

BY ANHALT UNIVERSITY OF APPLIED SCIENCES

PROCEEDINGS OF THE 13TH INTERNATIONAL CONFERENCE ON APPLIED INNOVATIONS IN IT

VOLUME 13
ISSUE 3

EDITION

Hochschule Anhalt

Proceedings of the 13th International Conference on Applied Innovations in IT

Volume 13 | Issue 3

Koethen, Germany
26 July 2025

EDITORIAL BOARD

Editor-in-Chief

Prof. Dr. Eduard Siemens (Anhalt University of Applied Sciences), Germany

Topical Editors

Prof. Dr. Marija Kalendar (Ss. Cyrill and Methodius University in Scopje), North Macedonia

Prof. Dr. Halina Falfushynska (ENERTRAG SE), Germany

Prof. Dr. Hassan Bevrani (University of Kurdistan), Iran

Prof. Dr. Ralph Gottschalg (Fraunhofer IMWS), Germany

Prof. Dr. Iryna Stankevych (Odessa State Academy of Civil Engineering and Architecture), Ukraine

Prof. Dr. Douglas Hall (St. Mary's University), USA

Technical Editor

Kateryna Shulakova (State University of Intelligent Technologies and Telecommunications, HSA)

The Conference Proceedings Journal, volume 13(3), contains selected papers accepted for publication from those presented at the 10th International Conference: Education and Sustainable Development: Technological Advances and AI Impacts (ESDTAI-25), held on May 29–30, 2025, in Istanbul, Turkey. This issue presents research in information technology, artificial intelligence, digital transformation, communication systems, and sustainable development. The papers included in this volume are of particular interest to researchers and professionals exploring innovative technological solutions for education and global sustainability.

ISBN: 978-3-96057-186-5 (Online)

ISSN: 2199-8876

All papers published in this journal are full-open access:
immediately freely available to read and download.



Publisher: Edition Hochschule Anhalt
Phone: +49 (0) 3496 67 5611
Email: leiterin.hsb@hs-anhalt.de



Phone: +49 (0) 3496 67 2327
Email: conf@icaiit.org
<http://icaiit.org>



Köthen, Germany
Anhalt University of Applied Sciences
Bernburg / Koethen / Dessau
<http://www.hs-anhalt.de>

Postfach 1458 D-06354 Koethen, Germany

ABOUT THIS SPECIAL ISSUE

The Proceedings of the International Conference on Applied Innovations in IT (ICAIIIT 2025), Volume 13, Issue 3, features selected articles from researchers worldwide, highlighting the latest advancements in information technology, artificial intelligence, and sustainable development. This issue presents cutting-edge studies on communication technologies, cybersecurity, blockchain innovations, data processing, automation, and power engineering, reflecting the ongoing progress in these scientific fields.

About ESDTAI-25

The 10th International Conference: Education and Sustainable Development – Technological Advances and AI Impacts (ESDTAI-25) (<https://surl.li/fkkqdd>) was a distinguished international event that brought together scholars, researchers, and industry professionals from around the world. Held on May 29–30, 2025, in Istanbul, Turkey, the conference was organized by University Muhammadiyah Bangka Belitung (Indonesia) in collaboration with Scholar Worldwide (SCHWLAR).

The conference served as a dynamic platform for exchanging ideas, fostering academic collaboration, and presenting cutting-edge research on the intersection of artificial intelligence, education, and sustainable development. Participants engaged in expert discussions, workshops, and scientific panels that explored how emerging technologies can drive innovation and global progress.

This special issue features selected papers presented at ESDTAI-25, focusing on the following key research areas:

- 1) Communication Technologies and Hardware
- 2) Data Analysis and Processing
- 3) Data Transport Technologies
- 4) Control and Automation
- 5) Statistical Modelling and Analysis and Their Applications in Life Sciences
- 6) Information Technology in Economics

By integrating these diverse themes, the conference highlighted the essential role of technology and artificial intelligence in shaping the future of sustainable education, innovation, and scientific advancement.

Content

Section 1 Communication and Data Transport Technologies

<i>Wisal Jereis Alrabadi</i> Neural Network-Based Intelligent Routing for Secure VANET Communication.....	1
<i>Arif Abdul Sayel and Khalid Amer Abdul Razzaq</i> Performance Comparison of Energy-Efficient Routing Techniques in Wireless Sensor Networks.....	11
<i>Wayan Puguh, Haidir Amin and Arkan Adnan Imran</i> A Robust Wireless Sensor Network for Real-Time Soil Moisture Analysis in Precision Farming.....	19
<i>Wisal Jereis Alrabadi</i> Energy-Efficient Internet of Things-Based Wireless Sensor Networks through AI-Powered Resource Allocation.....	27
<i>Hema Muniyappan, Pavithra Sankarappan, La Ode Muhram, Muhamad Farih, Aarthi Saravanaperumal, Shubham Sharma and Arkan Adnan Imran</i> Advancements and Cybersecurity Strategies in V2X Communication and Transport System for Electric and Connected Vehicles Using IoT and AI.....	35
<i>Bollepogu Venkateswarlu, Muhamad Karnoha Amir, Raivel Raivel, Gudipati Murali, Rayala Sateesh and Sameer Hashim Raheem Al Deen</i> Low-Power Redundant-Transition-Free TSPC Dual-Edge-Triggered Flip-Flop with Single-Transistor Clocked Buffer.....	43
<i>Ahmad Al Yakin, Luis Cardoso, Ali Said Al Matar, Yasir Ahmed Mohammed Ridha Alsbbagh, Muthmainnah Muthmainnah, Ahmed J. Obaid and Mohammed Wajeesh Hussein</i> Rethinking Student-Robot Social Interactions in GenAI Capitalism and its Role in Shaping Cybersocialization through a Sociological Lens.....	55
<i>Ali Bnilam, Murtadha Ahmed Jawad and Hasan Abed</i> Privacy-Aware Machine Learning Techniques for Secure Internet of Things Systems.....	65

<i>Sura Abdulkareem Abbas, Hassan Jaleel Hassan and Ghaidaa M. Abdulsahab</i> Enhancing Healthcare Data Protection for Modern Digital Health Systems.....	73

<i>Ruksanan Ruksanan, Erni Danggi, Ebaa Abdulsattar Jaber and Arkan Adnan Imran</i> A Novel Encryption and Data Mining-Based Approach to Secure Sustainable Cloud Systems.....	81

<i>Irwan Lakawa, Hujiyanto Hujiyanto and Abbas Oudah Waheed Alomairi</i> Machine Learning Hyperparameters Optimization for Accurate Arabic Sentiment Classification.....	89

Section 2 Data Transport Technologies

<i>Abdul Nashar, Joko Tri Brata and Arkan Adnan Imran</i> Design and Development of an IoT-Blockchain Platform for Reliable Sensing Data Management.....	97

<i>La Ode Bariun, Haeruddin Haeruddin and Mahmoud A. Mahmoud</i> Blockchain-Enabled Data Analytics-Based Demand Service Optimization for Vehicular Safety.....	105

<i>Andi Bahrin, La Oge, Ahmed Zeyad Almhanna and Mohammed Abdul Jaleel Maktoof</i> Secure and Privacy-Enhanced Data Aggregation for IoT Using Fog Computing.....	115

<i>Hijriani Hijriani, Muhamad Yusuf, Reem Abduljaleel Khaleel and Banan Yousif Ahmed</i> A Novel Secure and Decentralized Model for Authentication and Data Integrity in IoT-Based Cyber-Physical Healthcare System.....	123

<i>Amna Arak Radeini Asal and Shaimaa Jabbr Ali</i> Privacy-Preserving Machine Learning Using Consortium Blockchain in Vehicular Social Networks.....	133

<i>Radhiya Sulaiman Nasser Alhabsi and Alla Salim Mohammed Almukhaini</i> Blockchain-Driven Security, Privacy and Reliability for Digital Healthcare Systems.....	141

Section 3 Control and Automation

<i>Basima Kzar Hassan, Khawla Rashige Hassen and Sakna Jahiya Faraj</i> An Efficient Intrusion Detection System for IoT Using Feature Engineering and Machine Learning.....	149
<i>Mustafa Hussein Zwayyer, Sajida Allawi Dawood and Ammar Bassem Saleh</i> Metaheuristic Optimization Algorithms for Deep Learning Model Design in Secure Internet of Things Environment.....	157

Section 4 Data Analysis and Processing

<i>Ashwini Lakshmipathy, Muthupandi Gurusamy, Siti Fatmawati Lalo, Nonia Sakka Lebang, Karthikeyan Selvaraj and Aws A. Abdulsahib</i> Advanced Network Security System: Honeypot-Based Intrusion Detection with Machine Learning and Visualization.....	165
<i>Shatha Hamead Othman and Huda Abdulaali Abdulbaqi</i> A Hybrid Static-Dynamic Analysis Model for Android Malware Detection: Design, Implementation and Comparative Assessment.....	173
<i>Sahlah Abd Ali and Ielaf Osamah Abdulmajeed</i> Privacy-Driven Webpage Fingerprinting Using Encrypted Traffic Packet Lengths.....	183
<i>Sufrianto Sufrianto, La Harudin, Abbas Oudah Waheed Alomairi and Mohammed Abdul Jaleel Maktoof</i> A Graph Neural Network Approach for Identifying Decentralized Applications in Encrypted Traffic.....	191
<i>Wasan Khairallah Ali</i> Enhancing Machine Learning Model Accuracy with Effective Data Scaling Strategies.....	199
<i>Shanmuga Shyam B., Selva Jothi Murugan, Chairan Zibar L. Parisu, Sasmin Sasmin, Janani Vignesh and Ahmed Mohammed Abdukareem</i> Revolutionizing Surveillance with Deep Learning-Powered Anomaly Detection and Real-Time Behavior Analysis.....	207

<i>Vedavyas Gurla, Sarlan Sarlan, Ilham Ilham, Aaluri Seenu, Vasavi Oleti, Danampalli Sravanthi and Hassnien S. Al Hashemi</i> Qualified Investigation of ML Models for Forecasting SPEC Benchmark Performance.....	215
<i>Ali Abdul Karim Jawad and Asaad Abbas Hindi</i> Intelligent Skin Cancer Detection from Dermoscopic Images with Machine and Deep Learning Approaches.....	225
<i>Anidi Anidi, Arna Juwariyah, Ahmed Aljuboori and Shahad Lateef Abdulwahid</i> Intelligent IoT-Based Data Analytics System for Precision Farming Using Regression Techniques.....	233
<i>Hussein Qahtan Al Gburi, Abeer Tariq Maolood and Akbas Ezaldeen Ali</i> AI-Enhanced Intelligent Healthcare: Advancements in Remote Monitoring, Predictive Analytics and Disease Diagnosis.....	243
 Section 5 Statistical Modelling and Analysis and their Applications in Life Sciences	
<i>Sura Ibrahim Mohammed Ali and Sura Zaki Al Rashid</i> Integrated Features Based on Graph Clustering and Gene Expression.....	253
<i>Athraa Abdul Ghani and Mohammed Shakir Mahdi</i> Using the Rooted Tree to Find the Optimal Solution for the Transportation Problems.....	263
<i>Laila Suwaid Said</i> Cryptocurrency Price Prediction Model Using GRU, LSTM and Bi-LSTM Machine Learning Algorithms.....	271
<i>Abdullah AbdulRazzaq Obaid and Haneen Kareem Hussein</i> Chromaticity of Normal Graphs and Cyclic Graphs.....	279
<i>Nandhakumar Dhatchinamoorthy, Vickky Anggara Ilham, Farra Sasmita and Reem Abduljaleel Khaleel</i> Data-Driven Crop, Fertilizer and Analytics Guidance through Machine Learning.....	287

Section 6 Information Technology in Economics

<i>La Ode Abdul Manan, La Panga and Mahmoud A. Mahmoud</i> The Role of Blockchain, Big Data and Cloud Computing in Modern ERP Implementation.....	297
<i>Nagham Ja'far Hussein and Abdulhussein Mansoor Ali</i> Machine Learning-Driven Predictive Analytics for Real-Time Supply Chain Risk Management.....	305
<i>Poojashri Pragalathan, Pavithra Sankarappan, Ridha Taurisma Lajaria, Nur Aisyah, Pavithra Sankarappan, Sujitha Vijayalekshmi Sadhasivan Nair, Shubham Sharma and Nazar Jabbar Alhayani</i> Advancing Intelligent Automation: Integrating Robotic Process Automation and Artificial Intelligence to Streamline Business Operations and Enhance Audit Processes.....	313
<i>Meeras Salman Al Shemarry</i> Machine Learning-Based Predictive Analytics for Sustainable Renewable Energy Investments.....	323
<i>Esraa Tayyeh Ehaimed and Mushtaq Abd Al Ghany</i> Using the Program Evaluation and Review Technique and the Critical Path Method to Help Project Managers.....	331
<i>Sabyasachi Pramanik, Abdulkhaleq Husham Yousif, Salah Jasim, Raushan Raj, Muskan Kumari, Atanu Roy and Ahmed J. Obaid</i> House Price Prediction Using Diverse Machine Learning Techniques.....	337

SECTION 1

COMMUNICATION AND DATA TRANSPORT TECHNOLOGIES

Proceedings of the 13th International Conference on Applied Innovations in IT (ICAIIIT)

JULY 2025

Neural Network-Based Intelligent Routing for Secure VANET Communication

Wisal Jereis Alrabadi

*Faculty of Physical Education, Yarmouk University, 21163 Irbid, Jordan
wrabadi10@gmail.com*

Keywords: Vehicular Ad-Hoc Networks (VANETs), Neural Network-Based Routing, Intelligent Transportation Systems (ITS), Machine Learning, Network Security and Attacks.

Abstract: Transportation Systems (ITS) enable seamless communication between vehicles and roadside infrastructure. This connectivity significantly enhances road safety, traffic efficiency, and overall driving enjoyment for users. However, router protocols in VANETs encounter substantial challenges due to the high mobility of vehicles and the rapid changes in network topologies. Traditional routing methods often suffer from delays and packet loss as a result of these dynamic conditions. To address these issues, we propose a novel algorithm that leverages machine learning techniques, specifically utilizing neural networks for intelligent routing in VANETs. This innovative approach dynamically optimizes routing decisions while also enhancing communication security. By effectively detecting and mitigating potential attacks, our algorithm improves routing efficiency, reduces communication delays, and strengthens data security. Simulation results indicate that our proposed system outperforms existing routing protocols, leading to improved network performance and a significant reduction in end-to-end delay, particularly in challenging scenarios such as black hole attacks.

1 INTRODUCTION

When vehicular ad-hoc networks (VANETs) are used in intelligent transportation systems (ITS), traffic can be managed better, and drivers are more comfortable by being able to communicate with one another and with roadside infrastructure. In an increasingly complex and scaled network, ensuring secure and efficient communication between vehicles becomes a more challenging task [1]. The key challenge facing VANETs is ensuring robust, adaptive, and secure routing mechanisms that can cope with the changing topologies, mobility, and communication conditions in vehicular environments.

A promising approach to addressing these challenges is to use neural networks-based intelligent routing. Neural networks can be used to optimize data transmission efficiency and security by using artificial intelligence and machine learning. In addition to improving the accuracy of route selection, minimizing communication delays, and enhancing the overall security of information exchange, these intelligent routing techniques can detect and mitigate potential threats in real time. Using a vehicle ad hoc network

(VANET), vehicles can communicate with roadside units (RSUs) [2]. In Vehicle Area Networks (VANETs), traffic congestion is reduced, road safety is improved, and traffic efficiency is increased, as shown in Figure 1 [3], [4]. Vehicle-to-vehicle networks provide data communication, either with or without fixed infrastructure [5]. With this communication, a variety of safety and infotainment applications can be integrated into the vehicle to enhance the overall driving experience and contribute to smarter transportation systems [6], [7]. Because VANETs are wireless and mobile, they are highly susceptible to security threats [8]. There is the potential for catastrophic results from malicious activities, especially when it comes to the propagation of emergency messages and the management of traffic [9]. To ensure the success and wide adoption of VANETs, security must be addressed at the highest level [10]. It is vital to ensure the security of data communication and that messages are trusted and reliable in VANETs in order to operate efficiently and safely [11], [12]. A secure and intelligent routing scheme is crucial to addressing these challenges [4]. Traditionally designed routing protocols for MANETs often do not work well in VANETs due to

their high mobility and dynamic topology [13]. Smart systems need to be designed for VANET routing protocols, and the network environment is rapidly changing. Due to malicious nodes and potential attacks, innovative routing algorithms are required to ensure efficient and secure data transmission [14].

A neural network and machine learning technique can be used to improve routing in VANETs. In addition to optimizing network performance and mitigating security threats, NNs can enable smart routing decisions based on data-driven predictions and adaptive learning. With neural network-based routing protocols, optimal paths can be predicted, anomalies can be detected, and VANET communications can be enhanced in terms of overall security [15], [16].

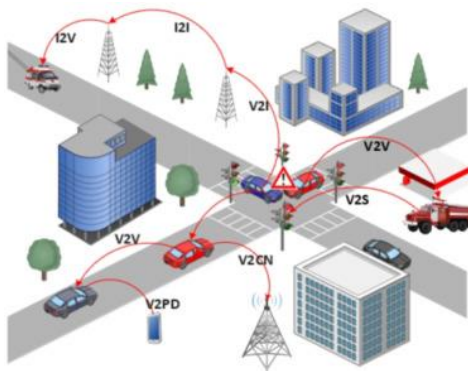


Figure 1: A VANET is composed of the following elements of communication.

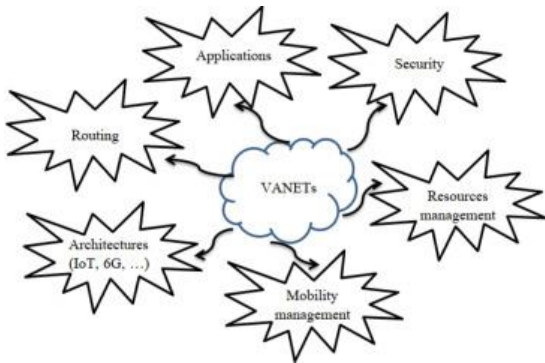


Figure 2: Topics for VANET research and issues.

A variety of VANET features, from entertainment features to driving assistance systems, have significantly improved automobile capabilities [17]. While these services are beneficial, they also present new challenges in terms of quality, security, and privacy. Since the 1990s, VANET research has encompassed a variety of topics, including application development, routing, security, and

privacy. The challenges remain despite technological advances, especially with the advent of IoT, cloud computing, fog computing, edge computing, and 5G/6G technologies, as shown in Figure 2 [18], [19].

2 LITERATURE REVIEW

VANET is an integral part of Intelligent Transportation Systems (ITS), which improve traffic efficiency, safety, and convenience for drivers. As a result, routing protocols struggle to keep up with the rapid changes in topologies and highly mobile vehicles of VANETs. This dynamic environment often results in delays, packet losses, and security vulnerabilities with traditional routing protocols. As VANET communications have become more secure, machine learning (ML) techniques, especially neural networks (NNs), have become increasingly important for developing smart and adaptive routing solutions [20].

2.1 Intelligent Routing Protocols in VANETs Machine Learning-Based Routing

VANET routing protocols have benefited from machine learning (ML) due to their ability to learn network behaviour on its own and in an adaptive manner [21]. This learning-based approach can address many challenges in VANETs, including mobility and dynamic topologies [22]. The use of machine learning algorithms can improve routing protocols' performance and security by making intelligent decisions based on real-time network conditions. Optimization of routing decisions can be achieved using reinforcement learning (RL). Using deep reinforcement learning, an author proposed a roadside unit (RSU) model that can maintain traffic information and predict vehicle movements to find feasible routes [23].

Further studies investigate using Q-learning and fuzzy logic to route VANETs hierarchically. The QFHR protocol consists of three phases: identifying traffic conditions, routing at intersections and routing at roads. In comparison with other routing protocols, QFHR improved packet delivery rate and reduced delay.

2.2 ANN-Based Routing

It has become increasingly evident that Artificial Neural Networks (ANNs) are powerful tools for

predicting routing in VANETs, enabling protocols to adapt to changing network characteristics. In accordance with the author [3], MPANN is a Multimetric predictive routing protocol for VANETs in urban environments. With the protocol, the best route to the destination is predicted by a neural network, increasing delivery probability and reducing delivery times. According to MPANN simulations, losses at different densities of vehicles are less than 20%, while delays are less than 0.4 ms.

To combine clustering algorithms with artificial neural networks, the author has developed an enhanced routing protocol [24] capable of detecting malicious nodes and establishing clusters. Through a modified version of the ad hoc on-demand distance vector protocol (AODV), it is possible to detect black hole attacks 98.97% accurately using an ANN classification model.

In this study, a neural network is used to provide a dynamic routing and switching scheme (NARSS) for Software-Defined Vehicular Networks (SDVNs) using artificial neural networks [25]. In simulations, NARSS is shown to outperform traditional protocols in terms of packet delivery ratios and end-to-end delays based on a neural network generating a scheme-switching model.

2.3 Ad Hoc On-Demand Distance Vector (AODV) Routing Protocol

The AODV protocol is used in VANETs for reactive routing. In his study of AODV in VANETs, the author [26] used SUMO, MOVE, and NS2 simulators to examine how it was implemented in those networks. AODV uses request-response mechanisms to establish routes only when they are needed. Based on a comparison of AODV's results with those of other studies, the authors evaluate packet drop rate, throughput, average end delay, jitter, and routing workload. AODV protocol was modified and incorporated into the intelligent cluster-based routing protocol. Modified AODV is used to discover routes on demand, select routes optimally and reduce packet transmission delays through enhanced request and reply protocols [27]. Other approaches, such as multicast routing protocols designed for delay-tolerant vehicular environments, have also been explored to enhance data dissemination [34/28].

3 METHODOLOGY

In VANETs, nodes move, and their topology often changes, which makes them more difficult to

distinguish from Mobile Ad Hoc Networks (MANETs) [28/29]. Moreover, VANETs are independently developed and have a short lifespan. Mobile nodes (cars) primarily use VANETs, but roadside units (RSUs) also operate on VANETs [29/30]. It is possible for nodes that want to interact with another node that is not immediately within range to do so by forwarding their target to neighbouring nodes. There are a number of routing protocols to choose from, but we focus on the AODV protocol, which is among the most common and important protocols used today. VANETs and other IoT devices are becoming increasingly vulnerable to security breaches as intra-vehicular communication becomes increasingly important to applications [31], [32], [33]. Using our solution to simulate attacks and demonstrate their effectiveness can help us improve prevention methods. By analyzing these simulations, we show that our countermeasure detects and prevents Black Hole attacks accurately and effectively. As a result of our approach, network performance was boosted proportionally in instances of Black Hole attacks.

3.1 Data Rescaling Techniques

Due to the different sizes of the input variables, rescaling the data is essential before applying the statistical methods cited above. Analyzing data this way might result in biased results. It is thus necessary to transform or rescale input data in a way that no single attribute dominates another [34]. This can be accomplished by transforming the original data into a standard range, such as $[-1, 1]$ or $[0, 1]$.

The original data must be transformed so that it falls within a smaller or standard range.

Using $V_1, V_2, \dots, V_n$ as an example V , we have n observations for this numeric variable

3.2 Z-Score Normalization

As a result, the data are rescaled to yield zero mean and unit variance [34]. To transform [27] into v'_i we follow the steps below:

$$v'_i = \frac{v_i - \mu}{\sigma}. \quad (1)$$

Among the variables, $V, i = 1, \dots, n$, μ , and σ reflect the means and standard deviations of the original values.

3.3 Min-Max Normalization

To map each value v'_i of v_i to a value within the series $[0, 1]$, this new value is computed as tracks:

$$v'_i = \frac{v_i - \min V}{\max V - \min V} \cdot \quad (2)$$

In this case, v'_i reflects its original value, $V, i = 1, \dots, n$, $\min V$ and $\max V$ is its maximum value.

A variable's minimum value becomes 0, and its maximum value becomes 1. If $= \min V$; then $v'_i = 0$. If $= \max V$; then $v'_i = 1$.

3.4 Normalization by Decimal Scaling

In this method, variable maximum values are determined by moving their decimal points. When variables have logarithmic variations, the approach is suitable [34].

Using the given data, instances v_i are rescaled into v'_i . As follows:

$$v'_i = \frac{v_i}{10^j} \cdot \quad (3)$$

Anywhere: $j = \log_{10} \max(v_i)$. A comparison of three different techniques was conducted using original research data [34].

3.5 Network Simulator 2 (NS-2)

To evaluate the performance of the proposed neural network-based routing scheme, we conducted experiments using Network Simulator 2 (NS-2). NS-2 provides a discrete-event simulation environment widely used for testing ad-hoc and vehicular networks due to its flexibility and support for wireless protocols.

In our setup, the mobility of vehicles was generated using traffic simulation tools (SUMO/MOVE), which were then integrated into NS-2. The simulation included vehicles (mobile nodes) and roadside units (RSUs), representing a realistic VANET environment. Each vehicle was equipped with wireless communication capability, and nodes were allowed to dynamically join or leave the network depending on their movement patterns.

The proposed routing model was implemented as an extension of AODV, where neural network-based decision-making was applied to detect malicious behavior (e.g., black hole attacks) and to select optimal routes. During simulation, packet headers and routing tables were updated dynamically to reflect real-time changes in topology.

