and control procedure

The magnetic actuation system shown in Figure 3 consists of seven electromagnetic coils to achieve 3D-magnetic field in order to cover 6-DOF motion of the soft robot. It provides three directions of homogeneous magnetic field across a cylindrical workspace (radius: 7.5 cm and length: 18 cm). This aspect guarantees a pure magnetic torque exerted to actuate the soft robots without any drifting caused by the wrench of magnetic force. Each coil is individually operated by seven current drivers (Dimension engineering; 25 kHz, 30V/10A), and electrically supplied by SIEMENS GR60 (40A/48V). A custom controller with 8-bit-packeted-serial communication commands the drivers to pass electrical current into the coils to generate magnetic field. A stationary CMOS camera with zoom lens (working distance: 6-120 mm and 1.6-mm depth-of-field) is mounted to observe locomotion of the robot, and localize the robot's position to feedback the coordinate into the control algorithm in order to adjust a proper magnetic field. Maximum magnetic field is 25 mT and 100 T/m at 15 A input electrical current, measured by a gaussmeter GM-08 Hirst.

The magnetic-soft robot responds to magnetic actuation due to having magnetic moments, $\vec{m}$, as active elements embedded in the soft structure. It acts as a motor-less mechanism to make the robot swim in fluid. Depicted in Figure 3, the electromagnetic actuation system [3] is specifically designed to power the robot by using three-dimensional magnetic field, $\vec{B}$, which is

$$\vec{B} = [B_x \quad B_y \quad B_z] \tag{3}$$

Once the robot is under magnetic field, its soft body is deformed toward the direction of the magnetic field due to an alignment of magnetic moments across the volume. Magnetic torque, $\vec{T}$, and force, $\vec{F}$, are exerted to the robot, expressed by

$$\vec{F}(x) = \nabla(\vec{m}(x) \cdot \vec{B}) \tag{4}$$

$$\vec{T}(x) = \vec{m}(x) \times \vec{B} \tag{5}$$

Magnitude and direction of magnetic field is adjustable by using the superposition technique resulting from sources of magnetic field. In this work, the oscillating magnetic field is applied to manipulate the deformable structure of the robot continuously, leading to the body wave propagation or undulation as if swimming. The oscillating magnetic field is a product of the superposition of the magnetic field in the x- and y-direction, B_x and B_y, which oscillates with frequency, f (Hz: cycle number in a second). As shown in Figure 4 of the signal sample, the eq. (3) is rewritten by

$$\vec{B}(t) = B[|cos(2\pi ft)| \quad sin(2\pi ft) \quad 0]^T \tag{6}$$

Frequency and magnitude of the oscillating magnetic field in the eq. (6) are adjustable by varying magnitude and direction of the electric input current supplied into each electromagnetic coil.

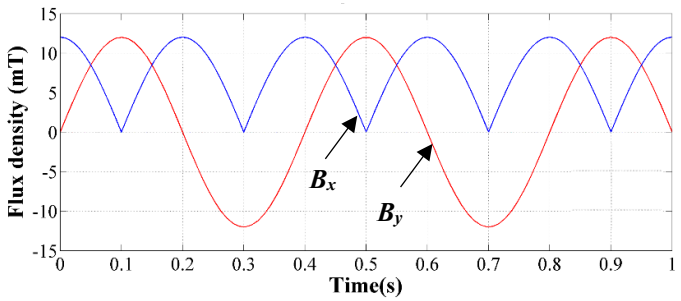


Figure 4: Signal sample of the x- and y-magnetic field, B_x and B_y, generates oscillating magnetic field with 2.5 Hz.

2.4. Modeling of deformation

Regarding the soft property of robots, Euler-Bernoulli beam theory is adopted to determine the local body deformation caused by magnetic field, depicted in Figure 5. Once magnetic torque, T, as bending moment, M_b, is exerted to the robot, expressed by

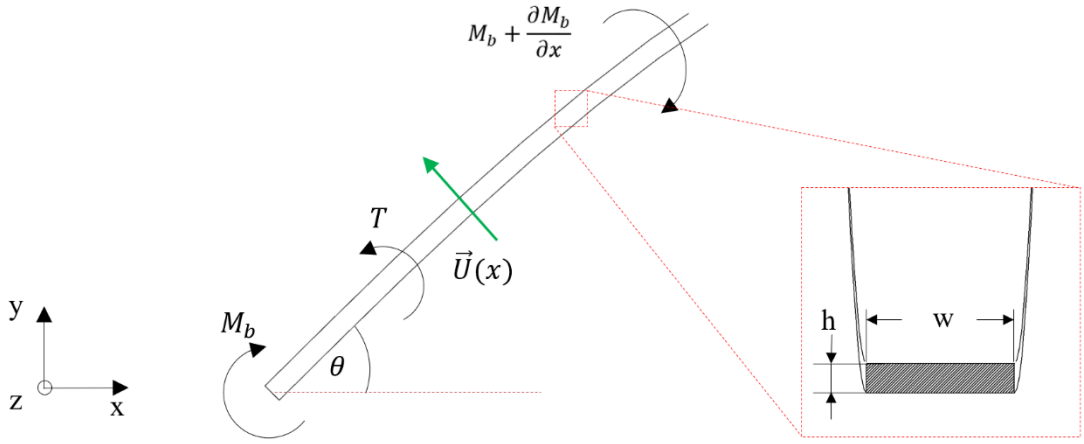


Figure 5: Physical interpretation of a deformable structure. When magnetic torque is exerted to deform a soft robot long, l, its local body segment, x, with thickness, h, and width, w, is bent with angle, θ, resulting in the moving of each local segment with velocity, $\vec{U}(x)$.

$$\boldsymbol{T}(x) = \boldsymbol{M_b} = -EI\frac{\partial^3\theta}{\partial x^3} \quad (7)$$

where E is Young modulus of the robot (Pa), θ is bending angle defined by deformation degree over length estimated by $\theta = \frac{\partial y}{\partial x}$, and I is area moment of inertia (m^4) which is a function of the thickness, h, and width, w. When the body of robot is long, l, the area moment of inertial at each local body becomes a function of the unit length, x, of the full length, l, expressed by

$$I(x) = \frac{h^2hw}{12} = \frac{h^2A}{12} \quad (8)$$

where A is cross-sectional area at the x position on the body length l. Substituting the eq. (5) and the eq. (8) in the eq. (7), so

$$(\vec{\boldsymbol{m}}(x) \times \vec{\boldsymbol{B}}) = -E\frac{h^2A}{12}\frac{\partial^3 y}{\partial x^3} \quad (9)$$

$$y(x) = -\iiint \frac{12}{h^2EA}(\vec{\boldsymbol{m}}(x) \times \vec{\boldsymbol{B}})\partial x^3 \quad (10)$$

As mentioned, considering that the whole body of the robot consists of many magnetic domains. Magnetic moment, $\vec{\boldsymbol{m}}$, at a specimen or a local is a function of magnetization and cross-sectional area. The eq. (10) expresses that deformation of a specimen at the position x, on the body length depends on magnitude and direction of magnetic moment significantly. Differentiating the eq. (10) over time to obtain swimming velocity, $\vec{\boldsymbol{U}}$, at that local, is expressed by

$$\vec{\boldsymbol{U}}(x) = -\frac{3fBM\lambda^3}{\pi^2h^2EA}\left[\sin\left(\frac{2\pi x}{\lambda} - 2\pi ft\right)\right]$$

and

$$\vec{\boldsymbol{U}}(x) = +\frac{3fBM\lambda^3}{\pi^2h^2EA}\left[\sin\left(\frac{2\pi x}{\lambda} + 2\pi ft\right)\right] \quad (11)$$

where f, B, M, λ, E, t is the oscillating frequency, magnetic field, magnetization, body wavelength, Young modulus of material, oscillating time respectively. Positive and negative sign express the deformation direction of a specimen of the body. From the eq. (11), in short, the swimming velocity, $\vec{\boldsymbol{U}}$, is proportional to magnitude and direction of magnetic moment which is a function of magnetization, M, at that specimen.

3. Experiments and Results

As mentioned, a key to have an effective swimming in fluid for a small-scaled robot is the use of an asymmetric body deformation, and the robots in this work follow such that concept utilizing a deformable structure triggered by magnetic field. The magnetic-soft robots are fabricated into three types, according to Table 2, and they are experimentally investigated in the term of swimming behavior and performance.

Experiments are all set up under the same parameters and conditions. Three types of the robots are all fabricated under the same process, but different in the post-fabrication, which is the cutting process to define quantity of magnetic moments via the final shape, and the magnetizing process to program the orientation of magnetic moments. A tank contains silicone oil to simulate viscosity of biological fluid, and it is inserted into the bore of the magnetic actuation system, depicted in Figure 3b. The system generates the oscillating magnetic field with three numbers of magnitude (5, 10, 15 mT) and fifteen numbers of frequency (1 to 15 Hz) to manipulate three types of the soft milli-robot.

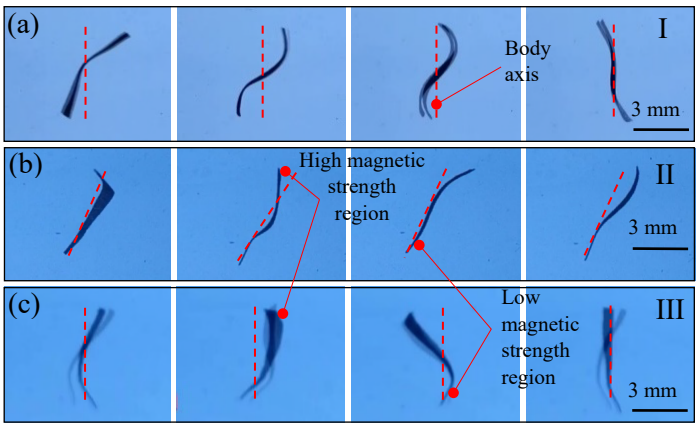


Figure 6: Swimming behavior of three robots holding different magnetic property. (a) The I robot symmetrically deforms as a wave. (b) The II robot sweeps at a specimen of high magnetic strength. (c) The III robot propagates a body wave along the body length.

3.1. The I robot

The I robot has a property of uniform magnetic strength and non-uniform magnetic orientation due to having a rectangular film shape (1 mm × 6 mm × 80 μm) and being magnetized with the sine profile, detailed in Table 2. Thus, its magnetic property is modeled by

$$\vec{\boldsymbol{M}}_I(x) = [M_x \quad M_y \quad 0]^T = M_I\left[cos\frac{2\pi x}{\lambda} \quad sin\frac{2\pi x}{\lambda} \quad 0\right]^T \quad (12)$$

The eq. (12) clearly details that magnetic orientation of the robot follows the sine profile depicted by the direction of arrows varying as a wave, and magnetic strength is uniform, depicted by number of arrows which is symmetric at each column along the body length.

Depicted in Figure 6a, once the I robot is actuated by the oscillating magnetic field, the magnetic moments embedded in the soft structure aligns with respect to the direction of the magnetic field, leading to the body deformation, and the deforming shape of the robot turns to be a sinusoidal curve. Next, when the direction of magnetic field oscillates with a frequency, the body of the robot is continuously deformed with respect to the oscillation of magnetic field under the same frequency, resulting in a lateral undulation, leading to swimming in fluid. It is noticed that the deformation of the robot is symmetric, and all of magnetic specimen in the structure simultaneous responds to the magnetic field at the same time because magnetic strength in each specimen is uniform.

3.2. The II robot

The II robot holds a property of non-uniform magnetic strength and uniform magnetic orientation, due to having a triangular shape (1 mm × 6 mm × 80 μm), detailed in Table 2. Number of magnetic moments in the triangular shape decreases along the body length, according to the eq. (2), resulting in non-

uniform strength across the whole structure. The larger the volume, the stronger the magnetic strength. It is magnetized to have the direction of magnetic moments aligned in the same direction, resulting in uniform magnetic orientation. Its magnetic property is expressed by

$$\vec{\boldsymbol{M}}_{II}(x) = [0 \quad M_y \quad 0]^T = M_{II}[0 \quad 1 \quad 0]^T \qquad (13)$$

The eq. (13) expresses that magnetic strength of the II robot is a function of a position along length *l*, and the magnetic moments all point to the y direction. Under this condition, once the robot is magnetically actuated, a specimen has a stronger magnetic strength firstly respond, and when the magnetic field is stronger, a specimen has a lower magnetic strength orderly follows. Thus, under the oscillating magnetic field with constant magnitude, the robot sweeps its body with respect to the oscillation of the magnetic field, making the robot propel as shown in Figure 6b.

3.3. The III robot

The III robot has a property of both non-uniform magnetic strength and orientation. Same to the II robot, number of magnetic moments varies along the body length due to having a triangular shape. It is maximum at the largest cross-sectional area, but minimum at the smallest area, resulting in non-uniform magnetic strength. The direction of magnetic moments aligned with respect to a sine curve, same to the I robot, resulting in non-uniform magnetic orientation. Its magnetic property can be expressed by

$$\vec{\boldsymbol{M}}_{III}(x) = [M_x \quad M_y \quad 0]^T = M_{III}\left[cos\frac{2\pi x}{\lambda} \quad sin\frac{2\pi x}{\lambda} \quad 0\right]^T \qquad (14)$$

The eq. (14) explicitly expresses that both magnetic orientation and strength are varying along the body length. Once the III robot is actuated by magnetic field, its body deforms as a curve, but the magnetic response of specimen where contains a higher magnetic strength is prior to another specimen where contains a lower magnetic strength. Under the oscillating magnetic field, it propagates the body as if a body waving for swimming interestingly and smoothly, depicted in Figure 6c.

3.4. Comparison in swimming behavior and performance

Under the actuation of oscillating magnetic field, each of robots (I, II, III) shows its own specific mechanism to swim in fluid. Magnetic alignment of magnetic moments embedded in the soft structure of each robot draws the deformable pattern differently due to being magnetized with the different profile. Sine profile magnetization causes a body wave propagation for the robot, resulting in more effective swimming, as appeared in the I and III robot. However, in experiments, one more interestingly magnetic aspect in the soft structure is figured out that quantity of magnetic moments has an influence to define the response degree to external magnetic actuation. If higher, magnetic strength is stronger, resulting in a fast response. If lower, magnetic strength is weaker, resulting in a slow response. When both high and low magnetic strength are together in one soft structure, leading to unequal magnetic strength of each specimen. Magnetic response of the entire body is different, and active orderly from the high to low magnetic strength. As apparent in the magnetic response of the I and III robot, both is profiled by sine pattern, but magnetic strength is different. The I robot transforms the entire soft body to

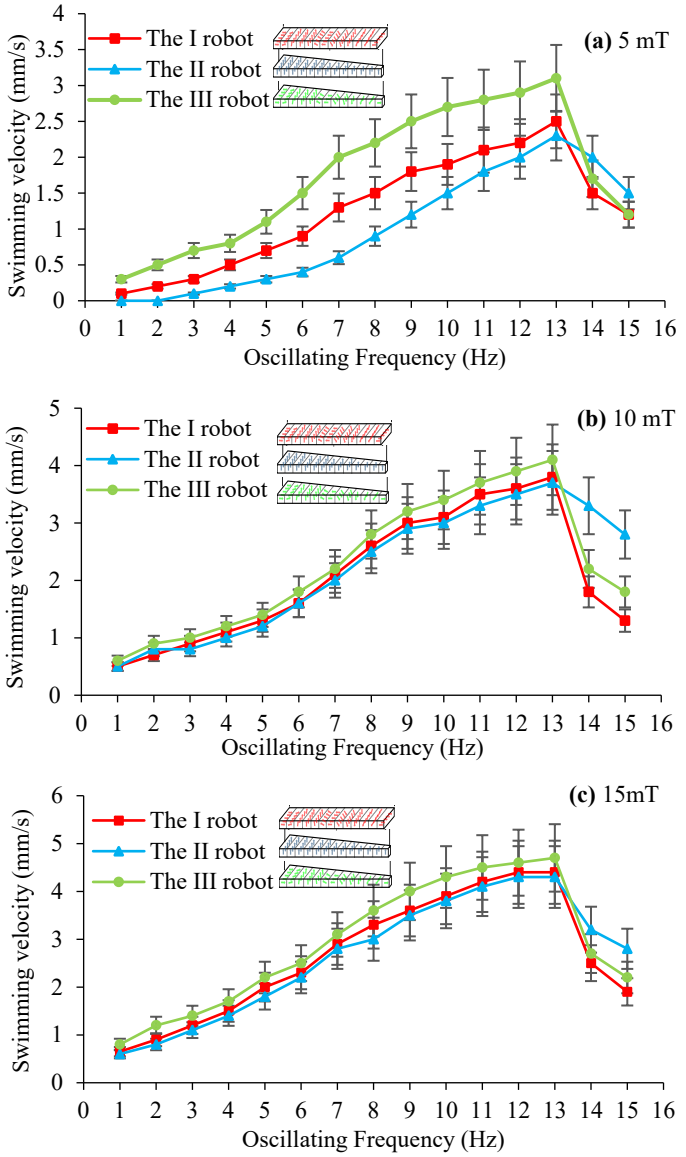


Figure 7: Under (a) 5 mT, (b) 10 mT, (c) 15 mT of magnetic field, the plot of swimming velocity against the different numbers of the oscillating frequency (ranged from 1 to 15 Hz). Three robots are powered by those input parameters to swim in fluid. The robots all swim faster with the increase of the actuating frequency, and drops after reaching the step-out point at 13 Hz.

a curve pattern immediately, but of the III robot, a specimen containing a higher magnetic strength is firstly attractive to magnetic field, but another with a lower strength follows orderly. Moreover, the III robot is better in manipulation because it allows another control of dynamic magnetic field. The rotating magnetic field can drive the robot to swim curvilinearly along clockwise or counter-clockwise direction. (Supplementary video 01)

In order to compare performance of three robots embedded with the distinguish magnetic property, swimming velocity is plotted against actuating input parameters; three numbers of magnetic field (5 mT, 10 mT and 15 mT) with fifteen numbers of the oscillating frequency (from 1 to 15 Hz), shown in Figure 7. The results report that velocity tends to continuously increase by an increase of the gaining frequency. The plot still expresses that the III robot can swim fastest at any magnitude of magnetic field,

including the oscillating frequency. All of the robots have the step-out point of the oscillating frequency about 13 Hz at all ranges of magnetic field. At this point, the robot lost in synchronization to the actuating magnetic field, resulting in lack of control and dramatical decrease in velocity. However, swimming velocity of the II robot does not drop fast interestingly if comparing with the others because its body deformation caused by magnetic alignment is the simplest, pointing toward only one direction. Compared with the I and III robot profiled with the sine-curve manner, the II robot's deformation has least lost in synchronization to the magnetic field. After the step out point at 13 Hz, it turns to be a better swimmer than others. Interestingly, at the lowest magnetic field; 5 mT, the III robot still shows a better control and performance than others at all ranges of the actuating frequency, including having the fastest response and performance even at the low oscillating frequency whereas the others cannot swim out due to lost in the magnetic synchronization.

4. Conclusion

In the research of biomimetic robots, fish-like swimming relies on a flexible body to achieve the higher performance, but this mechanism aspect is limited with the size of the robot down to millimeter or less. The use of magnetic field can be a solution to such that problem promisingly. Remotely magnetic manipulation of tiny robots is an effective and non-harmful technique to deal with biomedical applications in which the small-scaled swimming robots can contribute tremendous results as medical devices. In this works, the combination of both magnetic actuation and flexible structure is presented with the purpose of medical application. It enables the wireless power of the external magnetic field generated by the magnetic actuation system. This concept is beneficially applicable to the small-scale robot to employ motor less-mechanism. Three types of soft swimming robots are fabricated with distinguish magnetic property. Their swimming behavior and performance is investigated under the actuation of the oscillating magnetic field. The magnetic alignment of magnetic moments embedded in the deformable structure of the robot leads to maneuverability in fluid under the minimal control of the oscillating magnetic field. The experimental results report that the swimming performance of the robot mainly relies on two parameters; firstly, the strength of magnetic field to adjust the amplitude of the body deformation, and secondly, the frequency to gain the rate of the body deformation. Finally, we found out that quantity and orientation of magnetic moments in the soft structure function in swimming behavior and performance differently. Either of them can be employed solely to create a motion mechanism of the small-scaled robots. The contribution of this study is wide-opening and promising for a soft small-scaled robot to serve multi-purposes towards biomedical applications.

5. Discussion

There are several concerns to avoid error of the data collection, and prepare the robot to be ready for the application. During the experiments, the swimming velocity data of each robot at all input parameters is critical and needs a precise measurement in order to compare the swimming performance between those robots properly. Object tracking is applied to the camera mounted on the top of the magnetic actuation system, and the center of robot is captured to obtain the coordinate. Another issue of the work would be about the future work. According to the fabrication process, the robot allows us to add more functions to make the robot greater in performance. We definitely plan to extend this study to an advance experiment such as in-vitro and in-vivo experiment. Even though the results of the study are reliable, the robot still needs to improve biocompatibility. Typically, the magnetic active element used as a main component to fabricate the soft robot is not biocompatible. Consequently, the robot is partially biocompatibility. However, the use of biocompatible polymer (e.g., PEG, hydrogel) to wrap the robot is possible, and it does still not constraint the deformation degree actuated by magnetic field. Another would be about how to image the robot inside the blind area. There are two possible methods; Ultrasound imaging using a probe to detect the robot and PA (Photoacoustic) imaging visualizing the robot via the excited signal from additional components. Both imaging techniques can track the robot accurately. If these issues are managed properly, the robots will be ready for the medical applications.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work is supported by the Research Fund of Faculty of Engineering, Thammasat School of Engineering, Thammasat University, the Thammasat University Research Fund No. TUFT 49/2566, the Thammasat Postdoctoral Fellowship No. TUPD 19/2565, and the Thailand Science Research and Innovation Fundamental Fund fiscal year 2023.