For performance analysis, we measured:

- Packet delivery ratio (PDR) – the proportion of successfully delivered packets.
- End-to-end delay – the average time taken for a packet to reach its destination.

- Throughput – the overall amount of data successfully transmitted across the network.
- Energy consumption – estimated for both vehicle nodes and RSUs during communication.
- Attack detection accuracy – evaluated using confusion matrix-based metrics (Accuracy, Precision, Recall, F1-score).

These metrics allowed us to assess not only the efficiency of the routing protocol but also its ability to resist malicious nodes. By comparing results with baseline protocols (AODV, DSR, and trust-based schemes), we demonstrated that the proposed solution achieves higher delivery ratios, lower delays, and better resilience under attack scenarios.

3.6 Performance Evaluation

To evaluate the proposed neural network-based routing scheme in VANETs, we used both network-level metrics (alive and dead nodes, energy efficiency) and machine learning-based classification metrics (accuracy, precision, recall, F1 score).

3.6.1 Number of Dead Nodes

Since VANET nodes are mobile vehicles equipped with communication modules, their availability directly influences routing performance. In our simulations, we observed the number of alive nodes (vehicles that remained active and able to forward packets) and dead nodes (vehicles that could no longer participate due to depleted communication resources or disconnection). A higher number of alive nodes indicates better network sustainability and routing efficiency.

3.6.2 Energy Consumption

Even though vehicles have more energy resources than traditional sensor nodes, efficient energy usage is still important for VANET devices, especially roadside units (RSUs) and relay vehicles. We modeled energy dissipation during data transmission and reception. The total energy consumed by a vehicle node $E_{total}(i)$ is represented as the sum of residual energy and the energy harvested or supplied:

$$E_{total}(i) = E_{res}(i) + E_{harvest}(i). \quad (4)$$

Here, $E_{res}(i)$ - represents the remaining energy of the node after communication activities, while $E_{harvest}(i)$ - accounts for energy replenishment (for example, from vehicle batteries or hybrid systems).

The required transmitter power was estimated as:

$$TP = SNR/\alpha,$$

where SNR is the signal-to-noise ratio and α is the channel attenuation factor.

3.7 Model Performance Evaluation

Model performance was comprehensively evaluated using classification metrics derived from the confusion matrix framework. The confusion matrix is a tabular representation that summarizes the predictions of a classification model, making it easy to identify misclassified instances and calculate performance measures. For a classification task with multiple classes, four fundamental outcomes are recorded for each class:

- true positive (TP) represents correctly classified positive instances,
- false positive (FP) represents negative instances incorrectly classified as positive,
- false negative (FN) represents positive instances incorrectly classified as negative,
- true negative (TN) represents correctly classified negative instances.

Based on these four classification outcomes, four key performance metrics were calculated according to the formulas presented in [35].

Accuracy reflects the overall correctness of the model's predictions across all instances, providing a general measure of classification performance. Precision indicates the relationship between correctly predicted positive instances and all instances flagged as positive by the model, measuring the reliability of positive predictions. Recall, also known as the true positive rate, measures the proportion of actual positive instances that were successfully identified by the model, indicating the model's ability to capture all positive cases. The F1-score represents a harmonic mean of precision and recall, providing a balanced single metric that accounts for both false positives and false negatives, making it particularly useful when the class distribution is imbalanced.

For the attack detection task in this study, accuracy reflects the overall correctness of intrusion detection. Precision indicates how many nodes flagged as malicious were truly malicious, reducing false alarms. Recall measures the proportion of actual malicious nodes that were successfully detected by the model. The F1-score balances precision and recall, providing a reliable single metric for evaluating overall attack detection performance.

4 RESULTS AND DISCUSSION

With the help of the BHT dataset, a comprehensive set of outcomes was developed to provide analytical and predictive capabilities for evaluating the proposed system's performance. A significant improvement in the system's predictive capabilities has resulted from our training efforts, enabling us to compare its performance across different datasets in a meaningful way. A clear illustration of the model's accuracy can be found in Figure 3.

Using the proposed BHT dataset, Figure 4 shows the model's loss. A graph showing the system's loss across the specified dataset is shown below. Based on these metrics, iterative training and testing can be conducted on the proposed model. Figure 4 illustrates the model's loss graphically.

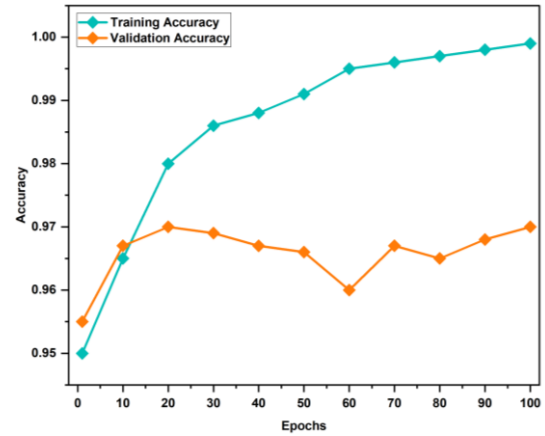


Figure 3: A model for BHT with high accuracy.

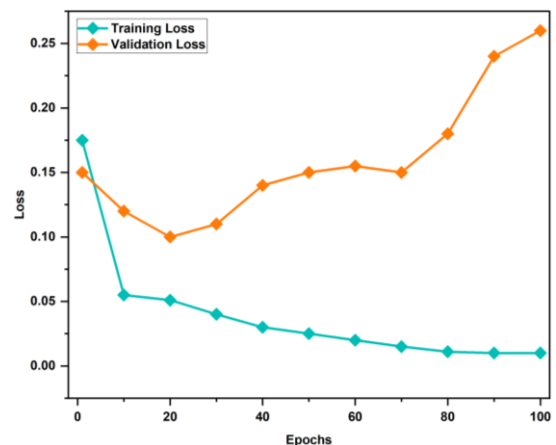


Figure 4: An analysis of the BHT dataset's model loss.

According to Figure 5, existing techniques are compared to the minimum number of under-attack scenarios that can occur. Compared to existing techniques, these occurrences occur less frequently, suggesting less impact on network routing performance. Data transfer efficiency in a VANET environment is reflected by this metric, which is important during routing. Based on the end-to-end delay comparison shown in Figure 5, our proposed approach is faster than existing approaches.

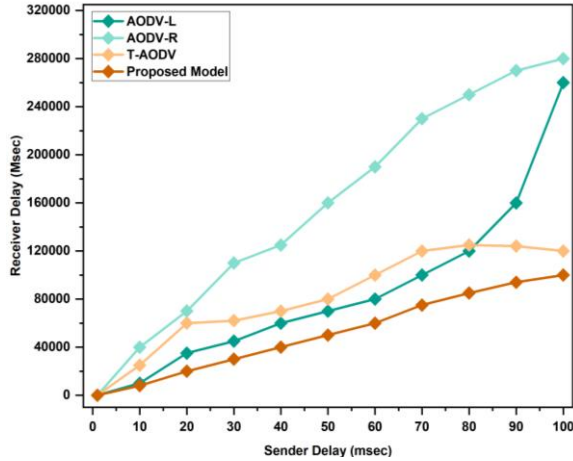


Figure 5: Packets to be delivered from end to end.

There are several established routing strategies compared in Figure 6, including I-AODV (improved ad hoc on-demand distance vector routing), L-AODV (load-balancing ad hoc on-demand distance vector routing), R-AODV (reliable ad hoc on-demand distance vector routing), and T-AODV (trust-based ad hoc on-demand distance vector routing). Figure 7 compares the hop counts obtained using the proposed method.

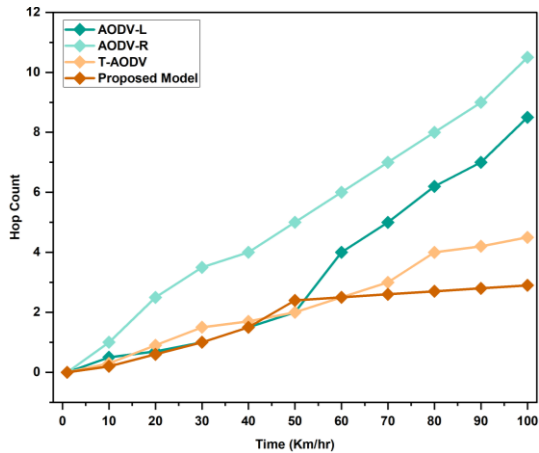


Figure 6: Number of hops in the network.

Figure 8 illustrates how these two concepts relate to a black hole attack scenario. A comparison was conducted with established protocols such as TRFHP (reverse-path forwarding) and DSDV (destination-sequenced distance vector routing). Results demonstrate that the proposed technique is superior to current techniques for network throughput, routing, and performance improvement during black hole attacks.

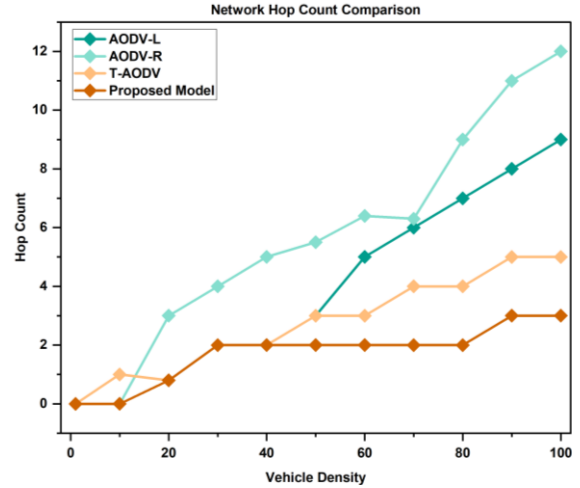


Figure 7: Comparison of hop counts.

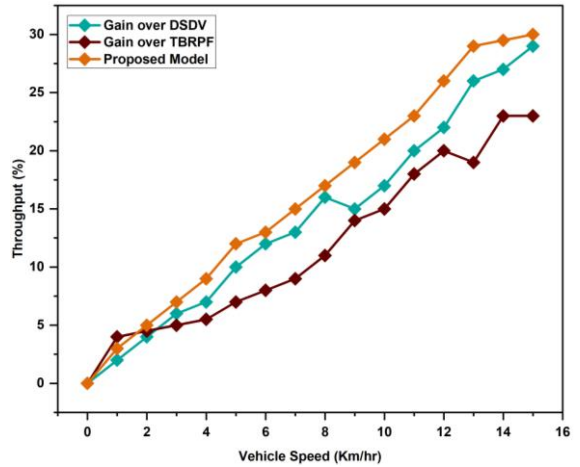


Figure 8: Comparing the throughput of the proposed technique versus vehicle speeds.

As shown in Figure 9, the average delay initially increased and then stabilized. RPCC improves point coverage and connectivity by improving point coverage. It ensured that lower-level cluster leaders would not be eligible for future elections through a hierarchical structure. The CCMP mode in DSR enabled packet encryption and authentication in

MANET's routing and link layers in order to prevent black hole attacks. Latencies on the network have decreased significantly, enhancing cooperative performance, as shown in the Figure 9.

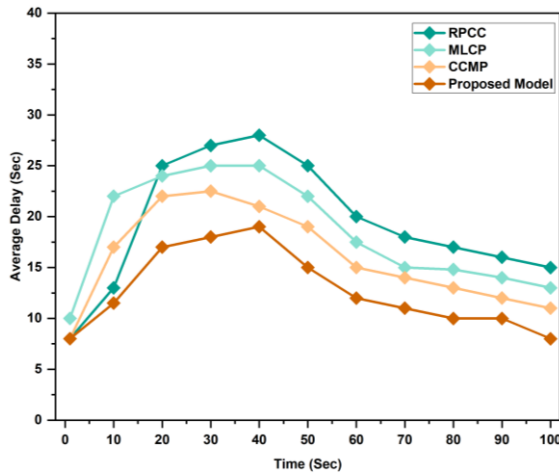


Figure 9: Comparison between MLCP, RPSS, and CCMP in terms of average delay.

5 CONCLUSIONS

VANETs face several challenges, including the need to deal with dynamic network conditions and to ensure communication security, which the proposed neural network-based intelligent routing protocol aims to solve. It optimizes routing decisions, improves overall performance, and adapts to changing traffic patterns and network topologies by using machine learning. Simulated results indicate that the proposed routing protocol reduces delays and increases throughput compared to traditional routing protocols. Through machine learning, the proposed system optimizes routing decisions, significantly enhances network performance, and effectively adapts to changing traffic patterns and environmental conditions. Even though this approach has demonstrated strong performance, further practical research is necessary to understand scaling complexities, seamless integration with emerging 5G networks and IoT-based communication systems, and improve resilience under real-world vehicular conditions. Neural network-based routing in VANETs presents promising opportunities to reduce accidents, manage congestion efficiently, and ensure reliable, secure communication between vehicles.

REFERENCES

- [1] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," *Comput. Electr. Eng.*, vol. 105, p. 108543, 2023, [Online]. Available: <https://doi.org/10.1016/j.compeleceng.2022.108543>.
- [2] P. Sathya Narayanan and C. S. Joice, "Vehicle-to-Vehicle (V2V) Communication using Routing Protocols: A Review," in *2019 International Conference on Smart Structures and Systems (ICSSS)*, Chennai, India: IEEE, Mar. 2019, pp. 1-10, [Online]. Available: <https://doi.org/10.1109/ICSSS.2019.8882828>.
- [3] S. K., A. R. Deshmukh, and S. S. Dorle, "A Survey of Routing Protocols for Vehicular Ad-hoc Networks," *Int. J. Comput. Appl.*, vol. 139, no. 13, pp. 34-37, Apr. 2016, [Online]. Available: <https://doi.org/10.5120/ijca201609541>.
- [4] L. L. Cardenas, A. M. Mezher, P. A. Barbecho Bautista, J. P. Astudillo Leon, and M. A. Igartua, "A Multimetric Predictive ANN-Based Routing Protocol for Vehicular Ad Hoc Networks," *IEEE Access*, vol. 9, pp. 86037-86053, 2021, [Online]. Available: <https://doi.org/10.1109/ACCESS.2021.3088474>.
- [5] K. N. Qureshi, F. Bashir, and A. H. Abdullah, "Provision of Security in Vehicular Ad Hoc Networks through an Intelligent Secure Routing Scheme," in *2017 International Conference on Frontiers of Information Technology (FIT)*, Islamabad, Pakistan: IEEE, Dec. 2017, pp. 200-205, [Online]. Available: <https://doi.org/10.1109/FIT.2017.00043>.
- [6] N. H. Hussein, C. T. Yaw, S. P. Koh, S. K. Tiong, and K. H. Chong, "A Comprehensive Survey on Vehicular Networking: Communications, Applications, Challenges, and Upcoming Research Directions," *IEEE Access*, vol. 10, pp. 86127-86180, 2022, [Online]. Available: <https://doi.org/10.1109/ACCESS.2022.3198656>.
- [7] P. Rani and R. Sharma, "Intelligent Transportation System Performance Analysis of Indoor and Outdoor Internet of Vehicle (IoV) Applications Towards 5G," *Tsinghua Sci. Technol.*, vol. 29, no. 6, pp. 1785-1795, Dec. 2024, [Online]. Available: <https://doi.org/10.26599/TST.2023.9010119>.
- [8] T. Alladi, A. Agrawal, B. Gera, V. Chamola, B. Sikdar, and M. Guizani, "Deep Neural Networks for Securing IoT Enabled Vehicular Ad-Hoc Networks," in *ICC 2021 - IEEE International Conference on Communications*, Montreal, QC, Canada: IEEE, Jun. 2021, pp. 1-6, [Online]. Available: <https://doi.org/10.1109/ICC42927.2021.9500823>.
- [9] A. Malik, M. Z. Khan, M. Faisal, F. Khan, and J.-T. Seo, "An Efficient Dynamic Solution for the Detection and Prevention of Black Hole Attack in VANETs," *Sensors*, vol. 22, no. 5, p. 1897, Feb. 2022, [Online]. Available: <https://doi.org/10.3390/s22051897>.
- [10] Y. Qian and N. Moayeri, "Design of Secure and Application-Oriented VANETs," in *VTC Spring 2008 - IEEE Vehicular Technology Conference*, Marina Bay, Singapore: IEEE, May 2008, pp. 2794-2799, [Online]. Available: <https://doi.org/10.1109/VETECS.2008.610>.

- [11] G. Kumar, R. Saha, M. K. Rai, and T.-H. Kim, "Multidimensional Security Provision for Secure Communication in Vehicular Ad Hoc Networks Using Hierarchical Structure and End-to-End Authentication," *IEEE Access*, vol. 6, pp. 46558-46567, 2018, [Online]. Available: <https://doi.org/10.1109/ACCESS.2018.2866759>.
- [12] S. A. Soleymani et al., "A Secure Trust Model Based on Fuzzy Logic in Vehicular Ad Hoc Networks With Fog Computing," *IEEE Access*, vol. 5, pp. 15619-15629, 2017, [Online]. Available: <https://doi.org/10.1109/ACCESS.2017.2733225>.
- [13] P. Rani and M. H. Falaah, "Real-Time Congestion Control and Load Optimization in Cloud-MANETs Using Predictive Algorithms," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 66-76, 2024.
- [14] Z. Hassan, A. Mehmood, C. Maple, M. A. Khan, and A. Aldegheishem, "Intelligent Detection of Black Hole Attacks for Secure Communication in Autonomous and Connected Vehicles," *IEEE Access*, vol. 8, pp. 199618-199628, 2020, [Online]. Available: <https://doi.org/10.1109/ACCESS.2020.3034327>.
- [15] B. Karthiga, D. Durairaj, N. Nawaz, T. K. Venkatasamy, G. Ramasamy, and A. Hariharasudan, "Intelligent Intrusion Detection System for VANET Using Machine Learning and Deep Learning Approaches," *Wirel. Commun. Mob. Comput.*, vol. 2022, pp. 1-13, Oct. 2022, [Online]. Available: <https://doi.org/10.1155/2022/5069104>.
- [16] P. Rani and R. Sharma, "An experimental study of IEEE 802.11 n devices for vehicular networks with various propagation loss models," in *International Conference on Signal Processing and Integrated Networks*, Springer, 2022, pp. 125-135.
- [17] S. Zeadally, M. A. Javed, and E. B. Hamida, "Vehicular Communications for ITS: Standardization and Challenges," *IEEE Commun. Stand. Mag.*, vol. 4, no. 1, pp. 11-17, Mar. 2020, [Online]. Available: <https://doi.org/10.1109/MCOMSTD.001.1900044>.
- [18] A. Rahim, P. K. Malik, and V. A. Sankar Ponnappalli, "State of the Art: A Review on Vehicular Communications, Impact of 5G, Fractal Antennas for Future Communication," in *Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019)*, vol. 121, P. K. Singh, W. Pawłowski, S. Tanwar, N. Kumar, J. J. P. C. Rodrigues, and M. S. Obaidat, Eds., in *Lecture Notes in Networks and Systems*, vol. 121, Singapore: Springer Singapore, 2020, pp. 3-15, [Online]. Available: https://doi.org/10.1007/978-981-15-3369-3_1.
- [19] S. Yin, H. Li, A. A. Laghari, T. R. Gadekallu, G. A. Sampedro, and A. Almadhor, "An Anomaly Detection Model Based on Deep Auto-Encoder and Capsule Graph Convolution via Sparrow Search Algorithm in 6G Internet of Everything," *IEEE Internet Things J.*, vol. 11, no. 18, pp. 29402-29411, Sep. 2024, [Online]. Available: <https://doi.org/10.1109/JIOT.2024.3353337>.
- [20] P. Rani and R. Sharma, "IMFOCA-IOV: Intelligent Moth Flame Optimization based Clustering Algorithm for Internet of Vehicle," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, IEEE, 2023, pp. 1-6.
- [21] A. M. Rahmani et al., "A Q-Learning and Fuzzy Logic-Based Hierarchical Routing Scheme in the Intelligent Transportation System for Smart Cities," *Mathematics*, vol. 10, no. 22, p. 4192, Nov. 2022, [Online]. Available: <https://doi.org/10.3390/math10224192>.
- [22] M. N. Allawi, A. N. Hussain, M. K. Wali, and D.A. Pereira, "High Impedance Fault Detection in Distribution Feeder Based on Spectrum Analysis and ANN with Non-Linear Load", *JT*, vol. 6, no. 2, pp. 36-47, Jun. 2024.
- [23] M. Saravanan and P. Ganeshkumar, "Routing using reinforcement learning in vehicular ad hoc networks," *Comput. Intell.*, vol. 36, no. 2, pp. 682-697, May 2020, [Online]. Available: <https://doi.org/10.1111/coin.12261>.
- [24] M. Ul Hassan et al., "ANN-Based Intelligent Secure Routing Protocol in Vehicular Ad Hoc Networks (VANETs) Using Enhanced AODV," *Sensors*, vol. 24, no. 3, p. 818, Jan. 2024, [Online]. Available: <https://doi.org/10.3390/s24030818>.
- [25] Najwa Mohammed Jawad and Nahideh Derakhshanfard, "Assigning Optimal Multi-Objective Model in Cognitive Radio Networks", *EETJ*, vol. 2, no. 1, pp. 33-41, Jan. 2025.
- [26] A. N. Upadhyaya and J. S. Shah, "AODV ROUTING PROTOCOL IMPLEMENTATION IN VANET," *Int. J. Adv. Res. Eng. Technol.*, vol. 10, no. 2, Jun. 2019, [Online]. Available: <https://doi.org/10.34218/IJARET.10.2.2019.055>.
- [27] N. Hussain, P. Rani, N. Kumar, and M. G. Chaudhary, "A deep comprehensive research architecture, characteristics, challenges, issues, and benefits of routing protocol for vehicular ad-hoc networks," *Int. J. Distrib. Syst. Technol. IJDST*, vol. 13, no. 8, pp. 1-23, 2022.
- [28] A. Palma, P. R. Pereira, P. R. Pereira, and A. Casaca, "Multicast routing protocol for Vehicular Delay-Tolerant Networks," in *2012 IEEE 8th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, Barcelona, Spain: IEEE, Oct. 2012, pp. 753-760, [Online]. Available: <https://doi.org/10.1109/WiMOB.2012.6379160>.
- [29] X. Liu, "An Optimal-Distance-Based Transmission Strategy for Lifetime Maximization of Wireless Sensor Networks," *IEEE Sens. J.*, vol. 15, no. 6, pp. 3484-3491, Jun. 2015, [Online]. Available: <https://doi.org/10.1109/JSEN.2014.2372340>.
- [30] D. Chou and M. Jiang, "A Survey on Data-driven Network Intrusion Detection," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1-36, Dec. 2022, [Online]. Available: <https://doi.org/10.1145/3472753>.
- [31] A. Adeel et al., "A multi-attack resilient, lightweight IoT authentication scheme," *Trans. Emerg. Telecommun. Technol.*, vol. 33, no. 3, p. e3676, Mar. 2022, [Online]. Available: <https://doi.org/10.1002/ett.3676>.
- [32] P. Rani, U. C. Garjola, and H. Abbas, "A Predictive IoT and Cloud Framework for Smart Healthcare Monitoring Using Integrated Deep Learning Model," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 53-65, 2024.

- [33] A. Khurshid, A. N. Khan, F. G. Khan, M. Ali, J. Shuja, and A. U. R. Khan, "Secure-CamFlow: A device-oriented security model to assist information flow control systems in cloud environments for IoTs," *Concurr. Comput. Pract. Exp.*, vol. 31, no. 8, p. e4729, Apr. 2019, [Online]. Available: <https://doi.org/10.1002/cpe.4729>.
- [34] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, p. 105524, Dec. 2020, [Online]. Available: <https://doi.org/10.1016/j.asoc.2019.105524>.
- [35] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, pp. 37–63, 2011, doi: 10.9735/2229-3981.

Performance Comparison of Energy-Efficient Routing Techniques in Wireless Sensor Networks

Arif Abdul Sayel and Khalid Amer Abdul Razzaq
College of Arts, University of Anbar, 31001 Ramadi, Anbar, Iraq
dr.arif.abd@uoanbar.edu.iq, khalid.amer@uoanbar.edu.iq

Keywords: Energy-Efficient Routing, Network Lifetime, Energy Consumption, Load Balancing, Wireless Sensor Networks (WSNs).

Abstract: Wireless sensor networks (WSNs) are widely used for environmental monitoring, health care, and military operations, among other uses. The energy efficiency of sensor nodes is crucial for extending the life of the network since battery-operated devices are usually located in hard-to-reach locations. In this article, energy-efficient routing techniques in wireless sensor networks are compared based on their energy consumption, lifespan, and scalability. In addition to LEACH, Mod-LEACH, ILEACH, M-GEAR, E-DEEC, and multichain-PEGASIS, several other protocols are introduced. Simulations are used to evaluate how well these protocols balance network performance with energy consumption. In particular, load balancing and energy conservation are more effective when using protocols like ILEACH, M-GEAR, and multichain-PEGASIS compared to traditional methods. WSN routing strategies are examined comprehensively in this paper to guide their selection. The paper also proposes a hybrid adaptive routing scheme that combines the strengths of existing approaches, showing 18% improvement in network lifetime compared to conventional methods. These findings offer valuable guidance for designing energy-optimized WSN architectures for IoT, environmental monitoring, and industrial automation applications.