References

[1] X. Tang, L. Manamanchaiyaporn, "Magnetic-Powered Swimming Soft-Milli Robot Towards Non-Invasive Applications," The eighth (8th) edition in the series of the International Conference on Control, Decision and Information Technologies (CoDIT'22), 1562-1566, 2022, doi: 10.1109/CoDIT55151.2022.9804108.

[2] B.J. Nelson, I.K. Kaliakatsos, J.J. Abbott, "Microrobots for Minimally Invasive Medicine," Annual Review of Biomedical Engineering, **12**, 55–85, 2010, doi: 10.1146/annurev-bioeng-010510-103409.

[3] L. Manamanchaiyaporn, T. Xu, X. Wu, "An Optimal Design of an Electromagnetic Actuation System towards a Large Homogeneous Magnetic Field and Accessible Workspace for Magnetic Manipulation," Energies, **13** (4), 911, 2020, doi: 10.3390/en13040911.

[4] L. Manamanchaiyaporn, T. Xu, X. Wu, "Magnetic soft robot with the triangular head-tail morphology inspired by lateral undulation," IEEE/ASME Transactions on Mechatronics, **25** (6), 2688-2699, 2020, doi: 10.1109/TMECH.2020.2988718.

[5] S. Palagi, P. Fischer, "Bioinspired microrobots," Nature Review Materials, **3**, 113-124, 2018, doi: 10.1038/s41578-018-0016-9.

[6] L. Zhang, J.J. Abbott, L.X. Dong, B.E. Kratochvil, D.J. Bell, B.J. Nelson, "Artificial bacterial flagella: fabrication and magnetic control," Applied Physics Letters, **94**, 2009, doi: 10.1063/1.3079655.

[7] H. Huang, M.S. Sakar, A.J. Petruska, S. Pane, B.J. Nelson, "Soft micromachines with programmable motility and morphology," Nature Communication, **7**, 12263, 2016, doi: 10.1038/ncomms12263.

[8] E. Diller, J. Zhuang, G.Z. Lum, M.R. Edwards, M. Sitti, "Continuously distributed magnetization profile for millimeter-scale elastomeric undulatory swimming," Applied Physics Letters, **104**, 2014, doi: 10.1063/1.4874306.

[9] G.Z. Lum, Y. Zhou, X. Dong, H. Marvi, O. Erin, W. Hua, M. Sitti, "Shape-programmable magnetic soft matter," PNAS, **113**(41), 6007–6015, 2016, doi: 10.1073/pnas.1608193113.
[10] W. Hu, G.Z. Lum, M. Mastrangeli, M. Sitti, "Small-scale soft-bodied robot with multimodal locomotion," Nature, **554**, 81-85, 2018, doi: 10.1038/nature25443.
[11] M. Cianchetti, C. Laschi, A. Menciassi, P. Dario, "Biomedical applications of soft robotics," Nature Review Materials, **3**, pp.143-153, 2018, doi: 10.1038/s41578-018-0022-y.
[12] M. Sitti, "Miniature soft robots - road to the clinic," Nature Review Materials, **3**, 74–75, 2018, doi: 10.1038/s41578-018-0001-3.
[13] L. Manamanchaiyaporn, X. Tang, X. Yan, Y. Zheng. "Molecular Transport of a Magnetic Nanoparticle Swarm Towards Thrombolytic Therapy," IEEE Robotics and Automation Letters, **6**(3), 5605-5612, 2021, doi: 10.1109/LRA.2021.3068978.
[14] E. Gultepe, J.S. Randhawa, S. Kadam, et. al, "Biopsy with Thermally-Responsive Untethered Microtools", Advanced Materials, **25**, 514–519, 2013, doi: 10.1002/adma.201203348.
[15] W. Zhu, J. Li, Y. J. Leong, et. al, "3D-Printed Artificial Microfish", *Advanced Materials,* **27**, 4411–4417, 2015, doi: 10.1002/adma.201501372.
[16] C.W. de Silva, S. Xiao, M. Li, C.N. de Silva, "Telemedicine-Remote Sensory Interaction with Patients for Medical Evaluation and Diagnosis," *Mechatronic System and Controls*, **41**, 2013, doi: 10.2316/Journal.201.2013.4.201-2536.
[17] E.M. Purcell, "Life at low Reynolds number," American Journal of Physics, **45**, 3-11, 1977, doi: 10.1119/1.10903.
[18] N. A. Spaldin, "Magnetic Materials Fundamentals and Applications," 2nd ed. Cambridge University Press, USA, 2010.

Advances in Science, Technology and Engineering Systems Journal Vol. 8, No. 3, 218-230 (2023)

www.astesj.com

Special Issue on Computing, Engineering and Multidisciplinary Sciences

ASTESJ

ISSN: 2415-6698

How a Design-Based Research Approach Supported the Development and Rapid Adaptation Needed to Provide Enriching Rural STEM Camps During COVID and Beyond

Rebecca Zulli Lowe[1], Adrienne Smith[1,*], Christie Prout[1], Guenter Maresch[2], Christopher Bacot[2], Lura Murfee[2]

[1]*Cynosure Consulting, Apex, 27502, USA*

[2]*North Florida College, Madison, 32340, USA*

ARTICLE INFO

Article history:
Received: 28 February, 2023
Accepted: 20 May, 2023
Online: 25 June, 2023

Keywords:
STEM
Summer Enrichment
Virtual Instruction

ABSTRACT

Like many STEM research projects, the members of the National Science Foundation-funded STEM SEALS project dramatically shifted from in-person delivery of a summer institute to distance-learning with minimal time for preparation. However, the daunting challenge also offered the unique opportunity to apply Design-based Research within an exploratory study to inform and document the progression and supply counsel to other STEM providers contemplating a shift to a virtual platform. The goals of this exploratory study include (1) to make apparent the barriers to transitioning to virtual STEM enrichment programming in rural spaces during the COVID-19 pandemic, (2) detail important decisions made in the move online, along with the reasoning behind those decisions, and (3) share best practices that arose during the inaugural effort. Methods included the review of multiple data sources, including project meeting minutes, educator reviews of materials, and pre/post institute student and teacher surveys. to inform rapid-paced learning cycles. As a result, the team adopted a mindset that focused on high-quality STEM experiences. Strategies supported by the research include effective substitutes for in-person demonstrations, leveraging existing platforms, employing mechanisms for troubleshooting, and framing failure in ways that encouraged the development of a positive STEM identity.

1. Introduction

This paper is an extension of work originally presented at the 2021 Integrated STEM Education conference [1]. It utilizes state of the art research methods in the employment of an exploratory research study to understand best practices in transitioning high quality STEM learning environments from in-person to virtual.

Due to the COVID-19 outbreak, and the World Health Organization (WHO) officially labeling COVID-19 as a pandemic [2], K-12 education in the United States would change dramatically. When stay-at-home orders started going into effect in many states, many public schools were forced to close their doors and move from in-person instruction to online teaching. Soon after, many STEM educators and researchers, including the National Science Foundation-funded STEM Sea, Air, and Land (SEALs) team from North Florida College (NFC) in Madison, Florida, began to realize the pandemic would not be easily or quickly be curbed.

In early 2020, the STEM SEALs team was planning for some highly engaging educational activities for both educators and students in their six-county service region. The STEM SEALs team was led by STEM experts from NFC, a rural community college. The team included educational researchers from Cynosure Consulting, LLC (Cynosure). These activities included opportunities for participants to build rovers, boats, and drones, which they would learn to code and then maneuver to complete fun, yet rigorous engineering design challenges as part of a weeklong STEM camp hosted on the NFC campus. This inaugural camp would expose rural middle school students to hands-on engineering and computer science experiences. Earlier in the fall of 2019, the STEM SEALs team recruited nine middle school educators from the surrounding counties. This group formed the design team and spent one Saturday a month together, where they tested out the curriculum and shared input on the structure and design of the student experience.

As the pandemic began to unfold, it interrupted the project's spring plans in which the STEM SEALs team was in the middle of organizing. In March of 2020, the STEM SEALs team was

* Corresponding Author: Adrienne Smith, 1302 Applethorn Drive, Apex, NC 27502, 919-616-1565, adrienne.ann.smith@gmail.com

scheduled to host a large group of educators on the NFC campus to participate as a review team. As review team members, local educators would learn about the STEM SEALs experience content, give ratings as part of the feasibility testing, and offer critiques that would continue to inform curriculum development and revision. With the date of the on campus large review team meeting approaching, it was clear the impacts of COVID-19 were only increasing. Each day was met with considerable conversations and discussions, which started as predictions about when in the spring the review team would meet as planned to whether it would meet virtually or not at all. Sadly, the STEM SEALs team had to accept that holding a virtual review team was not a viable alternative. The decision was made to postpone the spring review team meeting altogether. As the COVID-19 numbers surged in the US and Florida's own cases began to spike, it became evident that the inaugural STEM SEALs summer camp was in jeopardy of cancellation.

Finally, on April 24th, the STEM SEALs team was compelled to make a decision about the fate of the inaugural STEM SEALs enrichment camp that was planned for June. From a national perspective, it became clear that the COVID-19 pandemic was affecting many informal STEM experiences scheduled for the summer. Despite months of developing curriculum materials, testing out prototypes, and carefully arranging an in-person STEM SEALs camp with safety as a priority, it became evident STEM SEALs could not overcome the effects of the pandemic. Knowing the importance of the project goals to learn about best practices in offering high quality STEM exposure and interest for middle grades students in rural areas, the STEM SEALs team reluctantly made the decision to go virtual.

1.1. The Need to Push Forward

Underrepresentation of Rural Students in STEM

The underrepresentation of rural students in STEM is not a recent phenomenon, but in the past two decades, the issue has been receiving greater attention. Studies and literature reviews have advanced the field's understanding of the barriers that rural students face. Rural students struggle with issues of geographic separation and insufficient internet bandwidth to support online access and complete access to many technological advances [3]. They face limited opportunities to engage in advanced coursework in mathematics and science [4], and financial obstacles that limit future employment and educational prospects [5]. School administrative data have shown low participation in advanced coursework among low income, rural students in comparison to students from high-income families. The result is an excellence gap that is evidenced as early as elementary school and persists through high school [6]. Education researchers assert that this excellence gap "represents a growing crisis requiring programmatic intervention" [7]. Students in rural schools, particularly those that are under-resourced, are less likely to reach advanced levels of academic achievement compared with their urban peers, even when they demonstrate high potential [8].

Challenges and Strategies for Rural Students

Rural students that show high potential for academic success confront barriers that limit options for academic acceleration, putting them at risk of becoming part of the "persistent talent underclass" [9]. Researchers studying this excellence gap have identified innovations that can mitigate access to advanced coursework in high school. They point to programming outside of school time designed for middle school students as a potential stop gap measure [7]. These kinds of programs are advantageous for multiple reasons [6]. Spending time socializing outside of school increases positive peer interactions and stimulates social development, in addition to academic learning for middle school students [10, 11]. These benefits are larger for at-risk students, for whom researchers have documented a link between extracurricular programming and educational success [12].