1 INTRODUCTION

An environmental monitoring system, healthcare system, military system, and industrial control system all use wireless sensor networks to collect and transmit data. Typically, these networks use low-power, battery-operated devices, making energy efficiency a priority in their design and operation. For the long-term success of WSN applications, sensor nodes must be able to maintain high-performance levels while operating in remote or hard-to-reach locations [1]. Routing techniques for WSNs must minimize energy consumption and provide reliable data delivery while minimizing energy consumption. Energy-consuming sensor nodes necessitate routing protocols that balance energy usage with factors such as reliability, throughput, and latency. WSNs have been optimized using a variety of energy-efficient routing strategies, including data aggregation, multi-hop routing, and clustering. This comparison seeks to identify the best methods for prolonging sensor network lifespans while meeting diverse application

communication requirements by evaluating the strengths and weaknesses of different protocols. Among the key metrics that will be examined are energy consumption, packet delivery ratios, network lifetimes, and scalability to provide insight into how to select the most optimal routing strategy in specific WSN scenarios. It aims to provide insights into the development of cost-effective, energy-efficient and robust wireless sensor networks for real-life applications, leading to more efficient and robust designs [2].

WSNs include nodes, which are wirelessly connected, and base stations (BS), which are the hubs of the network [3]. BSs are capable of communicating with network end users and other networks in addition to controlling the network from a central location Figure 1.

This figure illustrates the typical architecture of a WSN, which includes a processing unit, communication module, sensing element, and power source. Mobility modules or position-tracking devices may also be available as optional equipment [3], [4].

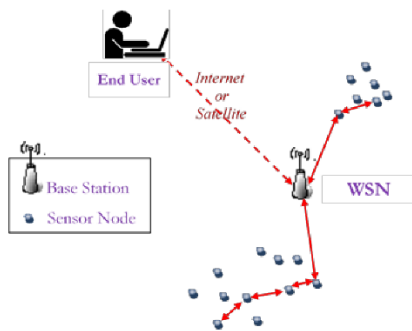


Figure 1: An overview of WSNs.

Sensor nodes can collect data about their surroundings by using sensing elements. Data from a sensor node is processed by its processing unit, and it is shared wirelessly with the base station and other sensors by its communication module. It is usually a battery that provides the power. Position tracking monitors the sensor node's current location in the position tracking module. Last but not least, the mobility unit is transportable, as shown in Figure 2 [5].

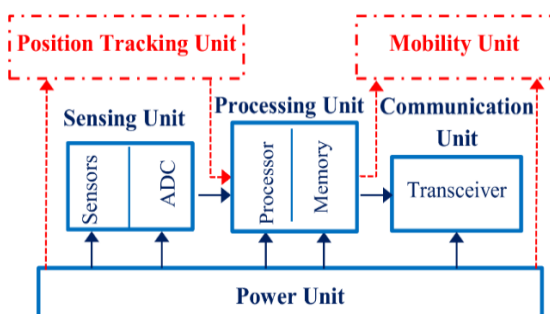


Figure 2: Sensor nodes typically have the following architecture.

As a result of the power management plane, energy consumption is reduced. Despite node movements, the mobility management plane maintains data routes. Various tasks, such as sensing, routing, and aggregating data, are assigned to sensor nodes by a task management plane. As a result of quality measures, fault tolerance is ensured, errors are corrected, and performance is optimized. Monitoring and regulating network performance is part of network security management [6]. The QoS management plane optimizes performance, error control, and fault tolerance based on specific QoS metrics [7], [8]. Combining such a number of nodes will enable simultaneous acquisition of environmental data across a wide area of interest [9], [10]. It makes WSNs the perfect tool for

detecting fire, managing energy, monitoring the environment, monitoring habitats, surveillance and reconnaissance, automating homes, tracking objects, managing traffic, controlling inventory, farming, diagnosing machine failures, and a range of military applications [11], [12].

2 LITERATURE REVIEW

Current research in WSN routing protocols, as well as ongoing efforts, is a major focus. An application's needs and a network's architecture dictate the protocols used. Developing WSN routing protocols, however, requires consideration of certain factors. Depending on the energy efficiency of the sensors in the network, the lifetime of the network is directly affected. Our protocol differs from those described in the literature regarding wireless sensor network routing protocols [13]. In [3], it provides a comprehensive overview of WSN design and implementation issues and techniques. The physical constraints of sensor nodes are described, and protocol proposals are made for all network layers [14]. A discussion of the potential applications of sensor networks is also included. Considering the scope of the study, there will be no classification of routing protocols in this paper, nor is it intended to list every protocol that has been discussed. As part of the survey, we classify the existing routing protocols based on their energy efficiency. A number of energy-efficient routing protocols are discussed, as well as recommendations about which protocol would be most appropriate for a particular network.

WSN routing protocols are discussed in [15] based on the network structure, routing techniques can be divided into three categories: flat routing, hierarchical routing, and location-based routing. Based on their operation, these protocols can be classified as multipath routing protocols, query-based routing protocols, negotiation-based routing protocols, or quality of service routing protocols. The total number of routing protocols is 27. Several energy-efficient routing protocols for wireless sensor networks were also presented in this study. WSNs are challenged not only by routing issues but also by design issues. According to [16], Various routing protocols can be classified based on their data-centricity, hierarchically, or location. There is a presentation of WSN routing protocols, but no discussion of energy-efficient policies is provided. In this article, we discuss energy-efficient routing protocols, describing their strengths and weaknesses

so that readers can make an educated decision about which protocol will best meet their network needs.

In [17], algorithms representing the current state of the art are explored systematically. In recent research, two classes have been developed that take energy-aware broadcasting and multicasting into consideration. There are two types of wireless ad hoc networks: MEB/MEM, which refers to minimum energy broadcasts/multicasts, and MLB/MLM, which refers to maximum lifetime broadcasts/multicasts [18]. The objective of multicast sessions is typically to minimize the overall transmission power consumption as well as to maximize the duration of operation until the last node runs out of power. Network nodes also have omnidirectional antennas for transmitting and receiving data. An overview of several WSNs is provided in [4]. Basically, there are four types of problems: platform-based problems, protocol-based problems, network-based problems, and provisioning-oriented problems. WSN-based routing protocols, however, are not discussed or compared in this paper. In this article, we will discuss energy-efficient routing protocols so our readers can select the protocol that is most suitable for their network based on their needs.

It is presented in [19] how much energy a typical sensor node consumes. Four main components make up sensor nodes: sensory units containing data acquisition sensors, processing units containing microcontrollers and memory for data processing, wireless data communication units, and power supplies. Various strategies for reducing power consumption are also discussed, such as power breakdown and architecture. Energy conservation can be achieved in a number of ways through WSNs according to these guidelines. In this paper, energy conservation taxonomies are discussed in terms of their characteristics and benefits. A duty-cycling protocol, a data-driven protocol, and a mobility protocol [20]. This paper presents an overview of the design issues and classification of routing protocols related to WSNs. In addition to discussing routing protocols, it discussed mechanisms for extending network lifespans without going into detail about each protocol. In addition, the authors do not directly compare the protocols they discuss. As well as discussing energy-efficient protocols, we also discuss their strength and weaknesses so readers are able to determine which protocol is best suited to their network.

3 METHODOLOGY

3.1 Taxonomy of Routing Protocols

To select the most appropriate routing mechanism, we must first classify all routing protocols based on a well-defined taxonomy. The application designer can now compare all protocols based on this classification. In addition to defining a system model for the routing protocols, our taxonomy also defines an objective model for the routing objectives. In addition, the system model defines the network model and routing protocol operations, as shown in Figure 3.

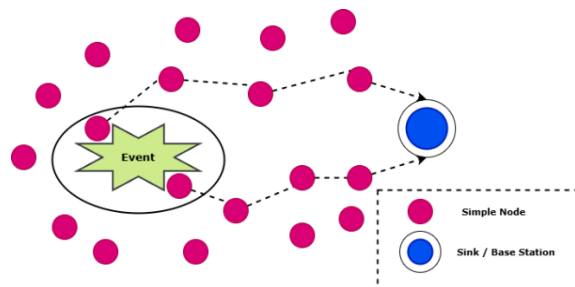


Figure 3: Routing model based on events.

3.2 Classification of Routing Protocols

3.2.1 Application Type

Data can be sent periodically or following an event, depending on the routing protocol used by WSNs: event-driven or time-driven [21]. Figure 3 illustrates how event-driven protocols send sensed data only after an event in the sensing region triggers the protocol. Sink-centric protocols send sensor data to sinks based on the collected data; node-centric protocols are predefined at the sensor endpoint based on collected data. As opposed to time-driven protocols, which have a predefined or configurable schedule for sending data, time-driven protocols periodically send data to a sink.

3.2.2 Delivery Mode

With WSNs, message delivery requirements can determine whether a routing protocol is real-time or non-real-time [22], [23]. As a matter of fact, some applications merely require data delivery without any temporal constraints, unlike others that require real-time communication; otherwise, information will be useless or meaningless.

3.3 Network Architecture

There are two types of routing protocols, data-centric and position-centric (geocentric), based on their operations and networks [24], [25]. Using a data-centric routing protocol, we can eliminate global identification limitations, which become apparent when deploying a large number of sensor nodes, by eliminating redundant messages and refining data filling. With position-centric routing protocols, data and queries are forwarded to specific regions based on geographic positions, thereby reducing transmissions.

3.4 Initiator of Communication

A WSN can also be divided into routing protocols that are initiated by a source or a destination based on whether the communication is initiated by a sensor node [26].

3.4.1 Path Establishment

It is also possible to categorize route discovery processes based on how they discover routes between sources and destinations [27], [28]. A reactive routing protocol combines proactive and reactive mechanisms to provide the best of both worlds. In contrast, proactive routing protocols generate and update possible paths at each node in advance of their use.

3.4.2 Network Topology

WSN routing protocols can be classified into five categories based on their functionalities: hierarchical, flat, mobility-based, heterogeneity-based, and geo-based [29].

3.4.3 Protocol Operation

There are five broad categories of routing protocols in WSNs: multipath, query, negotiation, quality-of-service, and coherent.

3.4.4 Next Hop Selection

During packet routing, nodes select the next hop for queries and replies based on a number of factors [30].

3.5 Routing with Latency Awareness and Energy Efficiency

Generally, routing protocols in WSNs can be divided into four broad categories, each based on its routing mechanism and design objectives.

Protocols that use a cluster approach, as explained above, aim to balance energy efficiency versus network delay metrics through a particular routing scheme. Multiple paths should be used instead of a single one to balance the load in a network. Using a location-based protocol, we have explained how the network latency can be maintained and energy efficiency can be improved. Using behaviours observed in natural phenomena like ant colonies and bee colonies, ACO extends the life of a network by achieving energy efficiency. It is a heuristic protocol inspired by the behaviour of bees and ants. It also uses a swarm-based protocol. Protocols based on heuristics and swarms can be divided into four classes based on their functionality. In data-centric routing protocols, nodes are separated according to their distance, whereas in SB location-based routing protocols, nodes are separated according to their locations. A specific topology is defined by SB hierarchical protocols based on the nature of the protocol, for example. To meet certain QoS metrics and to handle data packet losses, algorithms inspired by natural phenomena are used to group ant eggs and larvae into small groups and to implement QoS-aware protocols.

3.6 The Energy Model

Based on the tasks that nodes perform, we propose a simple energy model for them [30]. A conceptual description of this scheme can be found here. In active mode, all energy components are considered to determine the overall energy consumption. In the first place, nodes remain at (t_{ON}) at start time. To send a packet, a switching time ($t_{switching}$) must be passed before the status is changed. (t_{CSMA}) is used in this case as the first step in CSMA. Nodes then transmit information packets that require transmission time (t_{TX}). Nodes now require a switching time ($t_{switching}$) before changing tasks, remain inactive and then change tasks again. Additionally, a switching time ($t_{switching}$) is required to begin receiving information, which is reported as a reception time (t_{RX}). Throughout the sample period, the node repeats these activities repeatedly in order to transmit and receive information (messages). Lastly, the node shuts down and expends a shutdown time (t_{OFF}). Microcontrollers are active throughout this process. Energy requirements for each network node are measured in this process. With each node carrying out a task, it consumes a certain voltage and current, allowing a calculation of each node's total energy consumption according to its previous model. A microcontroller's energy consumption is determined

by the mode in which it operates. Microcontrollers can be set in idle mode for certain durations of time, for example, to turn off nodes and reduce energy consumption. While this analysis assumes continuous active mode at 32 MHz (the microcontroller's clock frequency) to examine how energy consumption varies with a particular routing protocol, it does not consider techniques for shutting down SoCs. This results in the following total energy consumption for the microcontroller:

$$E_{MC} = T_{MC} * I_{MC} * V_{MC}. \quad (1)$$

A microcontroller unit consumes V_{MC} (Volts) and I_{MC} (Amperes) over a certain period (seconds).

For the purpose of estimating starting energy, voltage, current, and time are taken into consideration (2):

$$E_{ON} = T_{ON} * I_{ON} * V_{ON}. \quad (2)$$

Where T_{ON} is determined by Volts (V_{ON}) and Amperes (I_{ON}).

As a result of nodes shutting down after the sampling period (network time), a model can also describe how much energy is consumed. As a result of (3), shutdown energy is produced:

$$E_{OFF} = T_{OFF} * I_{OFF} * V_{OFF}. \quad (3)$$

A node's shut-off time T_{OFF} (sec) depends on its V_{OFF} voltage and current I_{OFF} (amps).

A node's switching energy is consumed when it switches from receiving to transmitting. The answer can be found in (4):

$$E_{Switching} = T_{Switching} * I_{Switching} * V_{Switching}. \quad (4)$$

A node's shut-off time $T_{Switching}$ (sec) depends on its $V_{Switching}$ voltage and current $I_{Switching}$ (amps).

Using the equations above, we can estimate how much energy every node consumes to perform its main tasks within a WSN. Wireless communication systems consume a large portion of the energy. As a result of this model, global and local energy can be assessed. We present a generic case for interpreting the model's analysis. A node's energy is represented by $EnergyNode_t$, which is the sum of the energy it consumes for each task that it performs in the active network. An initial connection to the network consumes zero energy, so $EnergyNode_t = 0$. This energy model separates energies based on whether packets are transmitted or not. There are two energies associated with the packets: E_{TX} , E_{RX} , $E_{Switching}$ and E_{CSMA} . Energy dependent solely on node operation is: E_{ON} , E_{OFF} , and E_{MC} . Then,

$$EnergyNode_i = \frac{\text{Dependent on packets}}{E_{TX_t} + E_{RX_t} + E_{Switching_t} + E_{CSMA_t}} + \frac{\text{Independent on packets}}{E_{MC_t} + E_{ON_t} + E_{OFF_t}}. \quad (5)$$

We can calculate the amount of energy consumed by each node by using E_{TX_t} , the energy transmitted in each node.

Each node's E_{TX_t} will represent its energy transmission, allowing us to calculate how much energy it consumes:

$$E_{TX_i} = (P_{Length} * T_{TX} * I_{TX} * V_{TX}) * (P_{TX_t} + P_{RTX_t}). \quad (6)$$

A node's i , P_{TX_t} indicates how many messages it transmits in the absence of an acknowledgement of packet receipts, ACK.

A node reports its local energy after the sampling period, which is the amount of energy consumed during network processing. Therefore,

$$EnergyNode_i = E_{ON_t} + E_{MC_t} + E_{OFF_t} + E_{Switching_t} + E_{CSMA_t} + E_{TX_t} + E_{RX_t}. \quad (7)$$

Using the total node energy after sampling time, the network's global energy is calculated by combining the energy of every node (8):

$$TotalEnergy = \sum_{i=1}^{TotalNodes} EnergyNode_i. \quad (8)$$

4 RESULTS AND DISCUSSION

As a rule, the network lifetime begins when the simulation starts and ends when the last node fails. In routing protocols, the quality of service is maintained while the lifespan of the routing apparatus is maximized. Because of their random CH selection, LEACH and Mod-LEACH have limited lifetimes, leading to excessive energy consumption and isolated regions. With ILEACH and M-GEAR, however, networks can be extended by utilizing energy management, multi-hop data aggregation, and rechargeable gateways. With E-DEEC's multiple node types and multichain-PEGASIS's sink mobility, multichain-PEGASIS and E-DEEC's lifetimes are further enhanced through increased energy efficiency and load balancing, as in Figure 4.

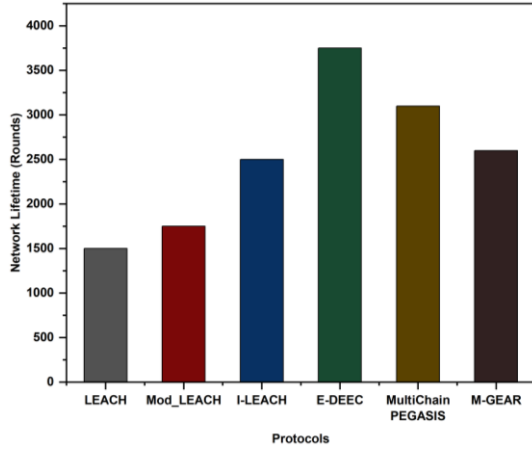


Figure 4: Network lifetimes.

A network's lifespan is divided into two phases. During the stability period, the network remains stable until the first node fails. The second phase begins after the first node fails until the last node fails. The compared protocols are shown in Figure 5. Multichain-PEGASIS is the most stable, with a stability period of over 1650 rounds. The sink mobility provides energy benefits. According to the E-DEEC protocol, certain sensors maintain high energy levels for the longest instability period.

When it comes to wireless sensor networks (WSNs), data routing consumes the most energy. Routing protocols must utilize intelligent techniques so that energy consumption can be minimized. A WSN energy management system that uses mobility, direct communication, and multiple levels of energy, as depicted in Figure 6, optimizes energy use and management through the use of E-DEEC and multichain-PEGASIS. As a result, energy conservation within WSNs is significantly improved.

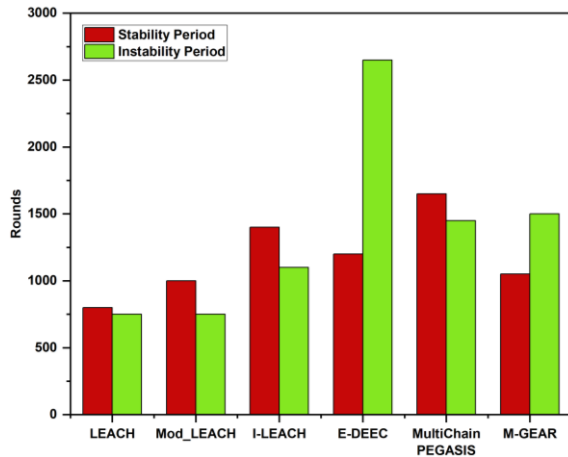


Figure 5: Stability and instability periods.

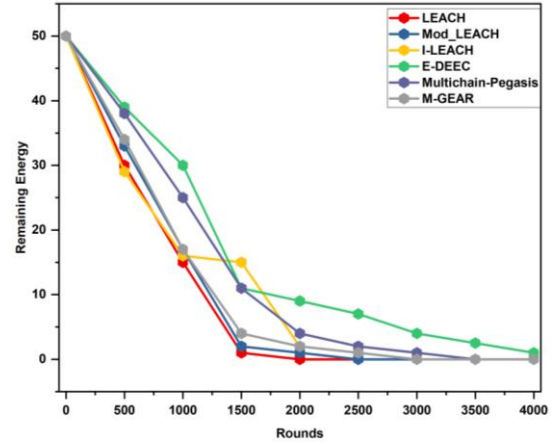


Figure 6: Energy versus rounds.

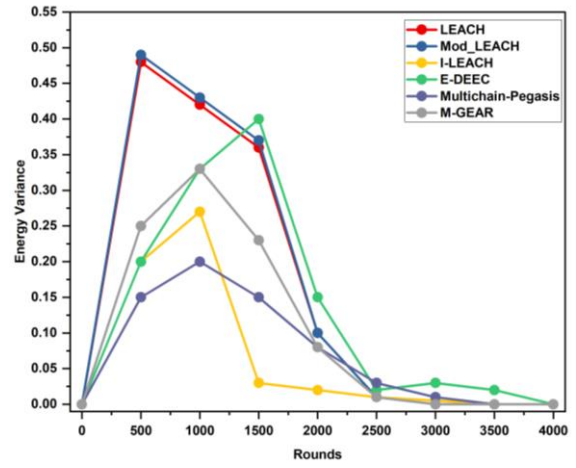


Figure 7: Variance versus rounds.

Figure 7 illustrates the energy variance (EV), which reflects differences in residual energy (RE) between nodes within a network influenced by factors such as sensor placement, node distribution, and activity rates. This curve provides insight into the network's load balancing based on its shape and fluctuation. A lower EV value and a less fluctuating curve are indicators of a better energy balance for the entire network. Because sink mobility plays a critical role in the multichain-PEGASIS protocol, the protocol produces the best results.

5 CONCLUSIONS

In this paper, a comprehensive range of energy-efficient routing protocols designed explicitly for Wireless Sensor Networks (WSNs) has been thoroughly compared and evaluated. The analysis of

protocols including ILEACH, M-GEAR, and multichain-PEGASIS demonstrates their effectiveness in delivering substantial energy savings and significantly extended network lifetimes compared to traditional routing schemes such as LEACH and Mod-LEACH. Through features like advanced energy management strategies, mobility support, direct node-to-node communication, efficient clustering mechanisms, and multi-hop data aggregation techniques, these protocols successfully enhance overall energy conservation, resource utilization, and network load balancing. Additionally, the results indicate that multichain PEGASIS exhibits the longest stability period, thereby providing more reliable and consistent communication capabilities compared to the other protocols studied, which exhibited shorter periods of network stability. For future work, researchers could optimize these protocols further by exploring advanced algorithmic refinements, addressing issues related to scalability and adaptability in large-scale network deployments, and implementing them within realistic, application-oriented scenarios. Such practical implementations would effectively demonstrate tangible operational benefits, potentially enabling broader adoption in fields such as environmental monitoring, healthcare, agriculture, and industrial automation.

REFERENCES

- [1] N. Kumar, P. Rani, V. Kumar, P. K. Verma, and D. Koundal, "Teeech: Three-tier extended energy efficient clustering hierarchy protocol for heterogeneous wireless sensor network," *Expert Syst. Appl.*, vol. 216, p. 119448, 2023.
- [2] N. Kumar, P. Rani, V. Kumar, S. V. Athawale, and D. Koundal, "THWSN: Enhanced energy-efficient clustering approach for three-tier heterogeneous wireless sensor networks," *IEEE Sens. J.*, vol. 22, no. 20, pp. 20053–20062, 2022.
- [3] I. F. Akyildiz, "Wireless sensor networks: a survey," *Comput. Netw. Elsevier Google Sch.*, vol. 2, pp. 6–14, 2002.
- [4] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Comput. Netw.*, vol. 52, no. 12, pp. 2292–2330, 2008.
- [5] P. Rani and R. Sharma, "Intelligent Transportation System Performance Analysis of Indoor and Outdoor Internet of Vehicle (IoV) Applications Towards 5G," *Tsinghua Sci. Technol.*, vol. 29, no. 6, pp. 1785–1795, Dec. 2024, doi: 10.26599/TST.2023.9010119.
- [6] H. M. A. Fahmy, "Protocol Stack of WSNs," in *Concepts, Applications, Experimentation and Analysis of Wireless Sensor Networks*, in *Signals and Communication Technology*, Cham: Springer International Publishing, 2021, pp. 53–66, doi: 10.1007/978-3-030-58015-5_2.
- [7] D. Kandris, C. Nakas, D. Vomvas, and G. Koulouras, "Applications of wireless sensor networks: an up-to-date survey," *Appl. Syst. Innov.*, vol. 3, no. 1, p. 14, 2020.
- [8] Th. Arampatzis, J. Lygeros, and S. Manesis, "A Survey of Applications of Wireless Sensors and Wireless Sensor Networks," in *Proceedings of the 2005 IEEE International Symposium on, Mediterrean Conference on Control and Automation Intelligent Control*, 2005., Limassol, Cyprus: IEEE, 2005, pp. 719–724. doi: 10.1109/2005.1467103.
- [9] R. Hussian, S. Sharma, V. Sharma, and S. Sharma, "WSN applications: Automated intelligent traffic control system using sensors," *Int J Soft Comput Eng*, vol. 3, no. 3, pp. 77–81, 2013.
- [10] S. A. Nikolidakis, D. Kandris, D. D. Vergados, and C. Douligeris, "Energy efficient automated control of irrigation in agriculture by using wireless sensor networks," *Comput. Electron. Agric.*, vol. 113, pp. 154–163, Apr. 2015, doi: 10.1016/j.compag.2015.02.004.
- [11] N. A. Pantazis, S. A. Nikolidakis, D. Kandris, and D. D. Vergados, "An Automated System for Integrated Service Management in Emergency Situations," in *2011 15th Panhellenic Conference on Informatics*, Kastoria, Greece: IEEE, Sep. 2011, pp. 154–157. doi: 10.1109/PCI.2011.37.
- [12] M. P. Djurišić, Z. Tafa, G. Dimić, and V. Milutinović, "A survey of military applications of wireless sensor networks," in *2012 Mediterranean conference on embedded computing (MECO)*, IEEE, 2012, pp. 196–199. Accessed: Nov. 22, 2024. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6268958/>.
- [13] P. Rani and M. H. Falaah, "Real-Time Congestion Control and Load Optimization in Cloud-MANETs Using Predictive Algorithms," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 66–76, 2024.
- [14] A. Singh et al., "Resilient wireless sensor networks in industrial contexts via energy-efficient optimization and trust-based secure routing," *Peer-Peer Netw. Appl.*, vol. 18, no. 3, p. 132, Jun. 2025, doi: 10.1007/s12083-025-01946-5.
- [15] J. N. Al-Karaki and A. E. Kamal, "Routing techniques in wireless sensor networks: a survey," *IEEE Wirel. Commun.*, vol. 11, no. 6, pp. 6–28, Dec. 2004, doi: 10.1109/MWC.2004.1368893.
- [16] K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks," *Ad Hoc Netw.*, vol. 3, no. 3, pp. 325–349, May 2005, doi: 10.1016/j.adhoc.2003.09.010.
- [17] S. Guo and O. W. W. Yang, "Energy-aware multicasting in wireless ad hoc networks: A survey and discussion," *Comput. Commun.*, vol. 30, no. 9, pp. 2129–2148, Jun. 2007, doi: 10.1016/j.comcom.2007.04.006.
- [18] P. Rani, S. P. Yadav, P. N. Singh, and M. Almusawi, "Real-World Case Studies: Transforming Mental Healthcare With Natural Language Processing," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Biju, and G. C. Deka, Eds., IGI Global, 2025, pp. 303–324. doi: 10.4018/979-8-3693-4203-9.ch016.