Not only do informal educational experiences serve as mitigating factors for poor academic outcomes, but they can also serve as a catalyst for the decision to seek advanced coursework in high school [6]. Researchers voice that for the strategy of improving high-potential rural students' STEM achievement through extracurricular programming to work, implementers "must also consider the inclusivity of identification models for such programming" [7]. This STEM SEALs project set out to model the potential efficacy of a widely inclusive outreach strategy with the purpose of reaching a broad pool of rural, high-potential students who are ready for STEM development opportunities.

1.2. Guiding Framework

The STEM SEALs project was always designed to be more than the creation of a high-quality STEM camp. The work of STEM SEALs was nested within a larger research design focused on efforts to develop rural STEM education pathways and building an evidence base for the emerging strategies and materials with the larger vision of creating broader access to high quality STEM experiences for students in rural parts of the country. The STEM SEALs project was not simply an outreach or STEM enrichment project, but instead, STEM SEALs was from the outset framed as Design-based Research [13, 14]. Design-based Research has been widely used in education, and curriculum development, in particular, where research and design activities are often inseparable parts of improving current practices and refining design theories and principles [15]. It has also been used extensively for researching and improving professional development [16]. This systematic methodology aims to improve educational practices through a cyclical process that involves iterative periods of design, testing, evaluation, and reflection between researchers and practitioners conducted in real-world environments [17].

Design-based Research has its roots in a larger movement near the beginning of the 21st century that looked to more effectively bridge the gap between research and practice. It acts as a practical methodology which serves dual roles for both developing and informing learning theory and the means designed to support that learning [13, 14, 18, 19, 20, 21]. In fact, it has been used to improve STEM education in a variety of ways, including developing new curricula and instructional materials for teaching science online to middle students [22]. In this example, the new curriculum was designed to engage students, including English-language learners and students with a disability. The Design-based Research process utilized data from multiple sources, including teacher logs, student and teacher surveys, and focus groups. Results showed the developed curriculum to be feasible, useful, and effective with a diverse student population. It also demonstrated that Design-based Research is a practical framework in such settings. The Design-based Research method has also been used in designing forensic science games for middle school students [23] and developing assessment tools for measuring

students' science critical thinking skills [24]. Supporting teachers in implementing new STEM curricula and instructional materials is another context that has utilized Design-based Research methods. In a recent example, a study by the authors in [25] used Design-based Research to develop a professional development program for supporting teachers in developing children's spatial reasoning. The professional development program was designed to help teachers understand the principles of the curriculum and to develop the skills they needed to implement it effectively. Researchers found Design-based Research to be a catalyst for epistemic change. Overall, Design-based Research is a promising methodology for improving STEM education. It is a flexible and iterative approach that allows for the development and refinement of interventions in real-world settings.

The structure of this Design-based Research study was bolstered through the use of a modified version of the Successive Approximation Model (SAM) [26] to ensure the iterative development process (a) occurs in small steps in association with ongoing evaluation that informs iterative changes, (b) supports productive collaboration among project team members, (c) directs energy and resources effectively in order to move efficiently with intervention development, and (d) allows for manageable completion of high quality projects both on time and on budget. Developers cycle through phases of analysis, design, and development supported by embedded research that routinely provides formative assessment and input to inform the ongoing development efforts. Central to this approach is the use of an iterative "development-revision-testing" process with teachers and students to ensure materials and activities are understandable, appropriate, and engaging.

Design-based Research has been heralded as a research approach that could help those looking to fill the research-practice gap by bringing educational research closer to the needs of educators in the field. The use of this approach by the current project provided clear evidence that supports this contention, with the STEM SEALs use of the Design-based Research approach emerging as a highly effective model for promoting the rapid innovation and adaptation needed to develop, implement, and build momentum under the spectra of the COVID-19 pandemic. Ultimately, the STEM SEALs research project outcomes and findings provide significant evidence in support of the effectiveness of Design-based Research framework for bolstering innovation. This article will provide an illustration of emergent innovations that resulted from its use by STEM SEALs to foster innovation within STEM enrichment in rural counties and a discussion of the key mechanisms of the approach associated with the bolstering of project success.

1.3. The Phases of Design-Based Research

The Design-based Research process "consists of four phases: (1) analysis of practical problems by researchers and practitioners working in collaboration; (2) development of new solutions informed by existing design principles; (3) iterative cycles of testing and refinement of solutions in practice; and (4) reflection to produce design principles and enhance solution implementation [27]. Each of these phases was operationalized by the project. See overview in Figure 1.

Phase 1: Analysis of practical problems by researchers and practitioners working in collaboration

Assemble Diverse Team of Researchers and Practitioners. To address the practical problems that emerged, a project leadership team was assembled with individuals bringing different expertise and skillsets. The STEM SEALs team included members with first-hand experience as STEM teachers in the rural area, education researchers, and content experts from a rural community college representing the fields of Physics, Engineering, Biology, and Advanced Manufacturing.

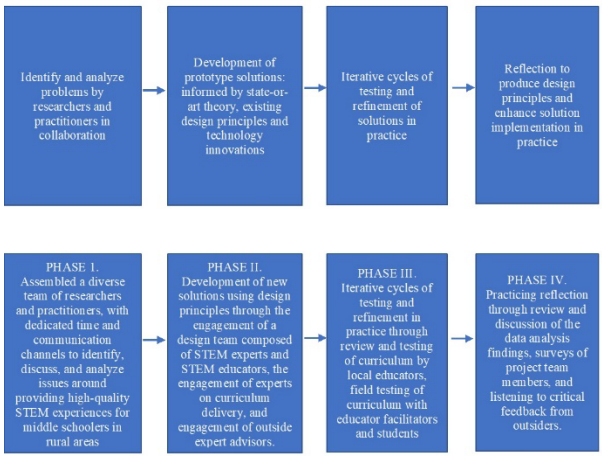


Figure 1: Design-based Research Process Operationalized

Routine and Ongoing Communication with Diverse Team around Issue of the Need for High-quality STEM Experiences in Rural Spaces at the Middle Grades Level. Through weekly meetings and ongoing project communication and activities, the team engaged in analysis of the existing problems of practice of access to STEM enrichment and capacity deficits in rural areas to come up with the multi-faceted model for addressing those issues within the context of designing a STEM experience that serves the needs of both students and teachers.

Phase 2: Development of new solutions informed by existing design principles

The second phase was addressed by the collaborative STEM SEALs team through engaging in iterative intervention development guided by the SAM model.

Design Team Collaboration for Curriculum Development. A key task for the project was creating instructional content that leverages an additional and important group of collaborators – current middle school educators within the rural area. To do this, a design team was comprised of nine rural educators with expertise in middle grades STEM instruction who hail from the five counties within the NCF service area. Design team members attended three day-long meetings where they listened to STEM experts, offered input, engaged with materials, and provided recommended changes in the development of the student institute experience. Design team members provided written and verbal feedback and observation of their use of the materials served as additional data to inform the design. Evaluators and research members from the STEM SEALs team spoke with members of the design team to further understand and be responsive to the ideas raised. The specific input received from content experts, best-practice research, and individuals who work with students in authentic education settings provided a strong foundation for the full scope of curriculum, although the details were subject to change during iterative development.

Engagement of Different Experts to Inform Curriculum Delivery Design. Social psychology research provides a wealth of information on techniques and strategies for supporting underrepresented groups in STEM. Led by Cynosure's social science and education researchers, the research team, a subset of the STEM SEALs project team, identified opportune moments within the sequence of the module delivery to elicit conversations and address misconceptions related to who does STEM and what a STEM career looks like. Additionally, activities supported by best practices in the education research literature were incorporated into the modules delivery to bolster the confidence and self-efficacy of rural students as well as develop a growth mindset and sense of identity within STEM. For example, students will watch as the experts troubleshoot, seeing productive failure as an integral part of the engineering design process.

Consultation with External Advisors. The project leaders also consulted with advisory board members to solicit feedback on the process of developing high quality materials and designing engaging STEM experiences. Advisory board members brough a wealth of expertise on informal STEM and the intersection between faculty, practitioners, and rural populations. Their feedback was used to refine drafts to maximize implementation with fidelity.

Phase 3: Iterative cycles of testing and refinement of solutions in practice

Feasibility Testing with Local Educators. To ensure the curriculum materials were understandable, appropriate, and useful for the intended population, middle grades teachers from the region were recruited to review and react to several of the developed modules planned for the institute. Reviewers were given the physical supplies and curriculum materials associated with each module to work through independently (with a STEM SEALs team member available for questions and trouble shooting.) After each module review, the reviewers completed a form where they rated the module, reflected on the feasibility of the module within a real-world summer institute with middle school students, and provided big picture and detailed feedback along with recommendations to improve the materials. Within the feedback, participants rated the intervention (from 1=Strongly disagree to 5=Strongly agree) to assess: (a) ease of use, (b) innovation, (c) value and need, (d) feasibility, (e) potential effectiveness for achieving intended goals, (f) usability, (g) advantages over existing methods; and (h) overall quality. Educators also rated the degree to which they: (a) would recommend the proposed intervention to schools, (b) would use the intervention themselves, (c) believe the intervention would be effective for preparing students in STEM, and (d) recommend continued development and testing.

As part of the review process, all reviewers were invited to the NFC campus to take part in the module activities during a day-long, more in depth exploration of the materials with the content experts. The purpose was two-fold. First, by implementing pieces of the institute with middle school educators the STEM SEALs team could gather feedback on the appropriateness of the language used as well as the assumptions made about students' pre-requisite content knowledge. Implementers would glean a stronger understanding of the extent to which guidance is needed for handling lab equipment and hear strategies for helping students stay on task and support them in their learning. Second, through the reactions and questions of teachers to the modules, implementers could assess the extent to which teachers have mastered the content and gain a better awareness of the content information that will be needed in module curriculum facilitator guides.

Lastly, review team members were invited to attend a focus group to gather quantitative and qualitative evaluation data. These data helped the team ascertain whether educators (a) view the intervention as demonstrating high quality, innovation, and value; (b) advocate use of the intervention as feasible and needed for schools; and (c) recommend continued development and testing. During the focus group sessions, the project team member also led group discussion to gather specific comments and suggestions, including review of the implementation guidelines to gather information on potential feasibility and fidelity challenges.

Data were analyzed by the research team to assess the degree to which the curriculum materials are acceptable. If any module failed to meet the team's standards, it was revised accordingly based upon feedback. Examples of revisions included removing confusing elements or adding clarifying directions for equipment use or assembly, substitution of more simplified code, and addition of videos or other resources to extend the learning.

Field Testing with Educator Facilitators. Local middle school educators were invited to serve as facilitators of the summer institute, under the direction and support of the community college content experts. During and upon the conclusion of the summer institute, facilitators provided feedback on the experience. Mechanisms were in place to collect data by researchers, evaluators, and content experts and designers. These feedback mechanisms allowed for just-in-time adjustments to the experience and served as a record for changes and recommendations for future institutes.

Field Testing with Students. To assess the usability of the materials, students in middle grades were recruited to participate in summer institutes that allowed them to engage with the STEM SEALs materials and culminate with a design challenge. Feedback from students was gathered through informal interviews with students during the institute, observation of their affect and behaviors during the institute, a survey soliciting written ratings and recommendations at the end of the experience, and analyses of pre- and post-institute assessments of key anticipated outcomes.

The iterative design-develop-test process involved multiple testing cycles. Early tests allowed the STEM SEALs team to enact the module curriculums (and design challenge) with the intention of gathering feedback to inform further revisions. Later tests serve more as a pilot, that is, is a more formal testing of the revised modules (and design challenge) to examine whether the intervention elicits the intended outcomes.

Phase 4: Reflection to produce design principles and enhance solution implementation

Review and Discussion of Data Analysis Findings. Data from all the processes described were analyzed and then shared with the STEM SEALs team, who reflected and shifted as needed based on the data. This phase involved the use of deliberate reflection activities to ensure that sensemaking happens routinely around the contributions, insights, recommendations, and lessons learned.

Surveying Project Members as Mechanism for Reflection. For example, STEM SEALs team members completed reflection forms separately. These data were analyzed, and common themes

shared during team meetings to solicit further discussion and to form the basis for revisioning efforts.

Reviewing Outside Reactions to Further Stimulate Reflection. Additionally, the convening of an advisory board and an Expert Teacher Material Reviewer helped to engage in further synthesis, sharing, and outside review and reflection based on experts from the field.

2. Method

While Design-based Research was a lens woven into the fabric of our study originally, its application was indispensable during the fast-shifting events that followed the advent of the COVID-19 pandemic. The designers and implementers of the innovation had to make rapid decisions. It was unclear whether the situation would be replicated, and it felt important to be mindful when cataloguing decisions and their rationale under the current context. Suddenly, there were new questions raised, ones whose answers could quickly and meaningfully contribute to the field. The researchers on the STEM SEALs team chose to adapt the Design-based Research approach to create a rapid learning process to align with the quick pace of the pivots made by those implementing. The results were an approach that situated rapid cycles of iteration within a modified Phase IV Design-based Research strategy, with the aim of engaging in strategic data-informed efforts to successfully navigate the pandemic-mandated pivot from an in-person STEM enrichment event to a virtual offering. What ultimately resulted was a series of rapid Design-based Research cycles. Within each cycle the barrier is identified, solutions explored, alternatives analyzed, and a decision is made. See Figure 2.

Figure 2: Design-based Research Rapid Cycle Iteration Process

Ultimately, Design-based Research, on overdrive, if you will, became the methodological mechanism for both informing and capturing the important learning during this time.

Thus, the over-arching research question guiding this effort was, "To what extent does a Design-based Research approach support the development and rapid adaptation needed to provide enriching rural STEM experiences?"

To answer this question, an exploratory study (described next under "Exploratory Approach") was developed, guided by Design-based Research, in which existing data (that is, data collected for other purposes) were leveraged along with the collection of additional data, tailored to addressing the specific needs of the Design-based Research rapid cycle iterative design effort (described under "Data Sources").

2.1. Exploratory Approach

The findings presented in this article were generated utilizing a qualitative framework at the level of exploratory analysis with the goal of garnering insights on the emergent success of the approach at promoting adaptation and innovation associated with the project's success at navigating the challenges and barriers presented by the sudden shutdown of schools and bans on gathering associated with the initial months of the COVID-19 pandemic. Exploratory studies are a type of preliminary research that provides initial information and sense-making. It serves as a foundation for later efforts poised to yield more conclusive findings [28]. Selecting an exploratory design offers several advantages aligned to the goals of this study, including: 1) affording researchers flexibility to adjust and adapt as research advances; 2) permitting researchers to recognize parts of a phenomenon that merit additional study earlier in the process; and 3) assisting other researchers by identifying potential causes of and solutions to a problem, that may be more intensively studied in the future. Additionally, the potential to encourage rapid learning, increase uptake, and share knowledge with the field means the study approach can be responsive to the larger national need for expertise on how STEM enrichment providers can transition from traditional approaches to innovations such as virtual, asynchronous, or blended content delivery models.

2.2. Design-based Research Rapid Cycle Data Sources

Data were collected from various sources to make explicit and unpack the STEM SEALs team's reasoning before, during, and after making the decision to pivot to on online delivery. The utilized data sources include the following:

Project Team Meeting Participation and Minutes

The STEM SEALs team convened on a weekly basis in order to document the project's activities and accomplishments, discuss challenges and issues, make key decisions, and plot out upcoming tasks. Discussions were recorded and meeting notes summarized to document important decisions and their rationale.

Project Team Reflection Survey Responses and Transcript

The STEM SEALs team members were each given a reflective survey that they completed individually. These surveys were compiled and analyzed by the research team and shared formally through a group reflection activity. The analysis and the transcript from the reflection activity meeting served as data sources for this study.

Internal Review of Virtual Delivery Methods and Materials

Instead of engaging an outside review team during the rapid cycle, the research team asked the STEM SEALs members who had developed the materials to engage in a virtual walk through of the rover experience. This included facilitating a session where the research team coded the Micro:bit, assembled the rover, and manipulated the rovers to accomplish tasks. The review was conducted through the online Zoom platform for two days, March

27th and April 3rd. It served as a critical initial field test of the material implementation virtually.

Research Team Facilitated Group Review, Analysis and Decision-making

Data that were collected related to barriers identified were routinely analyzed and presented at STEM SEALs team meetings. Typically, the analysis summaries were provided in advance and explored together during a facilitated virtual meeting. Ultimately, the meetings ended with a decision on a solution or response to the identified barrier.

2.3. Design-based Research Rapid Cycle Product Test Stage (the Camp) Data Sources

Data were leveraged from several sources to serve as evidence of success of the product of the Design-based Research rapid cycle process. In this case, the product was the camp itself. An inventory of these data sources included:

Observation of Virtual Camp Experiences

The virtual STEM SEALs experience was held July 8th through 17th of 2020. The virtual camp involved resources cataloged and accessible to participants using the Google Classroom platform. Support was offered by NFC staff through live in-person demonstrations and discussions and as-needed asynchronous support. These support sessions were recorded as part of the observation data collection.

Data from Teachers

Data generated from pre- and post-camp surveys captured information on educator backgrounds and perceptions of STEM. Teachers also completed reflection surveys at the end of activities during the STEM SEALs summer camp. The purpose of these reflections was to better understand how teachers and their students experienced the materials and to collect any recommendations for improvement.

Data from Students

Data was also collected through student surveys. These surveys asked students about their background and knowledge of STEM at the end and the beginning of the camp. Students also completed end of module assessments to document their learning. Lastly, students took part in an end-of-camp reflection exercise where they were asked to look back on their time at the camp and provide their impressions of the experience as well as offer any recommended changes.

2.4. Data Analysis

This exploratory study, situated within the larger project, involved multiple layers of data analysis. Three different perspectives were adopted to address the research question: 1) to delineate the role that emerged for Design-based Research within rapid applications, 2) to describe the contributions and outputs of the rapid Design-based Research cycles on the resulting product, and 3) to test the resulting product. In addressing these aims, qualitative data analysis techniques were predominantly employed. The data analysis strategy varied based on its purpose within the rapid Design-based Research cycle. The data were synthesized systematically and then objectively analyzed using mechanisms that identified key themes. In some cases, quantitative analysis techniques were further incorporated through summarizing program tracking and survey assessment data. Survey data were analyzed with the statistical software Stata (Version 13) and descriptive statistics were computed. Data verification strategies were incorporated within and across the data analysis activities. Triangulation of findings was conducted such that meeting documentation, field records, and transcripts were referenced as the researchers on the team utilized an iterative process of detecting and categorizing emerging themes, then cross-checking those themes with the various sources of data for confirming evidence.

3. Results and Discussion

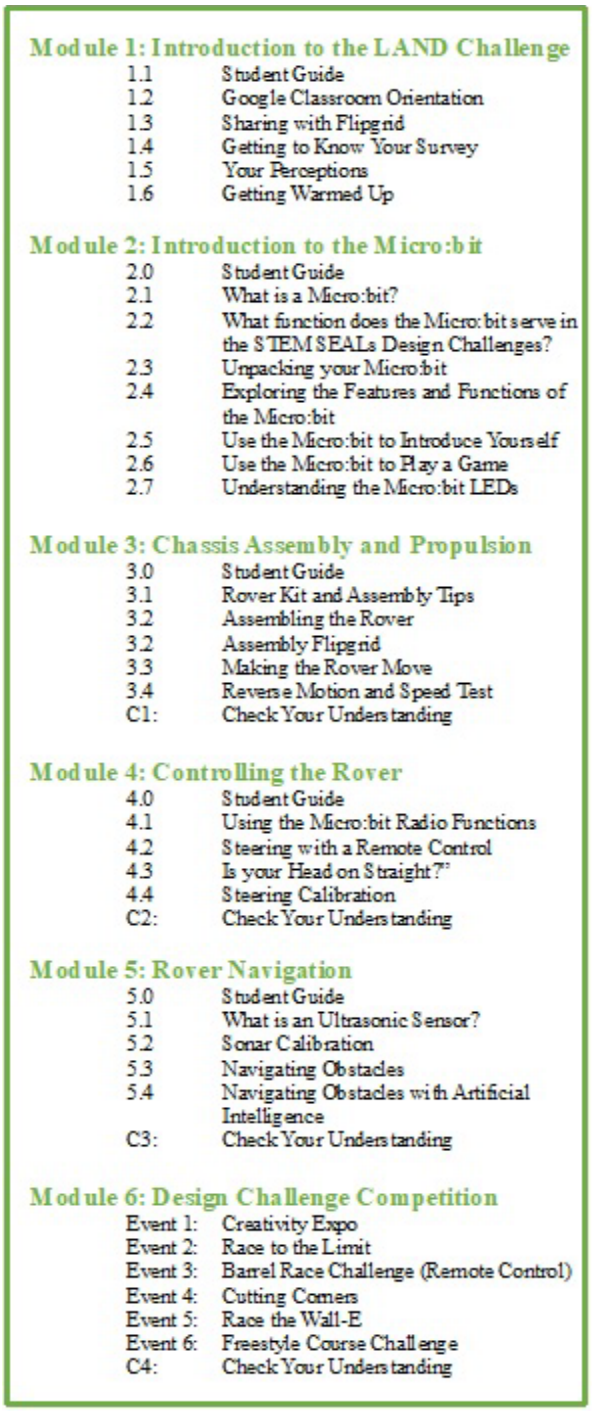
In early 2020, the STEM SEALs team began to finalize the first inaugural enrichment camp where student participants would be immersed in autonomous and remotely controlled robotic devices. The event was scheduled to reside on the NFC campus and would serve up to 48 participants. Then the pandemic struck. The team looked to regroup with the uncertainty of how to implement the camp while also complying with safety provisions that required social distancing. To even attempt, the STEM SEALs team narrowed in on the robotic rovers and decided to focus specifically on the activities that would culminate in a Land Challenge. Figure 3 displays an image of the rover students would assemble, code, and operate. To lessen obstacles in an already challenging time, recruitment for the first camp was restricted to those educators who had been involved with STEM SEALs as a member of the design or review team. For those educators, the student participant pool was assembled. The student pool was restricted to those individuals who resided with the educator or those whom the educator was in regular close contact (e.g., a grandchild).