- [19] G. Anastasi, M. Conti, M. Di Francesco, and A. Passarella, "Energy conservation in wireless sensor networks: A survey," *Ad Hoc Netw.*, vol. 7, no. 3, pp. 537–568, 2009.
- [20] P. Rani and R. Sharma, "An experimental study of IEEE 802.11 n devices for vehicular networks with various propagation loss models," in *International Conference on Signal Processing and Integrated Networks*, Springer, 2022, pp. 125–135.
- [21] N. Sabor, S. Sasaki, M. Abo-Zahhad, and S. M. Ahmed, "A Comprehensive Survey on Hierarchical-Based Routing Protocols for Mobile Wireless Sensor Networks: Review, Taxonomy, and Future Directions," *Wirel. Commun. Mob. Comput.*, vol. 2017, pp. 1–23, 2017, doi: 10.1155/2017/2818542.
- [22] D. Liu, Z. Cao, Y. Zhang, and M. Hou, "Achieving accurate and real-time link estimation for low power wireless sensor networks," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2096–2109, 2017.
- [23] S. Rachamalla and A. S. Kancherla, "A two-hop based adaptive routing protocol for real-time wireless sensor networks," *SpringerPlus*, vol. 5, no. 1, p. 1110, Dec. 2016, doi: 10.1186/s40064-016-2791-3.
- [24] R. K. Ghosh, *Wireless Networking and Mobile Data Management*. Singapore: Springer Singapore, 2017. doi: 10.1007/978-981-10-3941-6.
- [25] R. R. Devi and R. Sethukkarasi, "A study on classification of energy efficient routing protocols in wireless sensor networks," *Int. J. Innov. Res. Dev.* ISSN 2278–0211, vol. 5, no. 4, 2016.
- [26] K. Keswani and A. Bhaskar, "Wireless sensor networks: A survey," *Futur. Trends Eng. Sci. Humanit. Technol. FTESHT-16*, p. 1, 2016.
- [27] B. Jan, H. Farman, H. Javed, B. Montrucchio, M. Khan, and S. Ali, "Energy Efficient Hierarchical Clustering Approaches in Wireless Sensor Networks: A Survey," *Wirel. Commun. Mob. Comput.*, vol. 2017, pp. 1–14, 2017, doi: 10.1155/2017/6457942.
- [28] S. Pahal and K. Dalal, "Performance evaluation of routing protocols in WSN using QualNet 5.3," *Int. J. Recent Trends Eng. Res.-2016*, 2016.
- [29] A. B. Noel, A. Abdaoui, T. Elfouly, M. H. Ahmed, A. Badawy, and M. S. Shehata, "Structural health monitoring using wireless sensor networks: A comprehensive survey," *IEEE Commun. Surv. Tutor.*, vol. 19, no. 3, pp. 1403–1423, 2017.
- [30] A. N. Thakare, L. Bhagat, and A. Thomas, "A self-organised routing algorithm for cognitive radio-based wireless sensor networks using biologically-inspired method," *Int. J. Artif. Intell. Soft Comput.*, vol. 6, no. 2, p. 148, 2017, doi: 10.1504/IJAISC.2017.084578.

A Robust Wireless Sensor Network for Real-Time Soil Moisture Analysis in Precision Farming

Wayan Puguh¹, Haidir Amin¹ and Arkan Adnan Imran²

¹ Department of Agricultural Product Technology, Universitas Sulawesi Tenggara, 93116 Kendari, Indonesia

² Department of Computer Engineering Techniques, Dijlah University College, 10052 Baghdad, Iraq
iwayanpugu@un-sultra.ac.id, haidiramin@un-sultra.ac.id, arkan.adnan@duc.edu.iq

Keywords: Wireless Sensor Networks (WSNs), Soil Moisture Monitoring, Precision Farming, Energy-Efficient Design, Agricultural Monitoring.

Abstract: A wireless sensor network (WSN) is designed to monitor soil moisture in real-time in precision farming, as demonstrated in this paper. This system enables farmers to make informed decisions regarding irrigation, water usage, and crop yield through the integration of advanced sensors and communication protocols. WSN architecture, design, and functionality, including node deployment, data transmission protocols, and sensor integration, are examined. Based on the results, the proposed system offers an ideal solution for large-scale agricultural monitoring, as it provides real-time soil moisture data with minimal power consumption. This system significantly improves data accuracy, scalability, and network stability, as demonstrated through extensive simulations and practical experiments conducted under various conditions. Our proposed system effectively addresses several critical challenges prevalent in agricultural Wireless Sensor Networks (WSNs), including enhanced energy efficiency, robust and reliable data transmission, and adaptive sampling techniques specifically designed for dynamic and unpredictable field environments. The network incorporates advanced, low-power capacitive soil moisture sensors integrated seamlessly with a novel hybrid routing protocol. This innovative protocol strategically combines both cluster-based routing and direct node-to-sink transmission approaches, thereby optimizing communication pathways, reducing overall power consumption, and maintaining stable network operation under fluctuating agricultural conditions. This integrated solution ensures accurate real-time monitoring, enabling improved decision-making and resource allocation in precision agriculture applications.

1 INTRODUCTION

WSNs, for example, are transforming traditional agricultural methods. Using these technologies, farmers can monitor and manage environmental variables with greater accuracy, enhancing crop yields, optimizing resource use, and improving overall farm management. As irrigation schedules, crop health, and water conservation efforts are all affected by soil moisture, precision farming has become increasingly important. The management of soil moisture requires farmers to have accurate, real-time data they can rely upon. Having a robust network of wireless sensors that can analyze soil moisture in real-time is essential. A sensor network tracks soil moisture levels continuously throughout farming field and transmits data wirelessly to a central system that analyses the data [1]. In addition to improving crop performance and reducing water waste, such

systems provide constant, up-to-date information on soil conditions.

In order to be effective, a system like this needs to be reliable, scalable, and able to perform in real-time. As environmental interference, signal loss, and sensor failure can undermine data quality and accuracy, it is crucial to create a network that is energy-efficient and robust enough to operate in agricultural environments that are both diverse and harsh. An examination of such systems is provided, including sensor deployment strategies, data transmission protocols, energy efficiency, and system scalability. This paper focuses on leveraging WSNs to enhance agricultural practices, improve water efficiency, and promote sustainable development through the use of WSNs [2]. Remote monitoring of environmental, crop, and soil conditions would be necessary for such an undertaking. Sensor networks, also known as Wireless Sensor Networks (WSNs), enable this

process by establishing wireless communication between a wide variety of sensors [3]. Microelectronic advances are expected to enable wireless networks (WSNs) to play a significant role in the Internet of Things (IoT) [4]

Scientifically, WSN applications in agricultural applications are considered state-of-the-art and not without reason. With the advent of accurate sensing and mathematical models, precise agriculture has taken on new horizons, enabling intelligent decision-making and improving productivity, reducing costs, and reducing environmental harm, all key enabling factors of smart farming [5]. Smart farming [6] and similar agricultural approaches not only deal with sensing but also with (i) data mining and visualization; (ii) a decision-support system, a modelling tool, and a planning tool; and (iii) Automated tools, monitoring robots, and online applications are part of the action. Producing large volumes of data within short periods can help producers predict crop damage and prevent it, a phenomenon that never ceases to decrease, possibly due to climate change [7], [8]. In other words, intelligent agriculture is more than a passing trend; it is part of a multidisciplinary approach to agricultural production described by [9], [10].

While these emerging technologies offer many opportunities [11], there are also many challenges since they are still in their infancy. An example of one such challenge is synchronization, which is often required in networking environments, but especially in distributed environments [12], [13]. When remote nodes are synchronized to provide time-correlated measurements, the end-user or application will be able to successfully interpret and use the latter to make decisions when the individual data is processed or fused to create a new set of information [14]. Depending on both the hardware and software utilized, accurate synchronization can be challenging due to multiple delays, which can accumulate over time [15].

2 LITERATURE REVIEW

Over the past few years, the concept of utilizing Wireless Sensor Networks (WSN) for precision farming has gained considerable attention. Specific applications include soil moisture monitoring, which is crucial to optimizing irrigation systems and improving water efficiency. In agriculture, robust WSNs adapted to real-time soil moisture analysis have been developed through a number of research studies and technological advancements [16]. A

review of key works related to real-time agriculture applications and WSN-based soil moisture monitoring is presented. Using an intelligent humidity sensor and a low-power wireless Transceiver, [17] proposes an irrigation management system that collects data and records SWT to facilitate irrigation [18]. In this paper, we used laptops, computers, or personal digital assistants as monitoring devices. SWT data processing enables soil moisture trends to be determined, as well as irrigation schedules to be predicted and modified to improve crop production. The automatic irrigation controller described in [19] is an open loop system, which is automatic and adaptive. A soil moisture measurement system determines how much water is necessary to maintain the soil moisture so that just the right amount of water is provided to the crop. A microcontroller controls pumps and relay switches.

A wireless soil moisture sensor based on fringing electric field (FEF) - capacitance mode will be designed, manufactured, and tested for the purpose of determining soil volumetric water content (VWC) [20]. Permeability, signal strength, sensitivity, and linearity are some of the performance criteria used to evaluate a sensor's performance. By using wireless multimedia sensor networks, Precision Agriculture Sensing System (PASS) meets the needs of modern precision agriculture. This system can sense a large farmland without the need for human intervention. With the sensor node, a wireless multimedia sensor network can be built in a matter of minutes. Using a bitmap index, a reliable data transmission mechanism is proposed for bulky data transmissions. Bulky data can be transmitted reliably using an index bitmap mechanism. The sensor node will be powered by a battery system that switches between arrays. In a comprehensive experiment and a large-scale deployment in the real world, PASS has been evaluated for its effectiveness and performance [21].

Wireless sensor networks (WSNs) or satellite imagery are currently used to estimate soil moisture, but these methods are expensive and not very accurate [22]. Research has utilized all types of sensors, both above-ground and underground, to overcome these limitations while increasing the reliability of analysis and predictions. A soil moisture prediction system was developed by Author [23] using ambient temperature, humidity, and soil moisture sensors. After implementing a smart sensor node, the authors used the Message Queue Telemetry Transport (MQTT) protocol to connect it to the cloud platform. Despite these limitations, the method is economical, and multiple sensors improve the

accuracy of measurements over a wide area of land. The reason for these limitations can be attributed to two factors: first, it is economically expensive to use soil moisture sensors. As another restraint, there may be sensor failures, most likely because of their greater quantity and measurement errors as a result of hostile environmental conditions and incorrect installation.

A support vector machine (SVM) can also be used to predict soil moisture with digital photography [24]. However, hardware costs and size may make a photography-based approach unfeasible in some cases. This paper presents another method based on soil pictures [25]. It is investigated how soil moisture can be predicted using various machine-learning methods. There is a remarkable overview of machine learning approaches presented here, but RNNs and their improvements, such as LSTMs, are not discussed. In [26] compares multiple linear regression, support vector regression, and random forest algorithms, assessing their effectiveness for soil moisture prediction [27].

3 METHODOLOGY

3.1 Network Architecture

The TelosB wireless sensor module connects soil moisture sensors, soil temperature sensors, or soil humidity sensors to the network nodes. During the measurement cycle, each node collects data about soil moisture, relative humidity, and soil temperature, which is then sent to a sink node for analysis. The following architecture is proposed to accomplish this goal.

Sleeping and waking up periodically is essential for the sensing nodes to extend the network's life. Network operations are ongoing. Most of the time, the nodes are sleeping (the sleep period lasts 26 minutes, the wake-up period lasts 4 minutes, and the time between two consecutive periods lasts 4 days). There will be a lot of sleep among the sensor nodes during the day. To initiate the transmission of a message, the destination node must also become active (wake up) when the source node becomes active (wake up). A periodic synchronization of nodes is necessary to coordinate this process, as explained in section II-C. As data is relayed from nodes further up the tree, nodes near the sink node will drain their batteries more quickly. A lack of precautions may lead to nodes near the sink node dying much earlier than those in the bottom layers. To make all the nodes' energy consumption consistent, a new SPT should be calculated, considering that a

greater number of energy-consuming nodes will be chosen as parents with a larger number of children.

Sensor networks that monitor environmental conditions, such as soil moisture, usually perform convergence casting. Data is gathered from all sensor nodes and stored at sink nodes with a tree-based routing architecture [10]. There are two types of convergent casts: aggregated data and raw data. At each hop, aggregated convergence casting uses data aggregation techniques, whereas raw data convergence casting relays sensor readings directly to the sink. Whenever parameters are highly correlated, an aggregated converge cast is used to detect them.

We designed our soil moisture sensor network not to be densely deployed, so it covers a large area with few nodes, making it an affordable solution to monitor soil moisture. In this situation, it is assumed that all sensor readings are shared equally between the sink and source nodes.

Coordinated, comprehensive, and sustained Earth observation strengthens the capacity of the international community to achieve this.

In precision agriculture, environmental conditions related to crop health can be monitored using ZigBee wireless communication [28]. Wireless connectivity is provided by IEEE 802.15.4. ZigBee was originally developed for personal area networks under the ZigBee alliance [29]. Besides its flexible network structure, it comes with a long battery life, provides mesh, star, and tree topologies, and supports multi-hop data transmission.

A prominent IoT application that uses BLE is agriculture, which is a suitable protocol for Bluetooth smart technology [29]. The device was designed specifically for IoT applications requiring low bandwidth, low latency, and short range. With BLE, you can connect unlimited devices faster, use less power, and setup is easier than with conventional Bluetooth. In its original form, 6LoWPAN is an IP-based communication protocol for IoT applications. The low bandwidth and low power consumption of 6LoWPAN also make it a low-cost network. 6LoWPAN, including star and mesh topologies, supports multiple topologies. A network adapter layer is used between the network and the MAC layer to ensure that IPv6 and IEEE 802.15.4 are interoperable [29].

3.2 Soil Moisture Sensor Based on Coils

Using the mutual inductance principle, this study proposes a sensor consisting of sensors. According to this principle, a magnetic field will be generated

whenever an electric current (EC) is applied to a powered coil (PC). Several parameters influence the magnetic field of a coil, including the number of spires (N), the diameter of the wire, and the diameter of the powered coil (PC), as well as the signal used to power the coil (including voltage and the). An electric coil's magnetic field is determined by a variety of factors, such as the number of spires (N), the diameter of the wire, and the signal used to power the coil (voltage and frequency). The (1) depicts how Ampère's circuital law determines a solenoid's magnetic flux density B as a function of the permeability of the core, N spikes, and I intensity. Although (1) applies to a free space infinite solenoid, in our case, we have a soil solenoid with relative permeability μ_r . As shown in (2), the solenoid's length l should be considered. Our soil's μ_r is modified when the moisture changes, which results in a change in its magnetic field generated. Increasing moisture will increase the soil's permeability. This will increase the magnetic field. Solenoid cores are generally placed in the centre, where the magnetic field is greatest. As a result of the powered solenoid's magnetic field being higher in the centre (with different moisture levels in different tests), the soil will be the core of our device.

$$B = \mu_0 NI, \quad (1)$$

$$B = \frac{\mu_0 \mu_r NI}{l}. \quad (2)$$

An induced magnetic field will be created if another coil is in close proximity to the PC. Mutual inductance refers to this phenomenon. Magnetic flux will be created by the lines going through the induced coil (IC) from the PC's magnetic field. According to the theory, mutual inductances are as follows to (3), the mutual inductance of two solenoids is M . A coil has an inductance $L1$ and, an inductance $L2$, and a coupling coefficient k . Equation (4) shows that $L1$, $L2$, N , l and a depend on the core $\mu_0 \mu_r$. As before, the result of mutual inductance is determined by the medium acting as a core, the soil. A soil's permeability affects its K . A perfect coupling causes K to be maximum (1), while an absence of inductive coupling causes it to be minimum (0). A value of k indicates that 100% of the lines of flux in PC will cut all the turns in the IC, so the soil is highly conductive, and the coils are perfectly symmetrical. As part of our experiments, we examined the core's permeability (the characteristics of the core). Furthermore, different prototypes (different N 1 and A) are tested for various geometries to determine how they affect performance. While coils are located in the same place in each prototype, PC and IC distances, as well as total coil lengths, differ.

$$M = k\sqrt{L1, L2}, \quad (3)$$

$$L1 = \frac{\mu_0 \mu_r NA}{l}. \quad (4)$$

3.3 Circuit Design

Figure 1 illustrates the circuit design for the soil profile moisture sensor. Circuits for sensor acquisition and control were essentially the same. Moisture acquisition circuits include annular metal electrodes, voltage-controlled oscillating circuits, amplifiers, and frequency dividing and conditioning circuits to measure soil moisture. An ATmega328p 8-bit microcontroller (MCU) is integrated with the sensor control circuit, as well as an NRF 24L01 wireless 2.4 GHz communication module and a time-sharing power supply circuit in addition to the battery powering the sensor, a solar panel supplied 7.4V output from a lithium battery.

This means that signal processors divide, condition, and shape frequency signals in order to be measured directly by MCUs, that is, frequency division and conditioning. Depending on the soil moisture content, equivalent capacitance varies. In the voltage-controlled oscillating circuit, output high frequencies are measured between 100MHz and 150MHz. The frequency divider circuit, as well as the circuit that converts high-frequency signals into low-frequency signals, is required for the microprocessor to read the measured output frequency. The original frequency signals were divided by 64 using a dual-mode pre-scaler and divided by 256 using an SN74HC393D digital chip. MB504 and SN74HC393D are typical frequency-dividing circuits that enable back-end processors to measure frequency signals directly. Texas Instruments has developed a dual four-bit binary counter with an SN74HC393D. The number of counts can reach 256. When 256 signal pulses are input, the output will flip, and the input frequency will become 256.

3.4 Network Design

During the Southern Sierra Critical Zone Observatory (CZO) (374' N, 11911' W), a 1.5-kilometer transect was surveyed along 23 locations to monitor water balance variables. There is a transition zone between rain and snow, with low elevations receiving more precipitation in the form of rain. In this sub-catchment of KREW, most precipitation falls, such as snow, in the northern part. There are 76–99% mixed-conifer forests in the catchment, mixed chaparral, and barren areas in WSN-instrumented catchments.

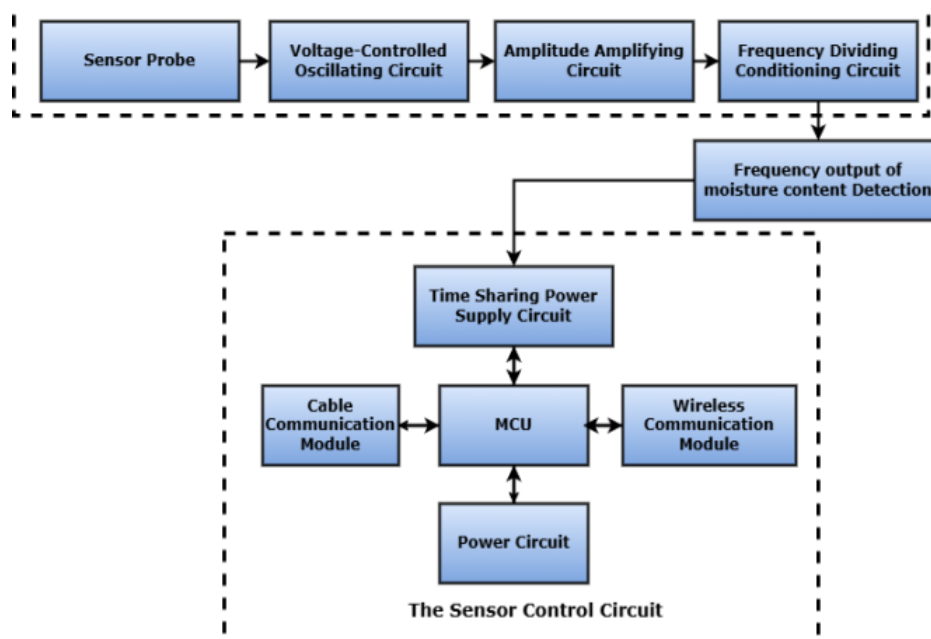


Figure 1: Principles behind the design of the soil profile moisture sensor.

As far as the network level is concerned, regular WSNs have the same function for all nodes. After measuring a parameter, each node processes it and then sends its findings to its neighbours. By sending sensed information to its neighbours, the base station or sink will be able to receive it. As opposed to passive WSNs, passive WSNs receive data from their environment and send it to a sink without any further processing. An active WSN will notify the sink and/or the management centre whenever an event occurs. It is then the responsibility of a few nodes to take action (send the information to other nodes, sense more variable information, etc.) as a result of the notification (or selection) of a few nodes.

There is a need for intelligent WSNs in many environments since passive WSNs are not useful. An event may require certain actions to be taken within a WSN to respond appropriately. A WSN used to detect fire, for example, should be able to measure a variety of variables to verify and even monitor the fire's progress. Once these variables are collected, a sensor may transmit the information to a node with a higher processing capacity to activate fire fighting mechanisms.

4 RESULTS AND DISCUSSION

Based on a 0.5 m emitter height, various deployment configurations were tested with a receiver at various

heights. X-axis numbers indicate tree positions. For all distances tested, the receiver positioned closest to the ground consistently produced higher RSSI values, proving to be the most stable configuration (Fig. 2). Trees 1 and 4 showed minor fluctuations, while trees 2 and 4 showed more significant fluctuations due to the density of foliage there. The multipath effect is likely to be responsible for these fluctuations.

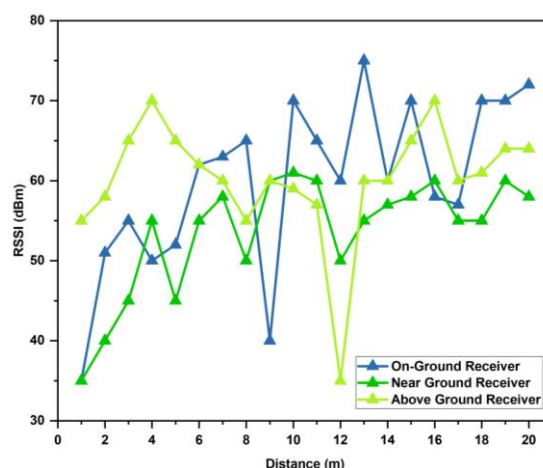


Figure 2: RSSI (dBm) versus distance (m) at 0.5 high emitter.

This Figure 3 shows the resonance frequencies (in kHz) and induced coil voltages (in mV) of the prototypes developed.

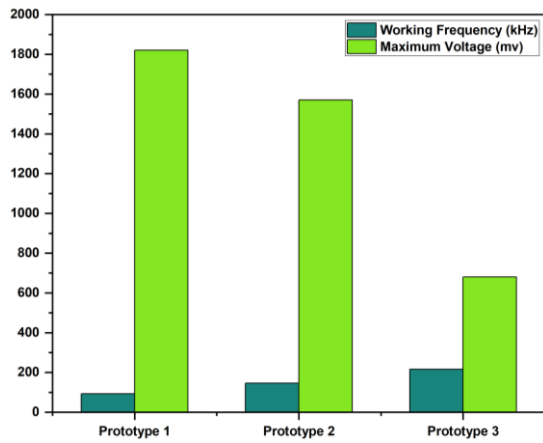


Figure 3: Moisture measurement prototypes.

Figure 4 shows the results of Prototype 1, showing 93 kHz working frequency and 1.82% maximum output voltage. As soon as the best sensor has been chosen, it is tested on several soil types to assess its versatility, aiming for maximum linearity. In the previous figure, linear behaviour is observed up to a humidity level of 18.75%, which corresponds to 750 mL of water for 4000g of sand. Sand on a beach is less sensitive than soil on a farm, so this needs to be considered when processing real-world results.

Figure 5 shows the results for a 1-meter-high emitter. A near-ground receiver best achieves performance. There is a reduction in signal quality between trees 2 and 4 in this scenario. However, after passing through a densely forested area, the near-ground receiver recovers its signal quality. There is no difference between the above-ground and below-ground deployment in areas with dense foliage, but there is a significant difference in signal quality in other areas. In addition, the on-ground receiver configuration performs poorly.

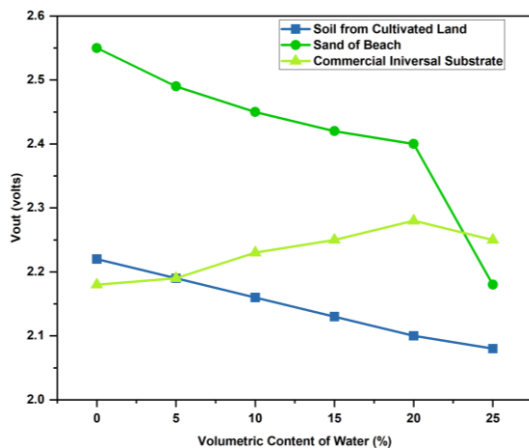


Figure 4: Various soil types and prototype 1's behaviour.

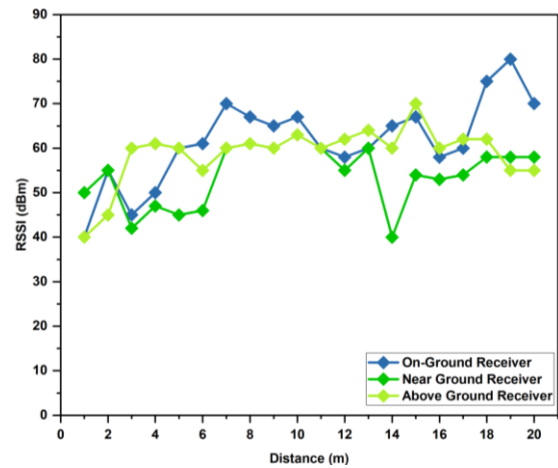


Figure 5: RSSI (dBm) versus distance (m) at emitter at height of 1 m.

Using a 1.5 m emitter height, Figure 6 illustrates the results. Almost all measurement points provide the highest level of signal quality, although there are some fluctuations between trees 2 and 3. This configuration maintains superior signal quality despite these fluctuations. In addition to near-ground receivers, above-ground receivers offer a more stable signal, but their quality still falls short. As a result of the on-ground deployment, the results were the poorest and fluctuated the most. A higher emitter height yielded a lower average signal quality than a lower emitter height, also worth mentioning.

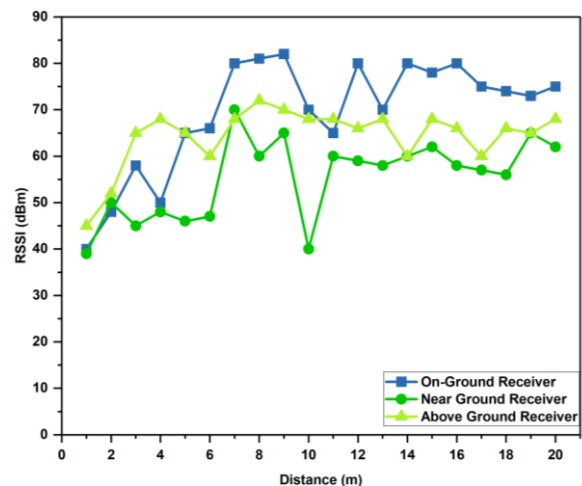


Figure 6: RSSI (dBm) versus distance (m) at 1.5-meter-high emitter.

The result at 2 meters is shown in Figure 7. All other emitter heights yield poorer signal quality than this one. Near-ground deployments deliver the best results, as with previous cases. At this emitter height,

however, all receiver configurations display similar performance. The on-ground configuration continues to be the least effective option in all scenarios.

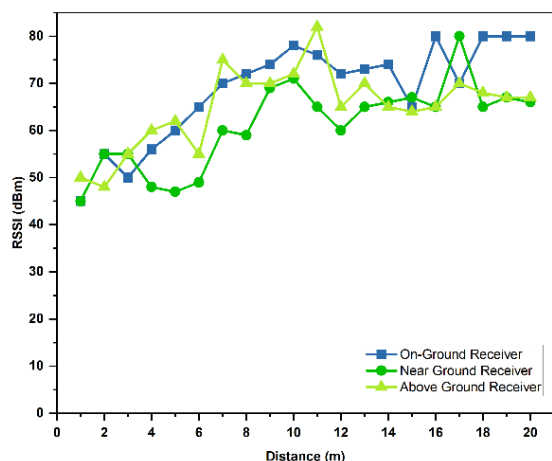


Figure 7: RSSI (dBm) versus distance (m) at emitter at height of 2 m.

5 CONCLUSIONS

As a reliable method for optimizing water use and improving agricultural productivity, WSNs have been integrated into precision farming to monitor soil moisture. The proposed system introduced in this paper, which integrates an energy-efficient architecture with periodic sleep-wake scheduling strategies and effective communication protocols, demonstrates significantly enhanced performance compared to existing agricultural monitoring systems. By deploying low-cost, highly responsive soil moisture sensors coupled with optimized routing techniques, the system achieves robust performance in large-scale agricultural environments, while ensuring minimal energy consumption and extended operational lifetimes. Future advancements should focus on scaling the system to handle more complex, heterogeneous agricultural scenarios, including diverse crop types, varied field topologies, and unpredictable environmental conditions. Additionally, enhancing sensor node reliability and durability would further ensure consistent and uninterrupted data collection. Integration of emerging technologies, such as cloud computing platforms for comprehensive data storage, analytics, and artificial intelligence for sophisticated predictive modeling, could greatly enhance real-time data interpretation, streamline decision-making processes, and ultimately improve agricultural productivity and resource efficiency in farming applications.

REFERENCES

- [1] N. Kumar, P. Rani, V. Kumar, S. V. Athawale, and D. Koundal, "THWSN: Enhanced energy-efficient clustering approach for three-tier heterogeneous wireless sensor networks," *IEEE Sens. J.*, vol. 22, no. 20, pp. 20053–20062, 2022.
- [2] N. Kumar, P. Rani, V. Kumar, P. K. Verma, and D. Koundal, "Teeech: Three-tier extended energy efficient clustering hierarchy protocol for heterogeneous wireless sensor network," *Expert Syst. Appl.*, vol. 216, p. 119448, 2023.
- [3] A. Baggio, "Wireless sensor networks in precision agriculture," in *ACM workshop on real-world wireless sensor networks (REALWSN 2005)*, Stockholm, Sweden, Citeseer, 2005, pp. 1567–1576. Accessed: Apr. 02, 2025. [Online]. Available: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=cdbc4cf0271a1049c8240ea70fd8579bdcdf0c97>.
- [4] P. Rani and M. H. Falaah, "Real-Time Congestion Control and Load Optimization in Cloud-MANETs Using Predictive Algorithms," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 66–76, 2024.
- [5] T. Grift, "The first word: the farm of the future," *Resour. Mag.*, vol. 18, no. 1, p. 1, 2011.
- [6] A. Kamilaris, F. Gao, F. X. Prenafeta-Boldu, and M. I. Ali, "Agri-IoT: A semantic framework for Internet of Things-enabled smart farming applications," in *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, Reston, VA, USA: IEEE, Dec. 2016, pp. 442–447. doi: 10.1109/WF-IoT.2016.7845467.
- [7] R. Adams, B. Hurd, S. Lenhart, and N. Leary, "Effects of global climate change on world agriculture: an interpretive review," *Clim. Res.*, vol. 11, pp. 19–30, 1998, doi: 10.3354/cr011019.
- [8] O. Westermann, W. Förch, P. Thornton, J. Körner, L. Cramer, and B. Campbell, "Scaling up agricultural interventions: Case studies of climate-smart agriculture," *Agric. Syst.*, vol. 165, pp. 283–293, Sep. 2018, doi: 10.1016/j.agry.2018.07.007.
- [9] H. Channe, S. Kothari, and D. Kadam, "Multidisciplinary model for smart agriculture using internet-of-things (IoT), sensors, cloud-computing, mobile-computing & big-data analysis," *Int J Comput. Technol. Appl.*, vol. 6, no. 3, pp. 374–382, 2015.
- [10] A. Singh et al., "Smart Traffic Monitoring Through Real-Time Moving Vehicle Detection Using Deep Learning via Aerial Images for Consumer Application," *IEEE Trans. Consum. Electron.*, vol. 70, no. 4, pp. 7302–7309, Nov. 2024, doi: 10.1109/TCE.2024.3445728.
- [11] M. Bacco et al., "Smart farming: Opportunities, challenges and technology enablers," in *2018 IoT Vertical and Topical Summit on Agriculture - Tuscany (IOT Tuscany)*, Tuscany: IEEE, May 2018, pp. 1–6. doi: 10.1109/IOT-TUSCANY.2018.8373043.
- [12] L. Lamport and C. aTime, "the Ordering of Events in a Distributed System," *o Comm.*, ACM, 1978.
- [13] F. Sivrikaya and B. Yener, "Time synchronization in sensor networks: a survey," *IEEE Netw.*, vol. 18, no. 4, pp. 45–50, Jul. 2004, doi: 10.1109/MNET.2004.1316761.

- [14] L. Tavares Bruscato, T. Heimfarth, and E. Pignaton De Freitas, "Enhancing Time Synchronization Support in Wireless Sensor Networks," *Sensors*, vol. 17, no. 12, p. 2956, Dec. 2017, doi: 10.3390/s17122956.
- [15] P. Rani and R. Sharma, "An experimental study of IEEE 802.11 n devices for vehicular networks with various propagation loss models," in *International Conference on Signal Processing and Integrated Networks*, Springer, 2022, pp. 125–135.
- [16] N. Kumar Agrawal et al., "TFL-IHOA: Three-Layer Federated Learning-Based Intelligent Hybrid Optimization Algorithm for Internet of Vehicle," *IEEE Trans. Consum. Electron.*, vol. 70, no. 3, pp. 5818–5828, Aug. 2024, doi: 10.1109/TCE.2023.3344129.
- [17] D. K. Shinghal and N. Srivastava, "Wireless sensor networks in agriculture: for potato farming," Available SSRN 3041375, 2017, Accessed: Apr. 02, 2025. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3041375.
- [18] P. Rani and R. Sharma, "Intelligent Transportation System Performance Analysis of Indoor and Outdoor Internet of Vehicle (IoV) Applications Towards 5G," *Tsinghua Sci. Technol.*, vol. 29, no. 6, pp. 1785–1795, Dec. 2024, doi: 10.26599/TST.2023.9010119.
- [19] M. R. Mohd Kassim, I. Mat, and A. N. Harun, "Wireless Sensor Network in precision agriculture application," in *2014 International Conference on Computer, Information and Telecommunication Systems (CITS)*, Jeju, South Korea: IEEE, Jul. 2014, pp. 1–5. doi: 10.1109/CITS.2014.6878963.
- [20] S. Sulaiman, A. Manut, and A. R. N. Firdaus, "Design, Fabrication and Testing of Fringing Electric Field Soil Moisture Sensor for Wireless Precision Agriculture Applications," in *2009 International Conference on Information and Multimedia Technology*, Jeju Island, Korea (South): IEEE, 2009, pp. 513–516, doi: 10.1109/ICIMT.2009.54.
- [21] P. Rani, U. C. Garjola, and H. Abbas, "A Predictive IoT and Cloud Framework for Smart Healthcare Monitoring Using Integrated Deep Learning Model," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 53–65, 2024.
- [22] R. Liao, S. Zhang, X. Zhang, M. Wang, H. Wu, and L. Zhangzhong, "Development of smart irrigation systems based on real-time soil moisture data in a greenhouse: Proof of concept," *Agric. Water Manag.*, vol. 245, p. 106632, Feb. 2021, doi: 10.1016/j.agwat.2020.106632.
- [23] H. Liu, Y. Yang, X. Wan, J. Cui, F. Zhang, and T. Cai, "Prediction of soil moisture and temperature based on deep learning," in *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Dalian, China: IEEE, Jun. 2021, pp. 46–51. doi: 10.1109/ICAICA52286.2021.9498190.
- [24] C. Saad Hajjar, C. Hajjar, M. Esta, and Y. Ghorra Chamoun, "Machine learning methods for soil moisture prediction in vineyards using digital images," *E3S Web Conf.*, vol. 167, p. 02004, 2020, doi: 10.1051/e3sconf/202016702004.
- [25] M. ElSaadani, E. Habib, A. M. Abdelhameed, and M. Bayoumi, "Assessment of a Spatiotemporal Deep Learning Approach for Soil Moisture Prediction and Filling the Gaps in Between Soil Moisture Observations," *Front. Artif. Intell.*, vol. 4, p. 636234, Mar. 2021, doi: 10.3389/frai.2021.636234.
- [26] S. Prakash, A. Sharma, and S. S. Sahu, "Soil Moisture Prediction Using Machine Learning," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, Coimbatore: IEEE, Apr. 2018, pp. 1–6, doi: 10.1109/ICICCT.2018.8473260.
- [27] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," *Comput. Electr. Eng.*, vol. 105, p. 108543, 2023.
- [28] K. Sarode and P. Chaudhari, "Zigbee based agricultural monitoring and controlling system," *Int J Eng Sci*, vol. 8, pp. 15907–15910, 2018.
- [29] S. Al-Sarawi, M. Anbar, K. Alieyan, and M. Alzubaidi, "Internet of Things (IoT) communication protocols," in *2017 8th International conference on information technology (ICIT)*, IEEE, 2017, pp. 685–690. Accessed: Apr. 04, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8079928>

Energy-Efficient Internet of Things-Based Wireless Sensor Networks through AI-Powered Resource Allocation

Wisal Jereis Alrabadi

*Faculty of Physical Education, Jadara University, 21110 Irbid, Jordan
wrbadi10@gmail.com*

Keywords: Internet of Things (IoT), Wireless Sensor Networks (WSNs), Artificial Intelligence (AI), Whale Optimization Algorithm (WOA).

Abstract: Adaptable, energy-efficient, and data-optimized environments are now achievable through the integration of wireless sensor networks (WSNs) and the Internet of Things (IoT). However, challenges such as resource constraints, high energy consumption, and inefficient communication protocols limit their effectiveness. To address these issues, this study employs the Whale Optimization Algorithm (WOA) in an AI-based approach. The proposed method optimizes resource allocation by minimizing communication costs, balancing network loads, and enhancing energy efficiency. Experimental results demonstrate significant improvements in network lifetime, throughput, and spectral efficiency. These findings highlight the effectiveness of AI-driven techniques in significantly boosting IoT network reliability, longevity, and resilience against operational challenges. By leveraging intelligent optimization, the proposed solution not only enhances WSN performance in IoT applications but also streamlines resource allocation and energy efficiency. This makes it a promising approach for next-generation smart homes, industrial IoT environments, healthcare monitoring, and other IoT-driven systems. The study underscores the substantial potential of AI-based optimization in overcoming key limitations of WSNs, paving the way for more sustainable, scalable, and efficient IoT deployments across diverse applications.

1 INTRODUCTION

With the convergence of the IoT, smart environments have become more connected, adaptive, proactive, and intelligent. Most adaptive environments are not designed with human emotions and preferences in mind, despite the fact that they are intended to facilitate people's daily lives [1]. Today, technology is mainly accessed through the senses of sight and sound. According to predictions, a full internet of senses will be established by 2025, including taste, smell, and sight. Also, by 2030, it may be possible to communicate thoughts digitally, which eliminates the need for mice and keyboards as a user interface [2]. Through the Internet of Senses, emotions and preferences will be transparently integrated without requiring direct feedback from users, resulting in a more personalized experience.

The Wireless Sensor Network (WSN) uses multiple sensors to monitor environmental parameters, including temperature, humidity, and moisture. There are a variety of sensors on the device, including electroacoustic sensors, pressure sensors,

motion sensors, image sensors, chemical sensors, weather sensors, pressure sensors, temperature sensors, and optical sensors. A wide range of applications can benefit from WSNs, including healthcare, military, defence, agriculture, and our everyday lives [3]. While WSNs have a wide range of applications, they face many challenges, including limited energy resources, computing speed, memory, and communication bandwidth, which result in a decrease in performance over time [4]. A variety of algorithms can be developed for different applications, which can be quite challenging. As part of the design of a WSN, designers should consider a number of factors, including data aggregation, clustering, routing, localization, fault detection, task scheduling, and event tracking, among others.

WSNs are multi-hop wireless networks that allow things to communicate in a monitored area, and the IoT depends heavily on them [5]. WSNs play a key role in IoT applications by enabling nodes to cooperate to provide high-functioning IoT applications [6]. IoT applications have high Quality-of-Service requirements, so deploying energy-

efficient nodes and adjusting resources in collaboration is critical in WSNs. As shown in [7], next-generation wireless networks incorporating artificial intelligence technology can improve machine learning. As a result of artificial intelligence's distributed capabilities, the wireless communication system as a whole has been surprised [8].

It has been proven that emotion-aware applications improve user experiences and system efficiency in wireless networks. In addition to cognitive radio ad-hoc networks [9], mobile cloud computing is also an example [4]. With the increasing use of emotional-aware applications, such as network personalization, services are being tailored to meet the needs and prospects of operators in continually changing environments.

Resource management is considered a key task for ensuring network connectivity because sensor nodes have inhibited power supplies. Additionally, in general, network operations consume a greater amount of energy due to communication between members. The efficient use of power can reduce the rapid exhaustion of energy in sensor nodes, improving their performance as a result. Energy efficiency, therefore, deserves a lot of attention. Many factors challenge the sensor network's functionality, such as power restrictions, insufficient processing power, etc. Sensor nodes in the network continuously gather data, which causes the battery to deplete extremely quickly. An effective strategy for managing data transmission and energy-efficient communication can significantly extend the lifetime of a network [10].

Clusters of sensor nodes are arranged hierarchically according to their type. There will be a distinction between nodes within a cluster according to their functionalities. Two-layer WSN architectures delegate nodes to collect and relay data from the surroundings to their respective heads. It is these heads that aggregate the inferred data and transmit it [11]. It is more important to schedule cluster update cycles at the appropriate time and place and to nominate sensor members to operate as the head nodes at the appropriate time. Without appropriate selection, the member nodes will experience premature energy shortfalls. In terms of the limitations of the available approaches, the cost of repeated message transmissions and the difficulty in exchanging messages through the formation of clusters are the key limitations.

2 LITERATURE REVIEW

The nearest neighbour algorithm and dynamic programming are common resource allocation algorithms in WSNs. According to the paper [12], resources should be allocated centrally. A high quality of service reduces energy consumption and enhances the utilization of broadband when it constrains resources. Using integer linear programming (ILP), a method for assigning tasks to wireless sensor networks (WSNs) in static environments is presented [13], and the nodes' overhead in communication and computation is accounted for in this calculation. In the works, the allocation of resources in the network is also based on the ILP method, but the primary consideration is the quality of the information [14]. In spite of this, the network requires efficient and dynamic resource allocation mechanisms to meet users' increasing demands. A market-based resource allocation method is used by [15] to optimize the utilization of sensor resources in competition.

A dynamic resource scheduling system has also been developed by the author [16] by applying game theory. Nodes in WSNs are assigned roles based on queuing rather than randomly. Typical WSN scenarios involve steady-state analysis to determine the optimal allocation scheme. The network remains energy-efficient and has a long lifespan when compared to other methods [17]. As part of the resource allocation process, all these methods assess the amount of energy consumed by the network. However, there is a need for different levels of quality of service for multiple publishing tasks. It is achieved by applying this concept to assign tasks to the corresponding nodes according to various quality of service requirements, thereby reducing resource utilization and improving resource efficiency.

A heterogeneous network presents a huge challenge for resource allocation due to its heterogeneous nodes. Heterogeneous WSNs can be allocated resources according to quality of service [18]. As a result of resource allocation constraints, maximizing network throughput is transformed into maximizing statistical quality of service [18]. An equitable network throughput and resource distribution scheme is proposed in the document. An optimal resource allocation scheme is presented in [19], which reduces interference and increases network throughput and speed. WSNs with heterogeneous topologies can benefit from layering protocols [20]. Author [7] illustrates how layering can effectively eliminate energy holes from networks. According to [21], the layered method can be

improved by implementing an agent that updates the network state.

Additionally, the functionality of active head nodes is interchanged in order to stabilize the entire energy usage and adapt to topological changes in the network. Researchers studied Wireless Sensor Networks to balance energy consumption using these approaches. Because the head nodes are primarily responsible for gathering, managing, and forwarding data, they consume a lot of energy. A TDMA schedule is also assigned to each node, irrespective of whether the node has sensed readings to send. With Cell LEACH, clusters are further partitioned into different cells. During the lifetime of the network, clusters and cells will persist in addition to those created during the setup phase. There will be a dynamic reorganization of clusters [22].

A regular feature of sensor networks is the adaptation of Machine Learning (ML) strategies to minimize the need for redesign [23]. By optimizing resource management and extending the lifespan of networks, machine learning can enhance both. In recent years, ML has focused more on algorithms that can be computationally feasible. In different applications, these techniques can be used for regression, classification, etc. A sensor network observes dynamic situations that rapidly change over time. Smart policies can be made with the help of the IoT, according to a recent application [24].

Wireless sensor networks can benefit greatly from machine learning. For intuitive learning, decision-making, and identifying complex models, classification and data learning are the most appropriate and widely used techniques. Graph embedding is another technique for teaching graphs, which preserves their topological properties by learning graph representations. However, the initial data is not guaranteed to be intact with these dimensionality-reduction techniques [25]. In this article [26], the author uses convolutional neural networks to obtain node features in graph structure data and to create atomic fingerprints based on the features obtained. With the original graph link sets as inputs, a full connection neural network (FCNN) [27] was trained with the WLNLM, a method proposed by the author [27] to build a link prediction model using the original graph link sets. A previously unstudied task, predicting link reliability, is addressed in this paper by applying the WLNLM. Deep learning algorithms were used to design, implement, and evaluate a link-reliability-based routing algorithm for wireless sensor networks.

Link reliability is predicted by WL-DCNN combining Weissfeiler-Lehman sub-graph labelling

with 2-branch convolutional neural networks. Moreover, the loss function was modified to account for imbalances in training samples and improve training set formation. Six public datasets demonstrate an efficient method for extracting topological features from complex networks, with up to 40% better predictions than baseline algorithms. WL-DCNN-based resilient routing algorithms ensure data integrity in sensor networks while reducing transmission paths and ensuring data integrity in vulnerable wireless links, as shown with our algorithm [28]. In addition to enhancing network durability and routing robustness, the nodes' power consumption is reduced as a result.

3 METHODOLOGY

The administration of IoT relies heavily on low latency and low energy consumption. As a first step towards achieving these goals, this paper focuses on the RA problem in IoT. A major goal of RA is to reduce communication costs and balance load. Network latency is reduced when communication costs are reduced. As a result of load balancing, network bottlenecks would be eliminated, and performance would be increased. Upon meeting these goals, this technology will be useful in many situations. This paper presents an algorithm based on particles.

Solving complex optimization problems has traditionally been done by heuristic algorithms. It is possible to devise a heuristic algorithm to find a solution to the RA problem because it is NP-complete. In this paper, we propose a new heuristic for solving the RA problem in cloud computing called WOA. Humpback whales are famous for their collective hunting technique, which inspired the whale algorithm. As a step toward solving the RA problem, we will examine WOA in detail.

3.1 Flowchart for the Proposed Algorithm

First, the 'whale-create' function is used to create whales. Scheduling problems are solved randomly by whales. Fitness analyses indicate that the BS is currently the optimal whale based on its fitness. RA patterns of whales are calculated based on the total communication cost. The following is an explanation of this function, which is titled 'fitness'. Whales begin moving after this step. A new x, X, Y, j, F value is calculated for each whale. In this Equation, X and Y are constants. Numbers are descending from $x \in$

[2.0]. There are two random numbers, $F \in [2.0]$ and $j \in [2.0]$. In the following sections, we explain how these numbers are used.

Distance functions play an important role in whale algorithms. Due to the fact that whale algorithms were designed for continuous problems, RA is discrete. Based on the algorithm, this function calculates distance. There are three functions involved in controlling whale motion. We will start by reducing the distance between the best whale and the current whale using the shrinking function. In addition, there's a "spiral" function, in which the whales rotate around the best whale. Furthermore, whales are known to approach random whales as part of their "search prey" behaviour.

3.2 Whale Creating

In this algorithm, whales represent solutions to RA. A node is a gateway or a resource in the described problem. There are many gateways, and each gateway receives information from a certain number of resources. The cloud's complete flow of information is created by connecting gateways based on a certain topology. Due to data transmission costs, gateways are connected by spanning trees or rings with the fewest edges. It creates two sets of communications for each whale: a set of edges representing the connections between gateways and a set of edges indicating the resources that are allocated to each gateway. It consists of an array w with a size $k + n$, where k is the number of gateways and n is the number of resources. As shown in array $w(w[1.k])$, the first k entries, gateways are related. In this case, $w[i] = j$ indicates that gateway $j(i, j)$ receives information from gateway $w[i] = j$. In array $w(w[k + 1.n])$ The second n entries represent each resource's gateway. Resources $F - k$ and q are connected through gateway q , as indicated by $w[F] = q$.