Figure 3: Image of Assembled Rover

The camp activities were organized into six modules (see Figure 4). These modules included lessons that introduced students to the overarching engineering design challenge which was the focus for the weeklong summer experience. Following the curriculum timeline, next students would be exposed to coding using the Micro:bit and then begin their construction of the rover. Once built, the students would learn how to use the Micro:bit to control and navigate the rover. The week culminated with a competition tied to the engineering design challenge.

STEM SEALs staff mailed to all participants kits that contained pieces of the rover and assembly tools, as well as binders that served as manuals for the camp. The camp officially began on July 8th and lasted until July 17th. A total of 29 teachers and students participated. The camp included online meetings each day. These meetings provided a space for students to receive help from the NFC expert team as desired. Teachers were also available

to support and aid students as best they could while also working through the STEM SEALs learning modules.



Module 1: Introduction to the LAND Challenge

1.1	Student Guide
1.2	Google Classroom Orientation
1.3	Sharing with Flipgrid
1.4	Getting to Know Your Survey
1.5	Your Perceptions
1.6	Getting Warmed Up

Module 2: Introduction to the Micro:bit

2.0	Student Guide
2.1	What is a Micro:bit?
2.2	What function does the Micro:bit serve in the STEM SEALs Design Challenges?
2.3	Unpacking your Micro:bit
2.4	Exploring the Features and Functions of the Micro:bit
2.5	Use the Micro:bit to Introduce Yourself
2.6	Use the Micro:bit to Play a Game
2.7	Understanding the Micro:bit LEDs

Module 3: Chassis Assembly and Propulsion

3.0	Student Guide
3.1	Rover Kit and Assembly Tips
3.2	Assembling the Rover
3.2	Assembly Flipgrid
3.3	Making the Rover Move
3.4	Reverse Motion and Speed Test
C1:	Check Your Understanding

Module 4: Controlling the Rover

4.0	Student Guide
4.1	Using the Micro:bit Radio Functions
4.2	Steering with a Remote Control
4.3	Is your Head on Straight?"
4.4	Steering Calibration
C2:	Check Your Understanding

Module 5: Rover Navigation

5.0	Student Guide
5.1	What is an Ultrasonic Sensor?
5.2	Sonar Calibration
5.3	Navigating Obstacles
5.4	Navigating Obstacles with Artificial Intelligence
C3:	Check Your Understanding

Module 6: Design Challenge Competition

Event 1:	Creativity Expo
Event 2:	Race to the Limit
Event 3:	Barrel Race Challenge (Remote Control)
Event 4:	Cutting Corners
Event 5:	Race the Wall-E
Event 6:	Freestyle Course Challenge
C4:	Check Your Understanding

Figure 4: STEM SEALs Camp Content Overview

Individually, students met at the competition site and had a chance to show off their rovers and compete in the challenge.

Feedback on the camp was offered in multiple ways, mutually supporting the value of the experience on student STEM learning.

For example, students said:

I personally liked learning the coding processes that went into coding the Micro:bit. Learning the code and seeing it work was really satisfying.

I learned a bit more about the electromagnetic scale and got a more in-depth description of how radio waves communicate with each other.

I learned about how even computers use a simulated sense of echolocation to decide how far an object is from it and the patterns it uses to get around the obstacle.

3.1. Findings

The findings section includes a presentation of the iterative design process outcomes along with documentation of the emergent Design-based Research roles and activities at each point in the process. The goal was to document the key decisions at each point in the process and the underlying STEM best practices that were instrumental in the success of the pilot. Ultimately, the findings are in service to a larger question of whether Design-based Research is a good fit for developing high quality STEM experiences, and especially so in situations that require rapid decision-making and significant pivots over time.

Making the tough call

Recognizing that it was now or never. Eventually, the time came when a decision would be needed before the window of opportunity for holding a virtual camp would close. Fortunately, the entire project team recognized that there was no more time to wait or debate. They made the final decision to abandon the idea of an in-person event and instead, move forward with a virtual camp.

Although the pandemic forced US schools to close their doors in late March and early April of 2020, the idea that offering an in-person summer camp opportunity might be in jeopardy was not a consideration, at first. The STEM SEALs team initially thought the county and schools would open in plenty of time to move forward with an in-person summer camp. However, as the pandemic dragged on, the STEM SEALs team began to doubt whether an in-person event would be feasible. When no break in the social distancing restrictions was visible on the horizon the team started to acknowledge that virtual might, in fact, be the only option available. Reluctantly, the team accepted that the inaugural STEM SEALs camp offering would be held as a virtual event.

Early in this process the research team recognized that they were going to be entering into uncharted territory. The traditional approach with well-laid out, pre-determined research activities would need to be paused in favor of strategies that would align to the rapid switch and emerging needs of the new design efforts. As had been the structure of the partnership since the beginning of the grant period, the research team continued to engage in weekly meetings with NFC expert team. It was through these meetings that the new Design-based Research approach began to provide support aligned to the rapidly shifting design efforts. Within these meetings, the research team would identify a welcomed and impactful role as an external sounding board as the NFC expert team collectively began confronting the realization that the social distancing mandates put in place to slow the spread of COVID, might potentially prevent them from hosting an in-person camp as planned. With so much work completed already and so much excitement building to kick off the seminal activities of the project, it was understandable how reluctant the NFC expert team was to accept that the event could not happen as planned.

The research team helped to nudge the STEM SEALs team toward what they recognized as a necessary pivot during the

meetings, sharing about other groups that they were working with who had already made the decision to pivot and providing encouragement that it could be done with STEM SEALs as well. Perhaps more importantly, the research team listened to the thoughts and concerns that were being voiced by the team members and adopted a formal role of providing formative feedback that redirected focus back to areas where issues had been raised, but not yet fully addressed. Ultimately, it was in this role of re-voicing that the research team significantly contributed to the making of a timely decision. The research team recognized that the logistical needs and timeline concerns frequently raised by the project manager needed to be highlighted. This Design-based Research activity was simple, but it proved essential through helping to direct the NFC expert team's focused attention on fast approaching deadlines before they passed. For example, the research team brought up the required timeline needed to successfully engage in recruitment - along with highlighting the long list of activities that would need to occur beforehand.

Expectation setting

Avoiding the "Anything is Better than Nothing" Mindset. Initially, when the STEM SEALs team realized that there was no way that the pandemic conditions would resolve in time to host an in-person event, two competing mindsets emerged: 1) *If we can't do it the way we envisioned, then there is no point in doing it,* and 2) *We need to do something and anything we do is better than nothing.* The team gradually began to embrace the notion that flexibility in the original vision was necessary. However, the team was also firm in not wanting to water down the student experience or alter the main activities that had been so carefully selected. Ultimately, it was the team's deep-rooted commitment to find ways to preserve the foundational elements of STEM SEALs, that propelled them to be able to do what at times seemed impossible, rather than reluctantly shifting toward the second mindset. The commitment to this mindset was an essential component of their persistence and willingness to innovate and adapt to find ways to engage students in a virtual experience that would afford opportunities for them to assemble a rover and to write and run the code that would allow it compete in a real challenge course.

As done previously, the research team engaged with the NFC expert team members weekly with the intention of documenting the process, collecting formative data, creating feedback loops for sharing back findings, and supporting the use of findings to inform continued adaptation and revision. Consistently, but in a much more fast-paced manner, the Design-based Research team had to sacrifice some aspects of rigor to ensure that thoughts were shared in time for decisions to be made. During this time, the research team listened and asked questions, trying to understand more about points of disconnect and indecision and quickly recognizing that the project had yet to engage in the best practices to establish a collective vision. The steps taken to elicit that vision were ultimately very worthwhile, because while many similar efforts pivoted with the "anything is better than nothing" mindset guiding their work, avoiding that mindset was a very important goal of the STEM SEALs team.

The fear that their inaugural effort would be something that lacked the flavor and rigor of what the team had been excited to offer in the original format, was something that permeated their early planning conversations. It was not until they recognized that there were central aspects of the STEM SEALs engagement that would have to be incorporated, or else the team preferred to abandon the idea rather than try to offer something less. Through questions designed to identify the components of the vision that individuals valued and for which there was significant consensus, the research team supported the NFC expert team in identifying the elements that were believed to be fundamental to the project:

1) Providing an opportunity for student participants to construct their own rovers,
2) Teaching them how to program a Micro:bit to use in controlling the rover, and
3) Allowing them to use their acquired skills to complete a challenge course in competition of some form with others in the camp.

Identification of the key challenges

Keeping Students on Track and Maintaining Pace. Quickly the STEM SEALs team noted how the shift from an in-person to virtual delivery would affect student pace and their ability to note slowdowns and intervene. In the face-to-face delivery as originally envisioned, students could work at their own pace, but that pace would be watched by the teachers and the NFC expert team so they might intervene with support when students got stuck or off-task. The teacher facilitators could assist in making sure all students were successfully able to complete the camp's essential elements. But, the virtual environment did not have the same level of oversight. The team wondered how the organization of the camp content could be structured in such a way that all students were able to follow along, stay on track, and proceed at their own pace.

Finding an Effective Substitute for In-Person Demonstrations. In the original, in-person design, the NFC experts planned to provide additional scaffolding for the teacher facilitators and their students through on-demand demonstrations. Thus, the STEM SEALs team had to grapple with the issue of defining and crafting further resources to set students up for success, but without the benefit of on-the-spot, step-by-step, in-person demonstrations. Another related loss was that the in-person demonstrations also afforded students the opportunities to watch other participants and learn through watching and benefiting from other students' efforts to progress through the process. Finding ways for students to receive the instruction needed for them to build their rovers successfully in lieu of the in-person strategies would be a challenge the project would need to overcome.

Providing Students with the Personalized Support Needed to Stay Engaged. The activities and learning tasks associated with the STEM SEALs modules are designed to be novel and challenging. In a face-to-face setting, students would be monitored and supported, but the team recognized that the virtual format potentially left students unprepared and without sufficient support to experience the desired level of success. Without a sufficient level of support, the STEM SEALs team feared the students would not stay engaged or be successful. This facet was complicated because the team recognized that the rigor of STEM exposure that was envisioned provided opportunities for students to grapple and problem-solve and with appropriate support to thrive. Without appropriate support most would likely end up frustrated or, even worse, quit in the face of what they perceived to be their failures. Ensuring sufficient support and scaffolding for students engaging in STEM is challenging in any format, but during a pandemic, through a virtual interface, within a rural setting, and maintaining

social distancing guidelines providing suitable support posed an enormous barrier.

As the team forged forward, much of the early conversations were around concerns about what would not work or what needs students would have that would not be easy to meet in a virtual format. In some ways, these conversations sounded a bit like a listing of the top ideas why there was no way to offer the STEM SEALs vision in a virtual format. Eventually, there would be recognition that all most all their concerns and what ifs emanated from four primary hurdles that the team would need to overcome to have confidence that the vision for the STEM SEALs camp could be carried out successfully in a virtual format.