3.3 Distance Function

In this algorithm, the distance function is used to calculate how far two whales are from each other, which is an important function to note. Whale algorithms are discrete algorithms, while RA problems are continuous ones, so distance needs to be redefined. Whale algorithms work based on distances. Using the "shrinking" function, we will reduce the distance between the best whale and the current whale in the first step.

Each whale has $k + n$ edges, similar to the RA graph. A gateway has k resources, and a resource has

n gateways. Therefore, we find that zero is the minimum distance, while $k + n$ is the maximum distance.

3.4 Spiral Function

There is a random number p in the interval $[0, 1]$ that determines the type of whale motion for each whale. Once this number exceeds half, whale motion is made by using spiral functions, meaning the best whale is spiralled around in a circle. A new definition of spiral is needed for the RA problem.

3.5 Shrinking Function

Every time the algorithm is run, p is generated at random. In the case of a number less than 0.5, it is recommended to consider another random number, A . A less than 1 indicates that the shrinking function should be used. A whale's current behaviour is due to this function, which urges it to seek out the best whale or prey. While spirals revolve around their prey, shrinking functions move directly and faster towards them. As a result, it would be helpful if the search prey function was executed more frequently prior to the algorithm's conclusion.

Whales and prey are separated by shrinking functions based on the position of the whales and their prey. Based on the distance calculation, a percentage of non-equal entries moves towards the best whale by 50%.

3.6 Search Prey Function

The generation of new solutions can help algorithms avoid getting stuck in local optimums. To accomplish this, we use the search prey function. This function should be performed more often at the beginning. As a result, more options will be available at the beginning of the exploitation phase and during optimization. Using the Search prey function, a whale is directed to a randomly chosen prey animal. The distance between whales determines the method of moving between them, as described below.

3.7 Whale Optimization Algorithm (WOA)

An algorithm based on the Humpback whales meta-heuristics is called the WOA [29]. A random population of whales is produced in this algorithm before optimization begins. This method involves whales finding (optimal) prey locations and encircling or using bubble nets to append them

(optimize). Humpback whales improve their current locations by following this method [29]:

$$d = |X \odot Z^*(t) - Z(t)|, \quad (1)$$

$$Z(t+1) = Z^*(t) - X \odot d. \quad (2)$$

This example uses d to represent the distance between prey $(t)^*$ and whale $Z(t) \odot$, while t opinions for the current iteration. In the following definition, A and C represent constant vectors.

$$X = 2x \odot r - x, \quad (3)$$

$$Y = 2r. \quad (4)$$

Each time r is iterated, Z is a random vector with indexes between $[0,1]$, with each index corresponding to the random number between $[0,1]$

Both methods can be used for bubble-net analysis. There are two ways to shrink encompassing: decreasing (3) value of a and decreasing A 's value. Humpback whales following their prey also use helix-shaped movements to update their position:

$$Z(t+1) = d' \odot e^{bl} \odot \cos(2\pi l) + Z^*(t). \quad (5)$$

A whale's distance from its prey is defined by $Z' = |Z^*(t) - Z(t)|$, while b describes the shape of the logarithmic spiral and J a randomly generated value $\in [-1.1]$.

$$Z(t+1) = \begin{cases} Z^*(t) - X \odot d & \text{if } F < 0.5 \\ D' \odot e^{bl} \odot \cos(2\pi l) + Z^*(t) & \text{if } F \geq 0.5 \end{cases} \quad (6)$$

Does $F \in [0,1]$ represent a random value that determines a determining the location of whales, either using the shrinking encompassing method or the spiral approach?

A Humpback whale explores for prey randomly in its discovery phase. Whales are positioned randomly rather than with the best search agents.

$$d = |Y \odot Z_{rand} - Z(t)|, \quad (7)$$

$$Z(t+1) = |Z_{rand} - X \odot d|. \quad (8)$$

How many positions are there in the current population where Z_{rand} is it chosen randomly? In the first algorithm, the WOA's structure is depicted as a whole.

4 RESULTS AND DISCUSSION

The proposed deep learning routing protocol was evaluated using MATLAB simulations. In the experiment, 200 to 1000 nodes were distributed over an area of 1000 x 1000 m. GEEC, TTDFP, and EADCR have been compared to WODNN HMOFA, as well as five other existing algorithms. In this analysis, a variety of performance metrics are evaluated, including throughput, energy efficiency, Quality of Service (QoS), and spectral efficiency. Figure 1 demonstrates the superior performance of 98% throughput, Figure 2 shows the 97% energy efficiency, Figure 3 illustrates the 77% QoS, Figure 4 demonstrates the 87% spectral efficiency, and Figure 5 shows the 93% network lifetime using the proposed technique. Based on these results, the proposed method was able to achieve optimized energy efficiency and data handling, confirming its efficiency in optimizing

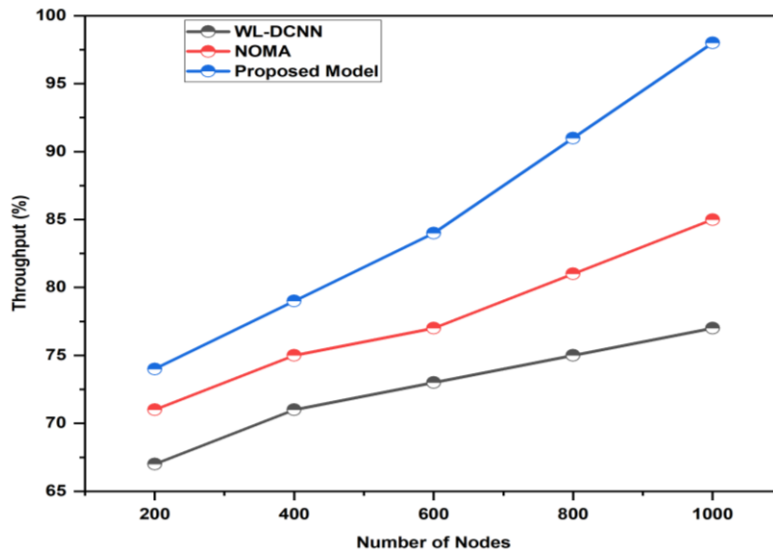


Figure 1: A comparison of throughputs versus number of nodes.

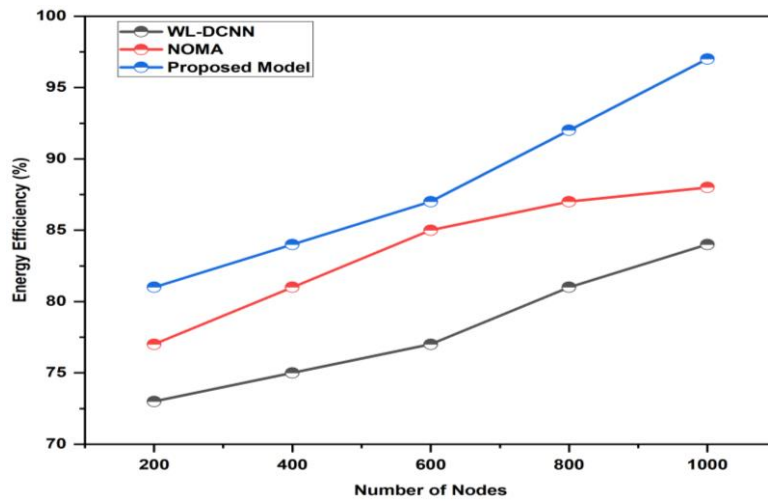


Figure 2: Energy Efficiency (%) versus number of nodes.

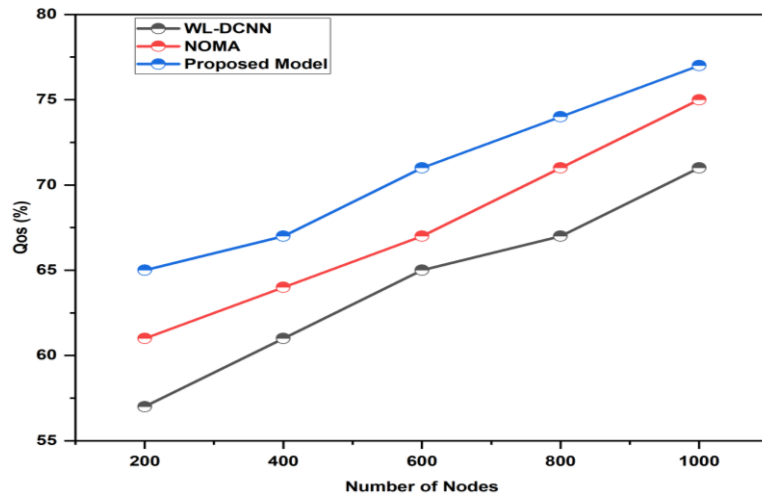


Figure 3: Quality of service versus number of nodes.

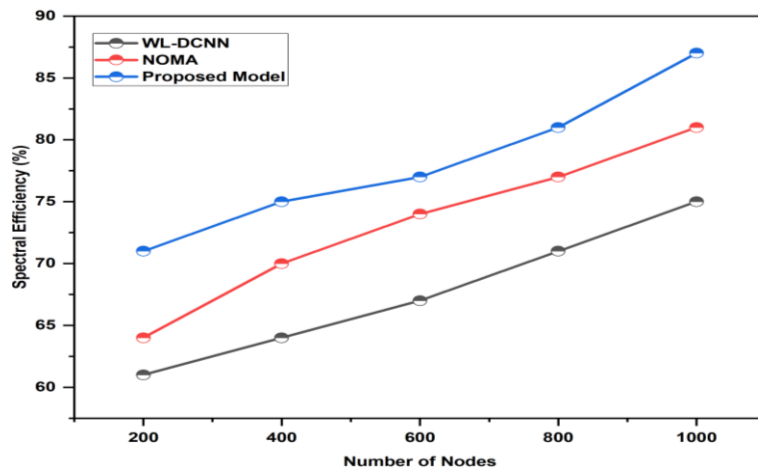


Figure 4: A comparison of spectral efficiency versus number of nodes.

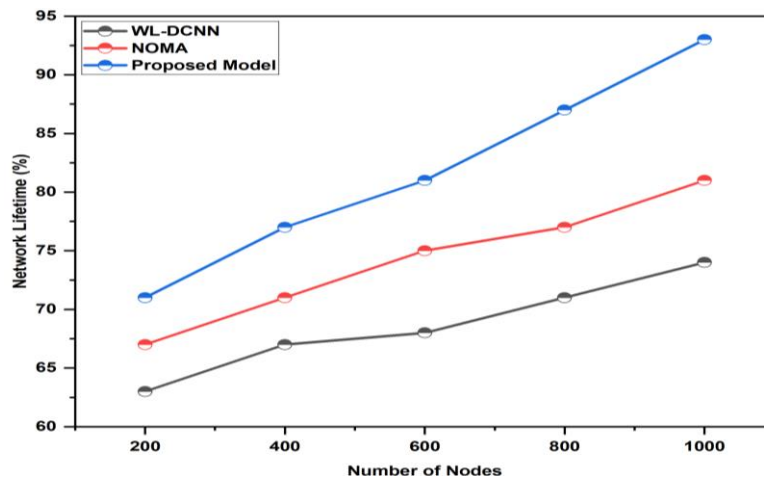


Figure 5: Analyzing the network lifetime versus number of nodes.

5 CONCLUSIONS

With the help of the Whale Optimization Algorithm (WOA), we have proposed an AI-driven approach to resource management in IoT-enabled WSNs that addresses critical challenges. A resource management strategy like this significantly improves energy efficiency, reduces communication latency, and enhances overall network performance and stability. According to detailed experimental evaluations and performance analyses, the proposed method demonstrates clear superiority in ensuring balanced load distribution across sensor nodes, maximizing energy utilization efficiency, and notably extending network lifetimes. These results highlight how IoT networks can substantially benefit from integrating advanced AI-based algorithms, thus enabling them to reliably meet the evolving demands of modern smart environments, including smart healthcare, agriculture, and industrial automation systems. Future research endeavors could further enhance the scalability, robustness, and performance of heterogeneous IoT networks by effectively integrating state-of-the-art deep-learning models with sophisticated optimization techniques, thus broadening potential applications and real-world effectiveness.

REFERENCES

- [1] N. Kumar, P. Rani, V. Kumar, P. K. Verma, and D. Koundal, "Teech: Three-tier extended energy efficient clustering hierarchy protocol for heterogeneous wireless sensor network," *Expert Syst. Appl.*, vol. 216, p. 119448, 2023.
- [2] E. ConsumerLab, "10 Hot Consumer Trends 2030," 2019, [Online]. Available: <https://artillery.co/wp-content/uploads/2019/12/10-Consumer-Tech-Trends-Ericsson.pdf>, [Accessed: Dec. 28, 2024].
- [3] N. Kumar, P. Rani, V. Kumar, S. V. Athawale, and D. Koundal, "THWSN: Enhanced energy-efficient clustering approach for three-tier heterogeneous wireless sensor networks," *IEEE Sens. J.*, vol. 22, no. 20, pp. 20053-20062, 2022.
- [4] M. Chen, Y. Zhang, Y. Li, S. Mao, and V. C. Leung, "EMC: Emotion-aware mobile cloud computing in 5G," *IEEE Netw.*, vol. 29, no. 2, pp. 32-38, 2015.
- [5] G. Kalnoor and J. Agarkhed, "Artificial Intelligence-Based Technique for Intrusion Detection in Wireless Sensor Networks," in S. S. Dash et al. (eds), *Artificial Intelligence and Evolutionary Computations in Engineering Systems*, in *Advances in Intelligent Systems and Computing*, vol. 517, Singapore: Springer Singapore, 2017, pp. 835-845, [Online]. Available: https://doi.org/10.1007/978-981-10-3174-8_69.
- [6] T. Van Nguyen, Y. Jeong, H. Shin, and M. Z. Win, "Machine learning for wideband localization," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 7, pp. 1357-1380, 2015.
- [7] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wirel. Commun.*, vol. 24, no. 2, pp. 98-105, 2016.
- [8] P. Rani, U. C. Garjola, and H. Abbas, "A Predictive IoT and Cloud Framework for Smart Healthcare Monitoring Using Integrated Deep Learning Model," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 53-65, 2024.
- [9] Hayder Jalo and Mohsen Heydarian, "A Hybrid Technique Based on RF-PCA and ANN for Detecting DDoS Attacks IoT", *IJDS*, vol. 1, no. 1, pp. 28-41, Jun. 2024.
- [10] M. S. Gharajeh, "Avoidance of the energy hole in wireless sensor networks using a layered-based routing tree," *Int. J. Syst. Control Commun.*, vol. 7, no. 2, pp. 116-131, 2016.
- [11] Ayas Talib Mohammad and J. Parchami, "Improving Diabetic Patients Monitoring System Using (NCA-CNN) Algorithm based on IoT", *JT*, vol. 6, no. 2, pp. 9-17, Jun. 2024.

- [12] Y. Zhang, Y. Zhu, F. Yan, Z. Li, and L. Shen, "Semidefinite programming based resource allocation for energy consumption minimization in software defined wireless sensor networks," in 2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), IEEE, 2016, pp. 1-6, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/7794902> /. [Accessed: Dec. 28, 2024].
- [13] V. M. Raee, D. Naboulsi, and R. Glitho, "Energy efficient task assignment in virtualized wireless sensor networks," in 2018 IEEE Symposium on Computers and Communications (ISCC), IEEE, 2018, pp. 976-979, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8538632> /. [Accessed: Dec. 28, 2024].
- [14] S. Hariharan, C. Bisdikian, L. M. Kaplan, and T. Pham, "Efficient Solutions Framework for Optimal Multitask Resource Assignments for Data Fusion in Wireless Sensor Networks," ACM Trans. Sens. Netw., vol. 10, no. 3, pp. 1-26, Apr. 2014, [Online]. Available: <https://doi.org/10.1145/2594768>.
- [15] A. T. Zimmerman, J. P. Lynch, and F. T. Ferrese, "Market-based resource allocation for distributed data processing in wireless sensor networks," ACM Trans. Embed. Comput. Syst., vol. 12, no. 3, pp. 1-28, Mar. 2013, [Online]. Available: <https://doi.org/10.1145/2442116.2442134>.
- [16] W. Guo, Y. Chen, and G. Chen, "Dynamic Task Scheduling Strategy with Game Theory in Wireless Sensor Networks," New Math. Nat. Comput., vol. 10, no. 3, pp. 211-224, Nov. 2014, [Online]. Available: <https://doi.org/10.1142/S1793005714500124>.
- [17] M. Okhovvat, M. R. Kangavari, and H. Zarrabi, "An analytical task assignment model in wireless sensor actor networks," in 2017 IEEE 7th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), IEEE, 2017, pp. 195-199, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8210626> /. [Accessed: Dec. 28, 2024].
- [18] Y. Gao, W. Cheng, H. Zhang, and Z. Li, "Optimal resource allocation with heterogeneous QoS provisioning for wireless powered sensor networks," in GLOBECOM 2017-2017 IEEE Global Communications Conference, IEEE, 2017, pp. 1-6, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8253949> /. [Accessed: Dec. 28, 2024].
- [19] M. Ali, S. Qaisar, M. Naeem, J. J. P. C. Rodrigues, and F. Qamar, "Resource allocation for licensed and unlicensed spectrum in 5G heterogeneous networks," Trans. Emerg. Telecommun. Technol., vol. 29, no. 10, p. e3299, Oct. 2018, [Online]. Available: <https://doi.org/10.1002/ett.3299>.
- [20] G. Han, X. Jiang, A. Qian, J. J. P. C. Rodrigues, and L. Cheng, "A Comparative Study of Routing Protocols of Heterogeneous Wireless Sensor Networks," Sci. World J., vol. 2014, pp. 1-11, 2014, [Online]. Available: <https://doi.org/10.1155/2014/415415>.
- [21] G. M. K. Dheepan and K. V. Anusuya, "A Status Update Agent for Cross Layered Approach to Increase Energy Efficiency in Wireless Sensor Networks," Sens. Lett., vol. 13, no. 12, pp. 1097-1105, Dec. 2015, [Online]. Available: <https://doi.org/10.1166/sl.2015.3639>.
- [22] P. Rani and M. H. Falaah, "Real-Time Congestion Control and Load Optimization in Cloud-MANETs Using Predictive Algorithms," NJF Intell. Eng. J., vol. 1, no. 1, pp. 66-76, 2024.
- [23] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," Comput. Electr. Eng., vol. 105, p. 108543, 2023.
- [24] M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," IEEE Commun. Surv. Tutor., vol. 16, no. 4, pp. 1996-2018, 2014.
- [25] P. Goyal and E. Ferrara, "Graph embedding techniques, applications, and performance: A survey," Knowl.-Based Syst., vol. 151, pp. 78-94, Jul. 2018, [Online]. Available: <https://doi.org/10.1016/j.knosys.2018.03.022>.
- [26] D. K. Duvenaud et al., "Convolutional networks on graphs for learning molecular fingerprints," in Adv. Neural Inf. Process. Syst., vol. 28, 2015, [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/hash/f9be311e65d81a9ad8150a60844bb94c-Abstract.html. [Accessed: Dec. 28, 2024].
- [27] M. Zhang and Y. Chen, "Weisfeiler-Lehman Neural Machine for Link Prediction," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax NS Canada: ACM, Aug. 2017, pp. 575-583, [Online]. Available: <https://doi.org/10.1145/3097983.3097996>.
- [28] N. Kumar Agrawal et al., "TFL-IHOA: Three-Layer Federated Learning-Based Intelligent Hybrid Optimization Algorithm for Internet of Vehicle," IEEE Trans. Consum. Electron., vol. 70, no. 3, pp. 5818-5828, Aug. 2024, [Online]. Available: <https://doi.org/10.1109/TCE.2023.3344129>.
- [29] S. Mirjalili and A. Lewis, "The Whale Optimization Algorithm," Adv. Eng. Softw., vol. 95, pp. 51-67, May 2016, [Online]. Available: <https://doi.org/10.1016/j.advengsoft.2016.01.008>.

Advancements and Cybersecurity Strategies in V2X Communication and Transport System for Electric and Connected Vehicles Using IoT and AI

Hema Muniyappan¹, Pavithra Sankarappan¹, La Ode Muhram², Muhamad Farih³,
Aarthi Saravanaperumal⁴, Shubham Sharma⁵ and Arkan Adnan Imran⁶

¹ Department of Computer Science and Engineering, Chennai Institute of Technology, 600069 Chennai, India

² Department of Law Science Study Program, Universitas Sulawesi Tenggara, 93116 Kendari, Indonesia

³ Department of Digital Business Study Program, Universitas Sulawesi Tenggara, 93116 Kendari, Indonesia

⁴ Center for Advanced Multidisciplinary Research and Innovation, Chennai Institute of Technology,
600069 Chennai, India

⁵ Center for Research Impact and Outcome, Chitkara University Institute of Engineering and Technology,
Chitkara University, 140401 Punjab, India

⁶ Department of Computer Engineering Techniques, Dijlah University College, 10052 Baghdad, Iraq
Hemam.cse2022@citchennai.net, pavithras@citchennai.net, laodemuhram@un-sultra.ac.id, muhammadfarih@un-sultra.ac.id, , aarthi.chem@citchennai.net, shubham543sharma@gmail.com, arkan.adnan@duc.edu.iq

Keywords: Vehicle-to-Everything, Automated Vehicle, V2X Communication, Bank Digital Transformation, Cloud Computing, Cyberattacks, V2X Security.

Abstract: Vehicle-to-Everything (V2X) communication has emerged as one of the transformative technologies that the automobile sector needs: it concerns with present day exchange of information between two or more vehicles, facilities, and anything that aims at enhancing safety on the roads, flow of traffic and driving in general. According to the specifications provided by 5G, edge computing, and IoT in V2X systems, the EVs gain the development opportunity of collision avoidance, traffic optimization, and energy management. On the other hand, these developments brought forward cybersecurity weaknesses: and the jamming and spoofing attacks and DDoS attacks with which reliability and safety of V2X systems is threatened. In the present paper, the particular attention will be given by the author to outlining the newest technology trends and cyber security frameworks responses to the threats found in defending the channels in the V2X system– current and future. It lists relatively broad topics as the following general topics: authentication schemes applicable for V2X networks, encryption techniques that can be used in V2X communication link, and IDS system that can be implemented for V2X networks. More importantly, it demonstrates how AI can enhance the security of the vehicular network and realistic legislation may construct a foundation for V2X cyber security. This work explains research toward developing a secure V2X environment for readiness in integrated transport system with safer, efficient and sustainable methods.

1 INTRODUCTION

Smart cities use ICTs in economy, environment, and social spheres, solving urbanization issues with new technologies [1]. Smart city services secured \$80 billion in 2018 compared to \$67 billion in 2017 and expected to reach \$135 billion in 2021 (IDC Research). Main investments concern ITC for transport, and environmentally friendly transport systems, though regional. The US and Europe are most concerned with transport while China is

concerned with watching citizens through cameras and Japan – with environmental control. The US and China are the biggest [2] spendings with \$22 billion and \$21 billion respectively in context with the Vehicle-to-Everything Communication for Electric and Connected Vehicles. In accordance with the smart cities with the transformative contemporary technologies in relation with the enhancing road safety and transport system is centered, the IoT solutions are predominant in smart city solutions and the overall global IoT expenditure is estimated to touch 123.8 Billion US dollars in the year 2021.

IoT extends value and disrupts the conventional industries especially the transport industry due to enhanced detection, Artificial Intelligence, and data analysis. The Spanish Recommendations for IoT show this growth; connected objects are predicted to reach 5 million in 2018 and 8 million by 2022. Vehicular communication, which includes V2X (Vehicle-to-Everything), I2X (Infrastructure-to-Everything), and P2X (People-to-Everything), is for smart mobility [3] though much work has been done on V2X, new communications such as I2X and P2X have not found standards. Present work examines industrial and scientific developments, bibliometrics and future targets from the major scientific databases with regard to further development of essentials for smart cities as the global population increases the demand put on them [4]. Figure 1 shows that the advantages of the smart citizens that leverage ICTs to enhance economic, environmental, and social aspects, addressing urbanization challenges through innovative technologies.

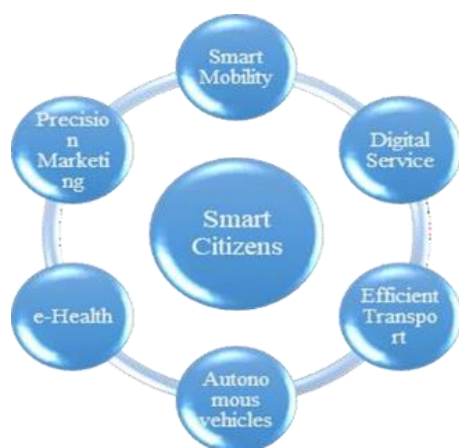


Figure 1: Advantages of smart citizens that influence ICTs to strengthen the technologies including the transport industry by using AI-IoT.

2 ADVANCES IN V2X COMMUNICATION TECHNOLOGIES

The progress in Vehicle-to-Everything (V2X) technology, focusing specifically on the future of transportation system and collaboration between vehicles, infrastructure [5], and their users with referenced real-time data sharing. This is highlighted by early research into banking and cloud security

highlighting its early-stage evolution and covering both risks such as phishing and malware as well as newer frameworks [6] such as MFA and hybrid cloud security. In the mid-2010s, there is blockchain for secure transaction and CASB for security management of the cloud vulnerabilities; in the late 2010s, it is AI-enabled fraud detection and the security concept zero-trust [7]. Although comparatively small, the trends emerging point to a future where IoT, quantum-safe cryptography, and biometric behaviors will be dominant and will require adaptable solutions [8].

Additionally, the study identifies areas of future development like the combination of AI and IoT for infrastructure in smart cities, quantum-proof cryptography for emerging banking applications. V2X security protocols that use such elements like encryption and etc. are critical to solving cybersecurity issues such as spoofing and DDoS attacks. Highlighted in the survey, V2X system protection requires effective frameworks and international cooperation; problems call for legislation, AI-based anomaly recognition, and blockchain-based data accuracy. The goal of this work is therefore to create a safe environment for V2X that leads to improved, safer, efficient and sustainable transport systems while at the same time considering emerging new trends in cybersecurity risks [9].

2.1 Unaddressed Challenges in V2X Technology

While AI and IoT are pivotal to smart cities, research lacks practical solutions for their secure integration. Issues like real-time data processing, scalability, and privacy remain underexplored, particularly in urban infrastructures. As hybrid and multi-cloud models expand in smart cities and banking, detailed insights into their unique security challenges are missing. Critical issues such as secure data synchronization, distributed access control, and interaction between public and private clouds need further study. Although promising, behavioral biometrics face limited research on their scalability and adoption in banking systems [10]. Additionally, ransomware attacks in cloud-hosted systems lack proactive and sector-specific detection frameworks. Despite growing awareness of quantum threats, there is insufficient research on transitioning banking and IoT systems to quantum-resistant cryptographic methods [11].

2.2 Applications of AI-IoT Technologies for Road Safety and Intelligent Transport in Smart Cities

The methodology of this review is to stand back and aim to look how key gaps in the literature can be improved when discussing cybersecurity issues and innovations in smart cities and cloud banking. This research investigates the application of AI-IoT technology in smart cities with particular focus on technical and privacy concerns in practice-based settings. Its purpose is to present the challenges related to security and data management in the systems based on hybrids and multi-clouds in both the urban and financial environments for building complex multimodal infrastructures [12]. Further attention will be paid to the use of quantum immunity in banking and Iot networks and possible transition from traditional cryptographic system to quantum immune. In addition to this, the review shall assess open proactive measures against ransomware in cloud infrastructure and use of behavioral biometrics for banking user identification. Pointing out these key challenges and the related actionable solutions, the present review should act as the basis for further research and enhanced security approaches to these very remarkable domains [13].

3 MATERIALS AND METHODOLOGY

Banking and cloud computing digitization creates highly advanced dangers with strong persisting threats, ransomware attacks, and insider attacks on the banking sector. In the banking industry, APTs target financial transactions [14]. On the other hand, ransomware attacks occurs frequently against cloud-hosted applications while exploiting shared environment vulnerabilities. Such threats point toward having a proactive defense system specially customized for each sector.

Figure 2 shows that the different types of threats of the evolving cybersecurity threats. Digitalization of banking, cloud computing results into high protection of advanced, continuous threats, ransomware, insider's threats to banking. In the banking industry the threat is done towards the financial transaction by APTs [15]. On the other hand, ransomware attacks are prevalent on the cloud hosted applications due to the reasons they exploit the shared environment vulnerabilities. Of course, such threats indicate that there should be a preemptive defense

strategy and this further should be industry specific. The use of algorithms for the attribute recognition of patterns and comparison of the obtained data with reference information contributes to the more efficient search for signs of risk and the use of outsourcing with cybersecurity companies for penetration Testing no more than once a year to ensure that hidden vulnerabilities of cloud infrastructures are revealed. Making systems more secure is the key reason preparing them for premium challenges by using protocols such as quantum-resistant algorithms [16]. The data transmission and storage, where a multilayer security system-including, endpoint security and Secure Application Program Interface-was attempted to lower risks. This makes the security policy always check and updated, thus, be strong enough to manage with the new form of attack of by hackers.

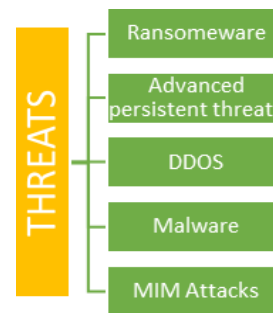


Figure 2: Different types of threats, including cybersecurity threats.

3.1 Enhanced Defense Protocols for V2X Communication in Connected and Autonomous Vehicles Using AI-IoT Technologies

V2X communication systems herald positive change in connected and autonomous vehicle systems enhanced by advanced technologies. V2X protocols incorporate 5G in the communication between vehicles and infrastructures and IoT for edge computing and integrated information sharing, primarily for the safety of life, property, and better traffic and energy management. V2X communication systems are introducing positive development and innovation in connected and autonomous applications enabling sophisticated technology. 5G and V2X, V2X hit control operate 5G [7] in the vehicle and infrastructure communication, and IoT governs the edge computing and the coordinated data exchange, mostly for safety of lives and property, efficient traffic, and energy usage. More effective encryption

and stronger privacy preserving pseudonym schemes also strengthen the defenses, which gets the communication secured and users' identities safeguarded. The integration of these protocols evidences that V2X systems can enhance safer, smarter and more sustainable transportation Systems is shown in the Table 1.

The existing V2X communication systems have transformed the fields of connected and autonomous vehicles with the help of technologies such as 5G and IoT to allow data transfer. These protocols had contributed to the improvement of traffic flow, energy usage and safety. Secure communications and privacy protection of user identity entail the use of sound encryption regimes, pseudonymity, and privacy-preserving measures. Also, AI and machine learning are critical in the V2X since they are used in alarming of anomalies or prognostication of likely system failure. Some of the benefits of predictive maintenance using AI include: cutting on time that could have been spent repairing or replacing components that are down. Blockchain core concepts [8], such as decentralization and anonymity, improve the reliability of data, as well as decrease fraudulent scenarios, thanks to an effective record of data partitions. The integration of 5G network slicing assures separated pathways and robust shields against cybersecurity threats to specific services. Everyone knows that firmware updates and patches are made periodically to protect systems from new risks. This allows automakers, regulators and tech providers to standardize the security platforms and entwine the defenses creating healthy and secure integration that protected the constantly transforming transportation systems.

3.2 Regulatory Compliance and Identity Management in V2X Systems

There is strong security that identifies and protects the vehicles from being accessed by unauthorized persons:

- Intrusion detection systems or simply IDS. These systems keep observing the vehicular communication for irregularities and threat.
- Regulatory compliance. GDPR compliance, as well as the use of the Zero Trust model and standards that are specific to V2X, allow for the alignment of cross-border security. These advances bear to the development of robust V2X systems to support safer, smarter, efficient, and sustainable transportation systems.
- Regulatory compliance in cybersecurity.

Protocols like GDPR, PCI DSS, and Zero Trust security model are necessary for protection of such data. Cloud providers and banks are protected for data control and transfer across borders by the GDPR regulation. Sequential verification of user identities used in Zero Trust strategies is essential for combating insider threats in contexts of hybrid cloud. Furthermore, it also confirms with the compliance and operational security in the cloud banking and also integrates the strong identity access management system. Cybersecurity compliance is vital for cushioning Hybrid Cloud-Specific in addition to banking security against risks, it makes sure security of vital information. Being in line with international standards like ISO 27001 makes it easy to combat cross border issues since measures are standard worldwide. Not only compliance with these standards can be observed in result of regular audits, but also potential threats can be identified on time.

Table 1: Key features and technologies in V2x communication.

Technology	Application	Benefits	References
AI and Machine Learning	Anomaly detection, collision avoiding	It enhances safety and the possibility of the right route planning.	[1] [8] [15] [9] [2]
Blockchain	Ensuring security for the exchange of vehicular data	The work of this layer is to maintain the accuracy of the data as well as reduce the risk of fraud.	[3] [10] [7]
Advanced Encryption	ECC and post – quantum security is Among the cryptographic protocols that need to be reconsidered in new communication networks	It improves the communication security.	[5] [6] [10] [11] [14] [2]
Privacy Mechanisms	Pseudonym- based frameworks	Protecting users' identification and their data	[12] [7]

Dissemination of complicated compliance frameworks should therefore be elastic in the face of constant cyber threats in order to improve organizational security. Industry, cloud providers together with the financial sector and the regulatory board develop synergy strategies to enhance compliance and security. In the same way, industry-specific frameworks in handling incidents reduce the loss of time due to the prompt restoration from the breach. Thorough training of the users makes awareness on TOP threats, and prevents insider threats, and strengthens identity and access management initiatives. These strategies in synergy work to align compliance measures with operational security [9] opening the doors to safer, smarter, and more efficient systems. To expand on the level of regulatory compliance, new and higher sophisticated risk management instruments are implemented into an organization that ensures that the compliance assessment is conducted concurrently with work. Both tools are particularly useful for understanding the existing security gaps and improving the processes of remediation. It is possible to safely achieve compliance by implementing SASE architecture in cloud-first designs that provide comprehensive network security. Moreover, work in accordance with the concept of an immutable audit trail enhances accountability that, in turn, contribute to better tracking and reporting of compliance activities. Working closely with the regulatory authorities and being a member of the compliance-based consortiums help organizations to be ahead in the current emerging policies. In addition, using various forms of encryption depending on the conditions in a country when transferring data across borders meets different needs. AI and machine learning are also slowly being used to fully automate compliance checks so that these can be checked in real time. Thus, the last element of the framework is promoting the transparency with stakeholders and issuing regular compliance reports that would create trust and show the continuing work on the cybersecurity best practices.



Figure 3: Types of V2X communication.

Figure 3 shows that the types of V2X communication that is used in the Compliance frameworks like GDPR, PCI DSS, and Zero Trust models are vital for securing sensitive data. It integrates robust identity access management systems ensures compliance and operational security in cloud banking for adjacent cooperation and energy efficient methods like platooning by giving additional information to the self-driving vehicle [2]. V2X is also useful to emergency response systems by providing priority to emergency vehicles and to notify other drivers to facilitate time reduction and safe operation. Besides, it has enhanced the users' experience, real-time information in terms of parking space and road status, and interacting with smart city systems in order to make the right pre-planned travel. This technology helps to encourage better, safer, and efficient means of moving people from one place to another as shown in the Table 2.

Table 2: Different V2X communication types.

V2X connectivity type	Description	References
V2I Vehicle to Infrastructure	Cars and road infrastructure (traffic lights, road signs, parking lots, etc.)	[1], [5], [8], [9], [12]
V2P Vehicle to Pedestrian	Cars and pedestrians	[10], [15]
V2V Vehicle to Vehicle	Vehicle(car) to vehicle (car)	[8], [11]
V2N Vehicle to Network	Vehicle (car) and networks	[1], [4], [5]

Compliance is crucial to the matters of data confidentiality and organizational security, most/values-of-all especially in case of the usage of hybrid cloud services and the banking sector. The setup of GDPR and implementations such as the PCI DSS provide strong systems for controlling data flow across borders establishing increased relevance due to the increase in international cooperation. Sequential verification processes in the recommendations of the Zero Trust security models are successful in identifying insider threats.

Additionally, compliance measures developed for the specific sector for the financial and transport sector tackle specific needs. Security measures like the multi-factor authentication do improve trust within the user base. Automized compliance check guarantees that the program continues to observe on regulations even as systems change. Technologies such as homomorphic encryption enable computation on data in a manner that legislation compliance and

anonymity of patients can be maintained. Contracts facilitate and promote cooperation and the functioning of multiple systems that connect countries. Also, once adopted, the internal and external audits together with adaptive security measures, provide a flexible work environment that addresses the ever-evolving nature of cyber risks and enhances the protection of compliance operations.

3.3 Benefits of V2X Communication

V2X stands for Vehicle to Everything and it is an advanced system out in transportation sector whereby the vehicles and other entities such as infrastructure, people on the ground and other networks exchange data and information in real time manner. Through collision warnings, alerts to pedestrians, and assistance for lane changing, V2X has gone ahead to enhance traffic safety by Avoiding Accidents leading to tremendous improvements in accident rates. Enhanced traffic flow provides the smart traffic control option, changes to dynamic roads, and even low traffic density in traffic as the result of less fuel utilization and pollution. It is applied to resolve the challenges of the vehicular environment

4 RESULTS AND DISCUSSION

This paper is a review paper which involves a review of review papers with regard to the dynamic security environment in the banks and cloud computing frameworks. This it attributes to rise digitalization and more complex forms of cyber threats. Among all the identified findings, the following can be highlighted: increased incidence of ransomware techniques, phishing, insider threats, and APT in the financial structures and customer databases. Misconfiguration particularly of popular resources in cloud architectures and ransom attacks are considered as threats in the cloud structures paradigms of defenders have been defined with use of artificial intelligence and machine learning in the identification of fraudulent and other behaviours in real-time. Besides, Cyber threat intelligence frameworks are available that can definitively assist in recognizing threat and control risks before they can harm the organization and become normal procedure over cloud banking data integrity through blockchain technology at on-boarding stage. NIST CSF and ISO/IEC 27001 have equally assisted in enhancing structure to CSI especially in HC environments; still, researchers lack works exploring cross-border compliance of CSI and its integration with Zero Trust

models. Some of these issues have been debated many of them have been addressed however, problems such as, how to scale AI, how to prevent malicious insiders, and how to move to quantum safe cryptography prove that there is still much more to be achieved. It is necessary that there should be more development and collaboration to mitigate cyber threats on banking and cloud computing today and in the future. This paper also highlights the increasing need to involve robust cybersecurity strategies in current and upcoming technologies such as V2X communication. In view of advancement and adoptions of connected and autonomous vehicles, real-time anomaly detection, secure encryption, and data exchange technologies are inevitable. Blockchain technology has been cited to have advantages on cloud banking and thus can be adopted and enhanced to fit vehicular networks on aspects of data integrity and availability. Additionally, by widely implementing AI and ML in vehicular systems, it is possible to dramatically enhance the possibility of identifying and mitigating cyber threats on the fly. Thus, the protection of the existing systems and development of the effective strategies for a further attack requires the international cooperation in the sphere, the easing of the international regulations, the transition to the quantum-safe cryptography.

V2X communication facilitates exchange of information between vehicles, infrastructure, and pedestrians in near-real time; hence – increases road safety and operational efficiency. Highlights such as the collision system, pedestrian sensing system, and a lane change system decrease the levels of accidents and enhance transportation safety. V2X improves traffic conditions through the initiation of route changes and the provocation of less traffic density, and better fuel consumption. In addition, V2X complements other advanced technologies such as IoT in the provision of smart navigational networks as well as efficient use of energy in the vehicle. Real-time weather and road condition information make a driver's decision-making process reliable by improving journey reliability. Platooning in combination with V2X boosts up fuel efficiency and decreases the number of carbon emissions due to coordinated movement of vehicles. In emergency response systems, priority routing gave fastest routes to ambulance and fire trucks which reduce response time. Furthermore, through machine to machine communication, V2X improves driver experience provisions such as parking spaces, traffic conditions, and compatibility with smart city services. All these advancements bring safety, intelligence and sustainability to transport, fostering mobility for the future.

5 CONCLUSIONS

V2X is one of the transformative ways, and its application laid down significant aspects in safety, efficiency in transportation and sustainability. It can permit experience to data exchange between vehicles, infrastructure, and pedestrians and reduce as many of the accidents as possible. Also, in turn, it would help to optimise traffic management. It is also future-ready since it can work together with other sophisticated systems like AI, and IoT so that it can tackle problems that are still going to occur with vehicle coordination and emissions. V2X is environmentally friendly: That reduces the use of fuel and emission in dynamic route finding, and the other cooperative driving methods like the leading and following formation. The technology is important in Emergency Response Systems since it increases efficiency in the execution of these operations through traffic prioritized services to emergency vehicles. Essentially, efforts to standardised V2X protocols require collaboration between governments, automotive industries and smart city designers at a national and international level and across manufacturers. To ensure public acceptance of the gain from the smart transportation systems, V2X-secure communication needs to be continuously researched and imbibed and investment is to be made regarding the cybersecurity and data integrity concerns.

ACKNOWLEDGEMENT

This work is partially funded by Center for Advanced Multidisciplinary Research and Innovation. Chennai Institute of technology, India, vide funding number CIT/CAMRI/2025/CFR/010.

REFERENCES

- [1] D. G. Costa, J. C. N. Bittencourt, F. Oliveira, J. P. J. Peixoto, and T. C. Jesus, "Achieving sustainable smart cities through geospatial data-driven approaches," *Sustainability*, vol. 16, no. 2, p. 640, 2024, [Online]. Available: <https://doi.org/10.3390/su16020640>.
- [2] K. I. Shah, M. Khan, S. Abbas, and Z. Hasan, "Intelligent transportation system (ITS) for smart-cities using Mamdani fuzzy inference system," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 2, p. 105, 2018, [Online]. Available: <https://doi.org/10.14569/IJACSA.2018.090215>.
- [3] M. Dibaei, X. Zheng, K. Jiang, S. Maric, R. Abbas, S. Liu, Y. Zhang, Y. Deng, S. Wen, J. Zhang, Y. Xiang, and S. Yu, "An overview of attacks and defences on intelligent connected vehicles," *arXiv preprint arXiv:1907.07455*, 2019, [Online]. Available: <https://arxiv.org/abs/1907.07455>.
- [4] S. Yogarayan, S. F. A. Razak, A. Azman, and M. F. A. Abdullah, "Vehicle to everything (V2X) communications technology for smart mobility in Malaysia: a comprehensive review," *J. Southwest Jiaotong Univ.*, vol. 56, pp. 534-563, 2021, [Online]. Available: <https://doi.org/10.35741/issn.0258-2724.56.4.47>.
- [5] M. Noor-A-Rahim, Z. Liu, H. Lee, M. O. Khyam, J. He, D. Pesch, K. Moessner, W. Saad, and H. V. Poor, "6G for vehicle-to-everything (V2X) communications: Enabling technologies, challenges, and opportunities," *Proc. IEEE*, vol. 110, pp. 712-734, 2022, [Online]. Available: <https://doi.org/10.1109/JPROC.2022.3173031>.
- [6] T. Butt and M. Afzaal, "Security and privacy in smart cities: Issues and current solutions," in *Smart Technologies and Innovation for a Sustainable Future*, pp. 317-323, 2019, [Online]. Available: https://doi.org/10.1007/978-3-030-01659-3_37.
- [7] I. H. Abed and S. Bahadori, "Using density criterion and increasing modularity to detect communities in complex networks," *IJDS*, vol. 2, no. 1, pp. 1-15, Jan. 2025, [Online]. Available: <https://doi.org/10.62051/ijds.v2i1.15>.
- [8] K. Ahmad, "Internet of Things-aided intelligent transport systems," *Wireless Commun. Mobile Comput.*, vol. 1, no. 1, p. 28, 2023.
- [9] K. Ansari and Y. Feng, "Design of an integration platform for V2X wireless communications and positioning supporting C-ITS safety applications," *J. Glob. Position. Syst.*, vol. 12, pp. 38-52, 2013, [Online]. Available: <https://doi.org/10.5081/jgps.12.1.38>.
- [10] J. Clancy, D. Mullins, B. M. Deegan, and J. Horgan, "Wireless access for V2X communications: Research, challenges and opportunities," *IEEE Commun. Surv. Tutor.*, 2024, [Online]. Available: <https://doi.org/10.1109/COMST.2024.3384132>.
- [11] T. Wang, A. Hussain, X. Wang, and S. Maharjan, "Artificial intelligence for vehicle-to-everything: A survey," *IEEE Access*, vol. 7, pp. 1-1, 2019, [Online]. Available: <https://doi.org/10.1109/ACCESS.2019.2891073>.
- [12] M. Houmer, "A secure vehicle to everything (V2X) communication model for intelligent transportation system," *Comput. Intell. Recent Commun. Netw.*, vol. 1, no. 1, p. 102, 2022.
- [13] S. A. Yusuf, A. Khan, and R. Souissi, "Vehicle-to-everything (V2X) in the autonomous vehicles domain – A technical review of communication, sensor, and AI technologies for road user safety," *Transp. Res. Interdiscipl. Perspect.*, vol. 23, no. 11, p. 100980, 2024, [Online]. Available: <https://doi.org/10.1016/j.trip.2023.100980>.
- [14] M. S. Alsabah, N. M. A. Aljarah, and S. V. Paşca, "Intelligent algorithmic approaches to ECG signal classification in heart disease detection," *Electr. Eng. Tech. J.*, vol. 2, no. 1, pp. 25-32, Jan. 2025.

- [15] K. Kim, J. Kim, S. Jeong, J.-H. Park, and H. K. Kim, "Cybersecurity for autonomous vehicles: Review of attacks and defense," *Comput. Secur.*, vol. 103, p. 102150, 2021, [Online]. Available: <https://doi.org/10.1016/j.cose.2020.102150>.
- [16] T. Yoshizawa, D. Singelée, J. Mühlberg, S. Delbruel, A. Taherkordi, D. Hughes, and B. Preneel, "A survey of security and privacy issues in V2X communication systems," *ACM Comput. Surv.*, vol. 55, no. 9, p. 35, 2023, [Online]. Available: <https://doi.org/10.1145/3558052>.

Low-Power Redundant-Transition-Free TSPC Dual-Edge-Triggered Flip-Flop with Single-Transistor Clocked Buffer

Bollepogu Venkateswarlu¹, Muhamad Karnoha Amir², Raivel Raivel³, Gudipati Murali⁴,
Rayala Sateesh⁵ and Sameer Hashim Raheem Al Deen⁶

¹Department of CSE - Data Science, Geethanjali College of Engineering and Technology, 501301 Hyderabad, India

²Department of Mining Engineering, Universitas Sulawesi Tenggara, 93116 Kendari, Indonesia

³Department of Geological Engineering, Universitas Sulawesi Tenggara, 93116 Kendari, Indonesia

⁴Department of CSE-Artificial Intelligence and Machine Learning, KKR & KSR Institute of Technology and Sciences, 522017 Vinjanampadu, Guntur, India

⁵Department of Electronics and Communication Engineering, MLR Institute of Technology, 500043 Hyderabad, Telangana, India

⁶Department of Medical Science, Dijlah University College, 10052 Baghdad, Iraq

bvenkateswarlu.cse@gcet.edu.in, karnohaamir@un-sultra.ac.id, raivel@un-sultra.ac.id, drgudipatimurali@gmail.com, sateesh_ece@mlrinstitutions.ac.in, sameerhashem@duc.edu.iq

Keywords: Dynamic Power, Dual-Edge Triggering, Flip-Flop.

Abstract: Designing low-power flip-flops with better reliability becomes crucial as the need for energy-efficient and high-performance integrated circuits continues to climb. A Low Power Redundant-Transition-Free True Single-Phase Clock (TSPC) Dual-Edge-Triggering Flip-Flop (DETFF) leveraging a Single-Transistor-Clocked Buffer (STCB) is developed and presented in this work as a new way to deal with these problems. With the goal of improving performance and reducing Low power consumption, the proposed design integrates the TSPC & DETFF optimums. architectures by reducing the number of redundant transitions during clocking and guaranteeing strong dual-edge triggering. An unique Single-Transistor-Clocked Buffer is also used to optimize the clock distribution network, which further contributes to power efficiency. Power consumption, speed, and dependability are some of the performance indicators measured by the flip-flop. These are assessed by comprehensive simulations conducted using technologies usually used in the industry. Testing the suggested architecture against current flip-flop designs shows that it is more efficient and reliable. Not to mention completely changing the game when it comes to low-power digital circuits design, the proposed Low Power Redundant-Transition-Free TSPC DETFF with Single-Transistor-Clocked Buffer has also made a substantial impact as an exciting new direction for future integrated circuits, which will require increased reliability and energy efficiency. Among the most potent building elements of processing in the GPU/AI era is the flip-flop (FF). An STC-based, dual-edge-triggered (DET) FF based on buffers is introduced to solve this problem. In the realm of data testing, STC support is great because it only needs one timed semiconductor and gets rid of all the internal and clock excess changes that other DET methods have. Approval for 22nm FD-SOI CMOS post-format recreations at 10% swapping movement.

1 INTRODUCTION

Power consumption is a major concern for CMOS advanced architecture, especially with GPU/AI brain network processor strain. calculate handling power used to make man-made intelligence copies 3.4 months from now [1]. Modern processor timing systems can account for up to 50% of the power consumption [2]. Streamlining power usage is a crucial aspect of addressing the previously mentioned power dispersion issue.

Clock distribution networks and flip-flops make up a processor's timing structure. A clock with only one stage FFs need more power to handle specific information since they process one clock edge at a time while addressing approaching information. Data is processed on both clock edges using dual-edge-triggering (DET) flip-flops. This allows them to continue operating at throughput even in the event of split clock recurrence.