The research team continued to engage in routine conversations with the group, listening to the discussions about moving forward with an ear toward extracting the core set of actionable barriers that were being alluded to by the team members across the discussions. Often in the team discussions, team members tended to talk about different challenges they could foresee:

- What if some students struggle and are not able to keep up while others may try to jump ahead before they are ready?
- We can't be there to be one-on-one with the students and so how are they going to be supported when they struggle?
- Not everybody might be able to be online at the same time, so what if some of the students need to engage asynchronously?
- How are they going to be able to follow along with a virtual demonstration and either do it at the same time or go back and do it later?

Employing a qualitative lens in analyzing the list of challenges that had been voiced, the research team was able to extract a much smaller number of key challenges to maintaining the foundational elements of STEM SEALs as a virtual offering; ultimately helping distilling concerns to identify the four key challenges listed above.

Identifying promising strategies to address the core challenges

Placing Significantly Greater Importance on Student Curricular Materials. The STEM SEALs team worked through each of the challenges, drawing on research, holding discussions, and providing time for reflection. First, how would the team provide the foundational elements remotely? There would need to be very close attention to the materials. STEM SEALs team members would have to think like a student and what would be available to students. For example, students will not have dual monitors. If students were using the computer to talk with an expert or watch a video, they would need directions, durable directions, to follow and track their progress for the rover assembly. The team used cards with detailed pictures to give students the support they would need for the technical assembly.

In the words of a dyad leader...

Activity cards pictures really helped clarify the steps of each activity and the coding required. Anywhere that more pictures can be added at the various stages of the rover building would help improve understanding of the required activity. The students tended to use the images on activity cards instead of instructions.

Taking Advantage of Existing Virtual Platforms. Ensuring students stayed on track when working independently through online materials seemed like a large issue initially, but ultimately the team noted it was likely a familiar challenge for anyone offering remote learning. Therefore, the team consulted existing platforms for virtual environments. They landed on Google Classroom, which had the advantage of being designed for students of this age and which students may already be familiar. As the team worked to integrate their materials into the Google Classroom platform, they noted a number of helpful features. There were features for presenting content in modules, similar to the design on the original face-to-face camp, and it possessed tools for controlling how far students could progress (e.g., the team set controls for what students could see and click on) and built-in monitoring to track students efforts (e.g., the team could see what students have clicked on and viewed as well as set mini-assignments that showed what had been completed).

Provision of Numerous Videos. Given the critical need for teachers and students to see and observe how to perform certain elements, the team turned first to existing, publicly available videos. Videos would be needed to explain and extend the camp content and could be used to demonstrate the larger context and relevance of the camp by featuring real-world applications. They culled videos from a number of various sources to create a video library for participants. They would also need videos to demonstrate camp-specific activities, so they began shooting videos of themselves performing STEM SEALs activities. These videos would be strategically placed within the STEM SEALs materials so that students could view the relevant videos as they navigated through the STEM SEALs experience.

In the words of dyad leaders …

The resources are very thorough. Dr. Maresch's video on propulsion was very interesting but the manual programming of the pins was a little confusing considering the knowledge wasn't necessary for the programming tasks in this module…The remainder of the video was very helpful, especially the speed and steer demo and explanation.

In fact, the research team received feedback that more and shorter videos would be helpful. In their words,

To keep the student's attention, the video length needs to be kept at a length of 10 minutes or less. For those struggling, a suggested more detailed tutorial video can be uploaded at the end of each module.

More short videos to explain what is happening at each step and what to expect would be great if available to instructors.

Using Dyads. Part of the goal of the STEM SEALs experience was to interest students in STEM careers. To do so, it was important to get and keep students highly engaged throughout the camp. Additionally, students' perception of the camp might be unfavorable if they do not perform well in the culminating Land Challenge. The camp activities were designed to build upon each other, so it was vital that students successfully complete every activity. The STEM SEALs team also recognized that with most of the student participants being middle-school aged, they would need to have a consistent adult to consult with when problems arose and to intervene when their pace slowed. In response to these factors, the team decided to use dyads, that is, each student was paired with an adult, preferably within their household, who could provide that more intimate, just-in-time support and encouragement, and who would help connect them with the experts and other resources needed to be successful.

Expert Office Hours. While the selection of a dyad deliver model addressed many of the issues around student engagement, the teacher facilitators could not be expected to have the technical expertise to solve all the potential challenges that could surface. Experts must be accessible to support the dyads, and available in a timely fashion for students to stay on pace. The team had to think through how to provide the technical expertise and flesh out a schedule of their availability to align with the workflow of students.

In the words of a dyad leader on the strengths of the camp…

The willingness of the team to assist with correcting programming. It takes a village:)

Armed with a clearer understanding of the challenges that threatened their ability to successfully engage students in a virtual STEM SEALs event, the team was positioned to move forward in rapid fashion. The research team leant their hands to the effort, increasing their role and contributions as design team members helping to search out aligned best practices, identifying existing web resources that could be leveraged, trying out PowerPoint for sharing instructional tutorials, and most notably, suggesting one of key strategies that would ultimately prove most critical to the success of the effort: the dyad model. In December of 2019, when the project was happily proceeding as originally proposed, one of the research team leads, engaged in an observational visit to attend the final meeting of the design team that had been convened to support the NFC expert team in developing curriculum materials aligned to middle school standards and students. At this meeting, a research team member spent the day interacting with the design team members and getting to know more about their backgrounds, interests, and motivations for participating. Several of the members were motivated to participate both by professional and personal desires wanting to be able to provide the best opportunities for their students while also wanting to learn more about ways they could position their own children to be more STEM-able. During the downtime, the researcher noted that several were looking at Amazon and placing orders for things that they had been using at the camp. When asked about the purchases, the researcher learned that they were things they were buying because they thought they would be great for their own children. Grounded in these observations, the research team recognized that one of the most significant threats to being able to provide a positive experience in a virtual camp was a lack of sufficient support for students as they grappled with very new and challenging content, would be best addressed by moving to a model that looked to recruit individuals who resembled the design team members. Individuals who possessed the instructional ability to support students and a desire for their children to have greater access to STEM enrichment. This practice of recruiting paired participation of a teacher/educator and a child was the foundation of the dyad approach that was utilized for the virtual offering.

Taking the virtual camp out for a test run

Having Mechanisms for the Expert Team to Troubleshoot would be Essential. After figuring out the strategies and refining the materials, the project team was ready to try out some of the planned camp activities with a group of novices who could provide feedback about what was working and where more development efforts could be directed. The process worked well and provided some good feedback. The initial review, however, was much more influential in preparing the project team for success by making sure they were prepared to address a need that had not previously been fully illuminated, "How could the team troubleshoot in a virtual environment?".

The research team, which typically includes expert review and feedback gathering as a key step in the Design-based Research process, recognized that providing an opportunity for the project team to engage in a test-run and collecting review feedback would go a long way toward strengthening the efficacy of the planned effort. With very little time and significant barriers to getting outsiders materials, sufficient contextual information, and requisite content knowledge, the research team, who did not have prior firsthand exposure to the actual rover materials, signed up as novices to engage in a virtual run through of the rover construction process. Starting with the same unassembled materials that would be provided to participants, the expert instructors led the research team members who were working at remote sites independent from one another in constructing the rovers. The session proved immediately fruitful as one reviewer struggled to easily identify the front of the rover body frame from the back, and as the challenges of trying to screw in the tiny screws were addressed with some additional tips that would be added to the materials before the camp. Gathering feedback of this nature is an explicit reason for engaging in the review and provided the value-add that was anticipated.

However, the ultimate value of the review would quickly surpass the value envisioned with disaster striking and from that disaster an extremely significant innovation would quickly take form. During the test-run review session that was conducted, when one of the two reviewers was testing out recently programmed code that had been transferred to the Micro:bit controlling the rover, the reviewer had the rover placed on their desk in front of them, but as soon as the code sequence was initiated the rover's motors controlling the wheels engaged and the rover quickly drove itself off the desk surface and came crashing to the floor. The rover was rendered inoperable, and panic started to set in for both the expert team and the reviewer. Solidified in this moment would be a need for the expert team and the participants to connect in ways that afforded opportunities to troubleshoot and come up with solutions when things did not work or go as planned. In this case, the panic quickly subsided as the expert instructor who was connecting with the reviewer via Zoom instructed the reviewer to position their camera such that he could survey the damage. After which, the instructor using the white board feature and his own camera position on his rover helped to talk the reviewer through the steps that would be needed to return the rover to working order. Ultimately, the review effort helped to solidify a need and strategy for how expert instructors would be able to connect with participants to provide the one-on-one technical assistance that would be needed to help diagnose problems and scaffold solutions.

Fulfilling the promise of a design challenge event

Finding an answer for, "Where is the Challenge in that?". With the relief that strategies had been incorporated to address all the barriers to success that that had once made a virtual STEM SEALs seem destined to fail, the project leadership did not pause to celebrate, noting that the full STEM SEALs vision centered around the entire experience of learning being grounded in the pursuit of competing with peers in tackling a related design challenge. So, they continued to iterate, recognizing that the virtual format would make it very challenging for students to compete on a similar course and really have the opportunity to apply and test out their

new learning and skills in pursuit of a context embedded design challenge. The design team noted that, to drive the rovers, students would need large, flat, smooth surfaces; something that is much harder to locate in rural areas. The team was greatly concerned that building rovers that could not be successfully driven would leave students disheartened. Ultimately, they believed that learning that does not culminate in a real-world application was below the rigor that they wished to ensure, and they further recognized that completing the camp and competing in isolation would further limit the impact of being part of a cohort of participants who had a shared experience. With that in mind, the team identified a strategy to ensure every student was positioned to be able to successfully engage in the design challenge competition, they arranged for students to compete individually at a parking lot, but live streamed so that students could watch each other and share in the excitement.

Documenting the unintended benefits and lessoned learned

In this step, the focus was on reflective learning toward documenting, 1) emergent benefits that were beyond those originally targeted by the STEM SEALs summer enrichment activity, and 2) emergent best practices that have merit for moving beyond just the current offering and incorporated into future project design and development efforts. As such, this process of taking stock to synthesize learning was one that was initiated and led by the research team which synthesized data from their observation of virtual camp activities, student surveys, dyad teacher surveys and focus groups, and project team reflections. From this work, important outcomes in the form of emergent benefits and best practices were identified and documented.

Increased Versatility. The shift to an online delivery model made the STEM SEALs land experience instantly more versatile and resulted in quicker progress towards leveraging virtual platforms. Access to a STEM SEALs experience remotely would be very beneficial given its goal of appealing to students in rural areas.

Strengthening the K-12 Education Connection. The dyad framework also had unintended advantages. By providing STEM exposure and professional development through the virtual camp, had led to a stronger continued engaged by educators who maintain interested in the STEM SEALs grant and larger mission to build a STEM ecosystem in the area. By connecting directly with teachers and then connecting teachers to each other and the STEM SEALs development work, the project has made swifter progress in creating a core group of educators who can serve as ambassadors as the College looks to strengthen its connection to regional K-12 institutions.

Emergent best practices identified to include in future development efforts were:

Elevating the Framing of Failure. Frustration can intensify quickly in virtual spaces when students are not able to observe others experiencing the same challenges or are spending more time grappling before they are able to receive support or technical assistance. The move to an online offering placed a larger spotlight on addressing failure and led to more proactive work to frame failure as normal and positive. In the words of a dyad leader …

Consider that middle schoolers' need consistency across directions, platforms, and materials. It is hard for them to have materials vary in information when in truth the purpose is for the material to match. Many of them get so frustrated they shut down. They love learning computer applications, doing their assignments online. coding is interesting to them; however, many get very frustrated. They do not understand that frustration is good/ a part of real learning.

New Approach to Differentiation. The STEM SEALs had known that the participants would be a diverse group, regardless of whether the camp was delivered in-person or remotely. This diversity would mean that students would be bringing different sets of pre-existing knowledge and experience and would vary in the kind and amount of support needed. Given the virtual format and much of delivery asynchronously, there was a need to keep instruction shorter, and more condensed. While in-person delivery might have allowed instructors to shepherd students down a more common pathway, the student pathway through the materials would likely be more disparate in the pivot to online. The team saw how alternate pathways that culminate in a singular outcome would need to be crafted. With multiple pathways available, instructors would have the opportunity to steer students toward the appropriate pathway so all students could feel successful progressing in a way that was aligned to their particular prior knowledge and current expertise.

Need to Connect Students to Each Other. The research team had noted the importance of belonging and community in fostering a sense of STEM identity and interest in STEM, leading to student choices to pursue STEM careers, a goal of the project. But, supporting emerging friendships and creating bonds between participants in a virtual environment was an anticipated challenge. While the team had great success in connecting students to facilitators with the dyad support model and in connecting students to experts with the videos and virtual meetings, they had less success connecting students to each other. The team tried to use participant self-introduction videos, videos to showcase participant work, and other strategies to promote team building and student to student connectivity, but these attempts ultimately fell short. New adaptations are currently being considered. In the words of a dyad leader …

With the online program this year, there was not a lot of peer student interaction which many teachers felt would have helped the students to work through the many difficulties they had. Most instructors agreed that the materials provided opportunities for students to express, clarify, justify, interpret, and represent their ideas (i.e., making thinking visible) and to respond to (some limited) peer and teacher feedback.

4. Conclusion

The contributions of this exploratory study of applying the Design-based Research approach within rapid change efforts in STEM enrichment resulted in important contributions to the field in two areas. The first of which is advancing best practice strategies for taking in person hands on STEM enrichment into asynchronous virtual delivery applications. The context of providing high quality STEM enrichment in isolated rural schools areas where connecting students with opportunities and expertise to promote rich engagement in STEM are limited and where creating greater access for rural students is further hampered by transportation and connectivity barriers. Ultimately overcoming these barriers may necessitate expanding opportunities for rural student participation asynchronously in virtual STEM offerings. Research reviews have shown that studies of both synchronous and asynchronous methods of online learning can be effective [29]. The findings of this exploratory study identify emerging best

practices for offering virtual STEM opportunities in a rural setting that are able to meet this high standard have many important implications. The findings lend evidence to support the success of STEM SEALs in offering a highly engaging and successful inaugural summer enrichment experience in an asynchronous and virtual format, and more importantly, provide rich insights to expand the literature on best practices for virtual STEM enrichment programming.

Secondly, the findings of this study are significant because they also provide initial evidence of the viability and value of expanding the knowledge base around best practices for embedding Design-based Research within rapid design efforts. While Design-based researchers have established well-supported best practices and a strong evidence base to promote its routine inclusion in STEM enrichment design efforts, the application of Design-based Research in rapid cycle initial design efforts has not been as well researched [30, 31]. While the pandemic forced widespread pivoting throughout the STEM landscape, the hope is that it will remain an isolated event and that no similar mass need for pivoting will occur. Yet, the rapidly advancing technological will constantly push for rapid development and rapid innovation within STEM enrichment offerings, amplifying the need for greater exploration of Design-based Research methods within rapid cycle design efforts similar to the example provided in this manuscript. The STEM SEALs project efforts to engage Design-based Research to support rapid change efforts represents an important contribution to the field with the findings from the initial test of the offering strongly supporting its effectiveness. The evidence gathered showed Design-based Research as a valuable tool for program improvement, even in the most extreme situation of a global pandemic, and in important contribution to best practices in developing STEM learning experiences.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This material is based upon work supported by the National Science Foundation under grant no. 1812913.

References

[1] R. Zulli Lowe, A. Smith, C. Prout, G.G. Maresch, C. Bacot, L. Murfee, & B. Eustace, "Taking STEM enrichment camps virtual: Strategies & reflections from quick pivot due to COVID-19" In 2021 Integrated STEM Education Conference, 83-90, 2021. DOI: 10.1007/s11423-020-09811-3

[2] World Health Organization, "WHO Coronavirus Disease (COVID-19) Dashboard[Online]," https://COVID19.who.int/?gclid=CjwKCAjwiMj2BRBFEiwAYfTbCod9hYnKI6jtuqIoHEVnStP_VrjFu3NFnfkKJSgh9rZcBuI7mtmfDhoCuYcQAvD_BwE, 2020.

[3] K. Spencer, "Not all towns are created equal, digitally: How a Colorado school district struggles to give its students a technology boost (The Hechinger Report)," Teachers College at Columbia University, 2017.

[4] National Science Board, "Science and engineering indicators 2014," National Science Foundation, 2014.

[5] R. Lapan, M. Aoyagi, M. Kayson, "Helping rural adolescents make successful postsecondary transitions: A longitudinal study," Professional School Counseling, **10**, 266-272, 2007, doi: 10.5330/prsc.10.3.u6j3j64h48p27w25

[6] J. Plucker, B. Harris, "Acceleration and economically vulnerable children," In A nation empowered: Evidence trumps the excuses holding back America's brightest students, Belin- Blank Center for Gifted Education and Talent Development, 181-188, 2015.

[7] S.G. Assouline, L.M. Ihrig, D. Mahatmya, "Closing the excellence gap: Investigation of an expanded talent search model for student selection into an extracurricular STEM program in rural middle schools," Gifted Child Quarterly, **61**(3), 250 –261, 2017.

[8] T. Kittleson, J.T. Morgan, "Schools in balance: Comparing Iowa physics teachers and teaching in large and small schools," Iowa Science Teachers Journal, **39**(1), 8-12, 2012.

[9] J. Plucker, J. Giancola, G. Healey, D. Arndt, C. Wang, "Equal Talents, Unequal Opportunities: A Report Card on State Support for Academically Talented Low-Income Students," Jack Kent Cooke Foundation, 2015.

[10] J.S. Eccles, B.L. Barber, M. Stone, J. Hunt, "Extracurricular activities and adolescent development," Journal of Social Issues, **59**, 865-889, 2003, doi:10.1046/j.00224537.2003.00095.x

[11] P. Olszewski-Kubilius, S.Y. Lee, "The role of participation in in-school and outside-of-school activities in the talent development of gifted students," Journal of Secondary Gifted Education, **15**(3), 107-123, 2004.

[12] R. Gira, "The challenge: Preparing promising low-income students for college," In Overlooked gems: A national perspective on low-income promising learners: Conference proceedings from the national leadership conference on low-income promising learners, National Association for Gifted Children, 69-74, 2007.

[13] T. Anderson, J. Shattuck, "Design-based research: A decade of progress in education research?" Educational Researcher, **41**(1), 16–25, 2012, doi:https://doi.org/10.3102/0013189X11428813

[14] T. Štemberger, M. Cencič, "Design-based research in an educational research context," Journal of Contemporary Educational Studies, **65**, 90-104, 2014.

[15] E. Oh, T.C. Reeves, "The implications of the differences between design research and instructional systems design for educational technology researchers and practitioners," Educational Media International, **47**(4), 263-275, 2010.

[16] D. Zinger, A. Naranjo, I. Amador, N. Gilbertson, M. Warschauer, "A design-based research approach to improving professional development and teacher knowledge: The case of the Smithsonian learning lab." Contemporary Issues in Technology and Teacher Education, **17**(3), 388-410, 2017.

[17] F. Wang, M.J. Hannafin, "Design-based research and technology-enhanced learning environments," Educational technology research and development, **53**(4), 5-23, 2005.

[18] T. Amiel, T. Reeves. "Design-based research and educational technology: Rethinking technology and the research agenda," Educational Technology & Society, **11**, 29-40, 2008.

[19] A. Bakker, "An introduction to design-based research with an example from statistics education," Approaches to qualitative research in mathematics education: Examples of methodology and methods, 429-466, 2015, doi:10.1007/978-94-017-9181-6_16, 2014

[20] S. Barab, K. Squire, "Design-based research: Putting a stake in the ground," Journal of the Learning Sciences, **13**, 1-14, 2004, doi:10.1207/s15327809jls1301_1.

[21] C. Pardo-Ballester, J.C. Rodríguez, "Using design-based research to guide the development of online instructional materials." Developing and evaluating language learning materials 86012," 2009.

[22] F.E. Terrazas-Arellanes, L.A. Strycker, E.D. Walden, C. Knox, "Development of a middle school online science curriculum: Lessons learned from a design-based research project." In Handbook of Research on Diverse Teaching Strategies for the Technology-Rich Classroom, ICI Global, 2020, doi: 10.4018/978-0238-9.

[23] D.M. Bressler, M. Shane Tutwiler, A.M. Bodzin, "Promoting student flow and interest in a science learning game: a design-based research study of school scene investigators," Education Technology Research Development, **69**, 2789–2811, 2021, doi:10.1007/s11423-021-10039-y

[24] A. Savran Gencer, H. Doğan, "The assessment of the fifth-grade students' science critical thinking skills through design-based STEM education," International Journal of Assessment Tools in Education, **7**(4), 690-714, 2020, doi: 10.21449/ijate.744640

[25] S. Fowler, S.N. Leonard, "Using design based research to shift perspectives: a model for sustainable professional development for the innovative use of digital tools," Professional Development in Education, 1-13, 2021, doi: 10.1080/19415257.2021.1955732

[26] M. Allen. Leaving ADDIE for SAM: An agile model for developing the best learning experiences, ASTD Press, 2012.

[27] M.W. Easterday, D.R. Lewis, E.M. Gerber, "Design-based research process: Problems, phases, and applications," International Society of the Learning Sciences, 2014.

[28] C. Jackson, "The advantages of exploratory research design," eHow, https://www.ehow.co.uk/info_8525088_advantages-exploratory-research-design.html, 2020.

[29] F. Amiti, "Synchronous and asynchronous e-learning," European Journal of

Open Education and E-Learning Studies, **5**(2), 2020.
[30] C. Hoadley, F.C. Campos, "Design-based research: What it is and why it matters to studying online learning," Educational Psychologist, **57**(3), 207-220, 2022.
[31] T.C. Reeves, L. Lin, "The research we have is not the research we need," Education Technology Research Development, **68**, 1991–200l, 2020, doi: DOI: 10.1007/s11423-020-09811-3