Flip-flop is essential to modern devices and PC design. Parallel data storage and sequential rationale

circuit. Registers, counters, and memory components consistently use goes back and forth.

In the dynamic field of integrated circuit design, finding solutions that save energy is essential for keeping up with the demands of current electronic systems. With the trend towards increasingly integrated systems and lower process nodes, Power consumption is undoubtedly a significant problem. Since they are essential parts of digital circuits, flip-flops are crucial in deciding how well and how efficiently these systems work as a whole:

- The clock and internal redundant switching are omitted entirely from the proposed first DET FF.
- Our FF design recommends using STC buffers with no duplicate transitions. T.

This research introduces a novel design approach that addresses the dual challenges of low power consumption and reliability in flip-flop systems. We propose introducing a novel Low-Power Redundant-Transition-Free TSPC DETFF that combines the advantages of True Single-Phase Clock (TSPC) and Dual-Edge-Triggering Flip-Flop (DETFF) techniques. This design not only seeks to minimize unnecessary transitions during clocking but also leverages dual-edge triggering to significantly enhance speed and efficiency.

A Single-Transistor-Clocked Buffer (STCB) is introduced as a key element of the proposed design. Reducing power consumption even further without sacrificing the robustness needed for dependable circuit functioning, this feature optimizes the clock distribution network. Together, TSPC, DETFF, and STCB form a novel and all-encompassing strategy for designing high-performance flip-flops with minimal power consumption.

The importance of efficient and dependable flip-flops is growing as the semiconductor industry keeps moving the needle on miniaturization and integration. Discover a new development in this field as this research delves into the proposed Low Power Redundant-Transition-Free TSPC DETFF with Single-Transistor-Clocked Buffer and how it could shape the future of digital circuits that are energy efficient. Complete simulations and comparison studies prove this design's effectiveness, opening the door to improvements in low-power, high-performance integrated circuit designs.

Standard signals that synchronise activity on computer circuits and on flip-flops are abided by by clock beats. In response to a clock signal, a flip-flop can store or transfer data. The SR (Set-Reset) and D (Data) flip-flops are the two most prevalent types.

1.1 Sr Flip-Flop

Two sources of information and outcomes regarding. The SR flip-flop includes inputs for Set (S) and Reset (R), as well as outputs for Q and its complement ($\bar{Q}$). It is possible to reset Q yield to 0 by activating the R input and to 1 by activating the S input. In contrast, Q^V yield is used to address Q yield. It is probable that an SR flip-flop will experience a "race condition". When two data sources are active simultaneously, it results in an undefined state.

1.2 D Flip-Flop

The D flip-flop produces two signals, Q and $\bar{Q}$, which are complementary to each other. D (data) input is its only input. D input determines the state of Q output. When the clock signal is engaged, the D input is changed to the Q output. In response to a high (1) or low (0) D input, Q's output shifts from high to low. The inverse of Q output is $\bar{Q}$ output.

Frequently utilized for control, synchronization, and data storage, flip-flops are integral to digital systems. Connecting multiple flip-flops enables the construction of complex sequential circuits, such as registers and counters. These circuits form the foundation of digital systems and devices like computers, calculators, and communication devices. How sequential logic circuits like flip-flops react to clock signals is the subject of ideas like dual-edge and single-edge triggering.

1.2.1 Single-Edge Triggering

To maintain single-edge triggering, a flip-flop must permit changes to its state only on Often called rising-edge triggering, positive-edge triggering occurs at one edge of the clock signal. In positive-edge triggering, the flip-flop can't change its state until the clock signal crosses the rising edge, which is a low-to-high transition. While going from low to high, the flip-flop stays in the same state it was in before the rising edge. To activate their operations, 90% of digital systems rely on the previously indicated trigger.

1.2.2 Dual-Edge Triggering

While flip-flops can only change their state on one of the clock signal edges, dual-edge triggering allows them to do it on both. Everything that has been said so far proves that the flip-flop is capable of sampling input and updating its state anytime there is a change in the clock signal, be it a rising or falling edge. If designers are working on sequential logic circuits, the aforementioned choice gives them more leeway (Fig. 1).

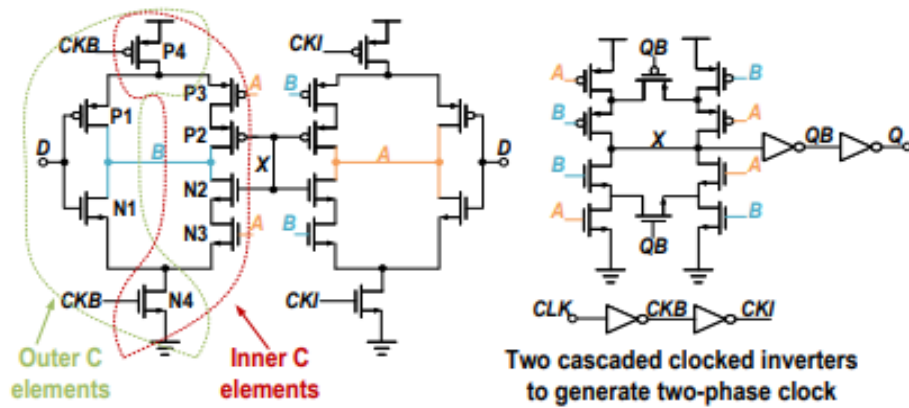


Figure 1: Element DET, floating node C, and FN C-DET.

Although dual-edge triggering isn't as common as single-edge triggering, it's still the go-to method for niche uses that demand pinpoint synchronization or timing.

The decision between single-edge and dual-edge triggering is determined by the specific application and design requirements. For most digital systems, single-edge triggering is effective and easy to implement. However, dual-edge triggering has its applications, among other things, designing fast circuits, implementing protocols, and controlling timing precisely.

1.2.3 Single Transistor Clocked Buffer

A basic digital circuit that is employed to amplify and transmit binary signals that are synchronized with a clock signal is a clocked buffer with a single transistor, sometimes called a single-stage clocked buffer or a latch with a single transistor. For driving and isolating signals across various circuit components, it finds extensive application in digital systems. Just one transistor Along with a clock signal, the single-transistor clocked buffer incorporates a metal-oxide-semiconductor field-effect transistor or complementary metal-oxide-semiconductor.

A periodic square wave, referred to as a clock signal, governs the buffer. The design specifies that the transistor will conduct and allow the input signal to flow towards the output whenever the clock signal is active, whether it's high or low. Transistor goes off, essentially separating, when clock signal is not active. An uncomplicated digital device capable of amplifying and distributing clock-synchronized binary signals is a clocked buffer that uses only one transistor, sometimes called a single-stage clocked buffer or a latch that uses only one transistor,

commonly referred to as a single-transistor clocked buffer. Its primary function in digital systems is to drive and separate signals among various circuit components. In a single-transistor clocked buffer, The only components are the rhythmic signal and a solitary transistor, usually a field-effect transistor (FET) or a complementary metal-oxide-semiconductor (CMOS). It is the timer signal that controls the buffer's functioning and is characterised by a recurring square wave. The transistor conducts while the clock signal is active (either high or low, depending on the design), and permits the input signal to go through to the output. The transistor turns off and is practically disconnected from the circuit when the clock signal is not active.

2 LITERATURE SURVEY

The desire for energy-efficient integrated circuits has encouraged low-power flip-flop design developments. Numerous methods and architectures have been developed to reduce power usage and preserve reliability. We review the literature that supports the proposed "Low Power Redundant-Transition-Free TSPC Dual-Edge-Triggering Flip-Flop using Single-Transistor-Clocked Buffer" and contextualize its contributions.

2.1 Traditional Flip-Flop Architectures

Key digital circuit components include master-slave and D-type flip-flops. These frequently used designs are power inefficient, especially with redundant transitions and static power waste during idle states.

2.2 The True Single-Phase Clock (TSPC)

The flip-flop architecture addresses power limitations in previous systems. TSPC saves power and runs faster. Minimising unnecessary transitions and boosting reliability remain difficulties.

2.3 Dual-Edge-Triggered Flip-Flops (DETF)

These architectures are renowned for their ability to operate on both rising and falling clock edges, resulting in a threefold increase in data transfer rates. This trait improves speed and performance, but power consumption and dependability must be balanced.

2.4 Clock Distribution Networks

Digital circuits rely on clock distribution networks, which impact power consumption and signal integrity. Single-Transistor-Clocked Buffers (STCB) have been studied to optimise clock distribution and reduce power overhead.

2.5 Low-Power Redundant-Transition-Free Flip-Flops

Many recent flip-flop designs aim to reduce redundant transitions during clocking occurrences. These designs aim to reduce dynamic power consumption and boost efficiency, paving the way for more complete flip-flop systems.

We identify gaps and obstacles that our suggested "Low Power Redundant-Transition-Free TSPC DETF using Single-Transistor-Clocked Buffer" addresses by reviewing the research in three important areas. We situate our work within the context of improving low-power, high-performance flip-flop topologies for next-generation integrated circuits by comprehending the state-of-the-art.

3 METHODOLOGY

In terms of intensity and concession, snares and flip-flops are the most important components of an arrangement. They classify different kinds of definite and express heartbeat ages, alongside static and semi-dynamic returns with a single edge set off. We provide an ip-DCO, an implicit pulse, semi-dynamic flip-flop, which incorporates the quickest delay of each flip-flop plus a large amount of non-positive

setup time information. On the other hand, for the majority of mission-critical pathways in an organization, an explicit-pulsed static flip-flop (ep-SFF) is the smartest energy-saving option [3].

Two edge-set off back-sells are being considered for potential power use. Certain conventional two-edge-set off layouts are found to be unfit for basic uses due to their dull appearance and excessive district discipline. One additional express beat double-edge-set off flip-flop Introduces a comparable value to a single-edge-triggered structure with much reduced power consumption in the networks that control flip-flops and clock movements [4].

Specifically, New Dual-Edge-Triggered (DET) flip-flops that are long-lasting and use less power are proposed in this study. By taking input data at both clock edges, DET-FFs can dramatically improve energy efficiency using the traditional Single-Edge-Triggered (SET) method. Dynamic voltage scaling, when used in tandem, should significantly improve efficiency. Previous DET-FF plans were limited to low voltage frameworks due to their inherent insensitivity to PVT (Process, Voltage, and Temperature) variations. For reliable performance at voltages as low as those in a neighboring edge framework, it is recommended to use a fully static certified single-stage coordinated DET-FF [5]-[7].

To address clock skew concerns and enable low-power operation, we employ a True-Single-Phase-Clocking (TSPC) architecture instead of the standard two-stage clocking strategy used in common DET-FFs. In addition, the low-voltage framework is guaranteed to operate reliably by the totally static implementation. This DET-FF is designed to work with 28nm CMOS technology. on the path to additional improvements in support arrangements, an intensified evaluation is planned that incorporates post-plan Monte Carlo simulation is employed to cover extensive Process, Voltage, and Temperature (PVT) ranges. Based on thorough testing and evaluation in comparison to previous DET-FFs, The proposed DET-FF demonstrates functionality at an impressively low voltage of approximately 0.28 V across a temperature range spanning from -40 °C to 120 °C. This is achieved while maintaining nearly optimal energy efficiency and power-delay product [8]-[12].

The utilization of dual-edge-triggered (DET) assisted operation presents an intriguing option for low-power, standard execution strategies. At the clock repeat stage, DET action can produce throughput comparable to that of conventional single-edge synchronous systems. In the workplace, an organization's regular contributions to overall power

usage can result in significant savings. Registers with specific model data on both clock edges should remain consistent and perform dual-edge-triggered (DET) operations reliably.

These registers are more disorganised and frequently encounter clock cross-talk between the internal revised clock and the primary clock, in contrast to their single-edge counterparts. The aforementioned crossing. Especially when operating with scaled power supplies and under process variations commonly observed in nanoscale advancements, might cause disagreements among cells and problems with reasoning. The unique static DET flip-flop (DET-FF) presented in the aforementioned paper completely avoids clock skew issues with a real single-stage clock because it doesn't require a changed clock edge for value. Although conventional DET-FFs fail miserably at low-voltage working locations, the suggested DET FF performed admirably when implemented in a conventional 40nm CMOS development. At a close edge supply voltage of about 500 mV, the proposed cell gives the smallest power-delay-product and a 35% reduction in CK-to-Q latency, all while ignoring any remaining DET-FF executions [13].

Our investigation into possible power savings entail transitioning from conventional single-edge-triggered (SET) flip-flops to dual-edge-triggered (DET) flip-flops, as discussed in the aforementioned work. In the aforementioned study, we introduce a new family of D-type double-edge set-off algorithms that, unlike previous plans, use significantly fewer semiconductors in their execution. The process of power dissipation in these flip-flops and single-edge-triggered flip-flops is investigated through building-level tests, logical reflections, and simulations. The effect of data successions on energy distribution regarding single and double edge set off back-pedals is remembered for investigation in an autonomous execution focus. We next look at the possibility of achieving framework-level energy reserve funds by switching from registers built about single-edge set off back-peddles to registers produced about two-edge set off back-peddles.

The results remain highly promising, illustrating that specific dual-edge-triggered flip-flops can achieve significant energy savings with minimal complexity. Low-energy dual-edge-triggered flip-flops are presented in the aforementioned paper [15] using a creative execution method. The new approach employs a clock branch-sharing plan in an effort to reduce the number of timed semiconductors in the design. There are two other approaches that have recently been suggested for use in limiting hinder and

exchanging actions: part way and contingent release. The latest CBS IP configuration outperforms other state-of-the-art dual-edge-triggered flip-flop designs in terms of power consumption (up to 20%) and power degradation probability (up to 12.4%) [16].

One significant drawback of integrating D-type dual-edge-triggered flip-flops (DET-FFs) into VLSI framework design is the need for a substantial quantity of semiconductors. Two new DET-FF circuits, one static and one dynamic, reduce the semiconductor count to a level comparable to that of specific conventional single-edge-triggered flip-flops (SET-FFs). In addition to functioning well at high frequencies, these novel circuits are easy to use, have a transparent architecture, and are very resistant to race and metastability problems (static and dynamic). The use of DET-FFs in the configuration of VLSI frameworks is more widespread, promising, and applicable due to these reasons [17].

According to the 2008 Global Innovation Guide for Semiconductors, power consumption is one of the three primary issues and a major cause of performance bottlenecks in systems. A significant amount of on-chip power is consumed by the clock framework in real-world use. This framework consists of failure flops and the clock circulation organisation. the aforementioned article discusses several near-term configuration draws for low-power timing frameworks. Using fewer timed semiconductors is one of the most effective strategies to To alleviate the clock load limit, we introduce a timed pair—an innovative flip-flop design that decreases the number of surrounding timed semiconductors by approximately 40%. This approach enables a 24% reduction in the power needed to drive the clock. Furthermore, through the integration of low swing and double-edge timing, these novel flip-flops can be seamlessly incorporated into timing frameworks [11]-[14].

The static dual-edge-triggered (DET) flip-flop designs described in the previous study exhibit remarkable circuit behavior since C-components are consistently used in them. Two better execution strategies improve the standard Hook MUX DET flip-flops, making them immune to signal intensity fluctuations at any point in their internal circuits. These enhancements help bring five first-rate DET flip-flops to market. The low-energy propagation that comes with flip-flops is a typical result of input errors. A comparison is made between new dual-edge-triggered (DET) flip-flops and existing DET flip-flops using replication in a high-performance 28 nm CMOS technology. The new flip-flops demonstrate better features, such as an indeterminate

power-delay product (PDP), across several exchange-related jobs. Through rigorous voltage scaling and Monte Carlo simulations, we confirm that the proposed designs are resilient against Process, Voltage, and Temperature (PVT) fluctuations [18].

The Advanced High-performance Bus, a high-performance member of the AMBA bus family, which stands for Advanced Microcontroller Bus Architecture. The components that make up a framework should be able to communicate with one another. ARM defines AHB standards, which provide communication between on-chip processors and external memory interfaces located off-chip. An AHB framework that can support four slaves and one expert is detailed in the aforementioned study, along with its design and implementation. the aforementioned work is the basis for the AHB Convention's building blocks, which include multiplexers, slaves, decoders, and bosses. The aforementioned AMBA-AHB convention can be used into any application so long as the strategy aligns with AHB principles. The following architectural components are commonly seen in Verilog systems: nurture experts, slaves, decoders, and multiplexers. To promote an environment conducive to confirmation, System Verilog (SV) is employed. Using Coach Designs' High Level Confirmation Apparatus, QuestaSim, the plan is duplicated, checked, and code and valuable inclusions are registered the third.

4 EXISTING STATE OF THE RESEARCH ENDEAVORS

Documentation of Just One Change Double-Edge-Triggered flip-flop is abbreviated as STCDET. Several benefits are offered by the aforementioned style of flip-flop design in comparison to more conventional designs.

Higher data transfer rates are still within reach when using STCDET flip-flops instead of single-edge-triggered ones. By collecting data on the rising

and falling edges of the clock in a single clock cycle, STCDET flip-flops allow for quicker and more efficient data processing, virtually doubling the data transfer rate.

Variations in the timing of clock signals, or clock skew, can affect the reliability and performance of flip-flops. With two clock edges used to capture data, STCDET flip-flops are less susceptible to clock skew. aforementioned mitigates the effect of clock skew on flip-flop timing needs. Improved Resistance to Noise: STCDET flip-flops exhibit enhanced resistance to noise.in opposition to flip-flops that just have one edge trigger. Their resistance to clock signal noise and hiccups is due to the fact that they capture data on both clock edges. the aforementioned noise immunity helps keep data intact and decreases the chance of incorrect data collecting.

In comparison to other flip-flop designs, STCDET flip-flops are able to achieve lower power consumption. It is still possible to get greater data transfer rates with STCDET flip-flops as opposed to single-edge-triggered flip-flops. Because STCDET flip-flops can capture data on both the rising and falling edges of the clock within a single clock cycle, they virtually double the data transmission rate, allowing faster and more efficient data processing. This achievement can be further enhanced with power optimization techniques like clock gating, made possible by a decreased vulnerability to clock skew. It is possible to do simpler timing analyses for STCDET flip-flops than for flip-flops with more intricate designs. Meeting time constraints is feasible because data is captured on both the positive and negative edges of the clock in a single clock cycle. The design process is streamlined and made easier. aforementioned may result in versions of the design that are both faster and more efficient.

One approach to making them backward compatible with STCDET flip-flops (Fig. 2) [19], is to implement single-edge-triggered flip-flops. All of the above makes it possible to incorporate them into preexisting digital systems without disrupting the system's overall architecture or necessitating huge changes.

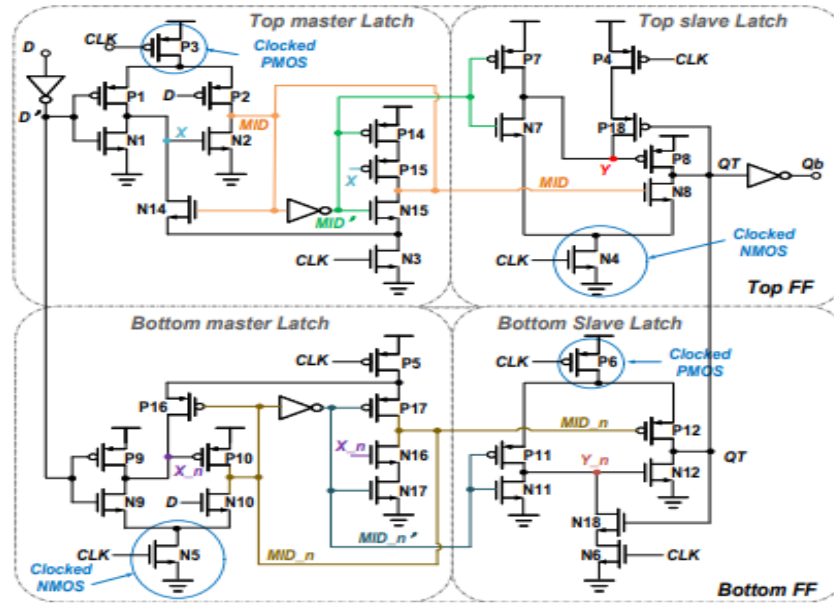


Figure 2: Schematic of the proposed TSPC-DET flip-flop.

4.1 Proposed

A decrease in the edge voltage of CMOS circuits brought on by voltage scaling increases the subthreshold leakage current and, eventually, static power scattering. We provide LECTOR, a short-circuite method for fabricating CMOS doors, which significantly reduces leakage current while increasing unique power dispersion. One method that has been suggested is to place two semiconductors for controlling leakage – a p-type and an n-type transistor are placed within a logical gate, with one set of terminals for each type being connected to the other's source. In the mentioned architecture, one of the LCTs is consistently "close to its end voltage" irrespective of the mix of data sources. This reduces leakage currents by reducing the path's resistance from V_{DD} to ground, as indicated before. A static CMOS complex door execution is the initial step in creating a spillage controlled circuit from an entry-level netlist regarding the given circuit. Once added, LCTs remain. LECTOR's primary value is in the fact that, in contrast to competing methods, it effectively reduces leakage regardless of whether the circuit is active or not (Fig. 3). In complementary metal-oxide-semiconductor (CMOS) circuits, voltage scaling reduces the edge voltage, which increases the subthreshold leakage current and the static power dissipation.

The LECTOR method, which we propose, provides a novel approach to reducing leakage current in CMOS doors while enhancing dynamic power scattering. Integrating a p-type and an n-type leakage control transistor (LCT) into a logical gate is a critical part of the previously described approach. The gate terminal of each LCT is controlled by the source of the corresponding semiconductor. Following the aforementioned approach, LCTs are typically "close to their end voltage" regardless of the data mix. This reduction in leakage currents is fundamental, the aforementioned increases blockage in the way from V_{DD} to ground correspondingly.

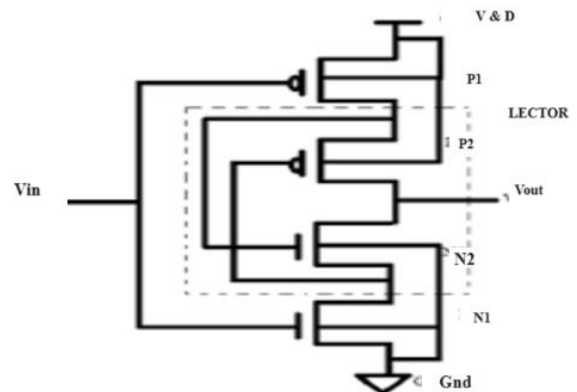


Figure 3: Schematic about LECTOR.

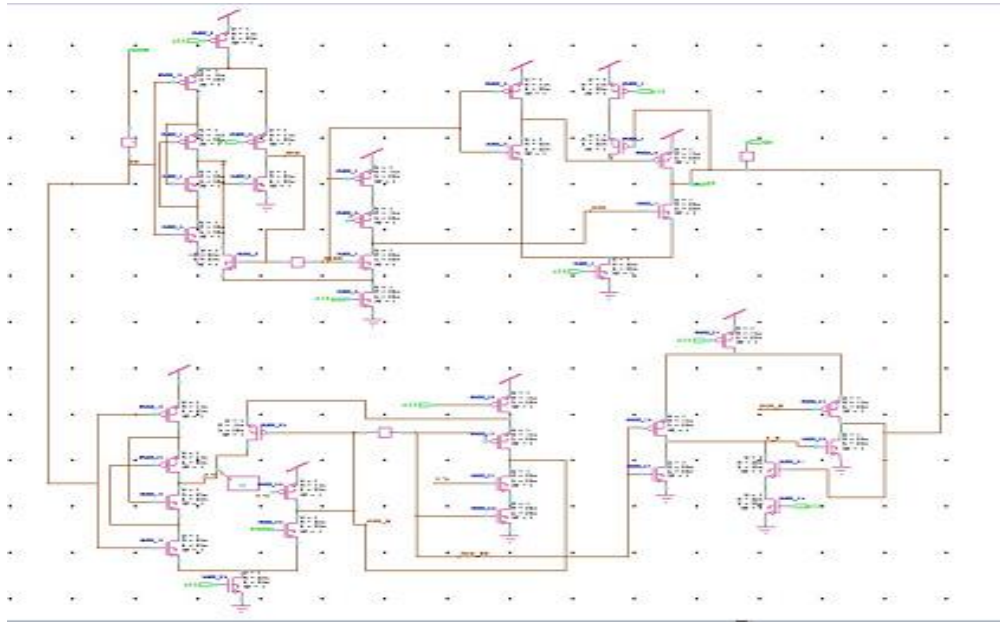


Figure 4: Schematic about TSPC STCDET using LECTOR.

4.2 Operation about Top FF in STC-DET

When $CLK = 0$, as shown in the worked-on reasoning diagram in Figure 5a, The transistors (P2, N2) effectively function as a virtual inverter. The upper master hook's timed PMOS P3 activates when $CLK = 0$, causing node X to transition to D'. Conversely, in the upper slave hook of Figure 4, the timed NMOS N4 remains off due to $CLK = 0$, indicating that node Y will not reach 0, and PMOS P8 is also off. Later on, data is sent to MID via the top expert lock (bolt in the top left corner of the previous illustration). Due to the inability of VDD and QT to interface with either GND, the top FF's QT is drifting (refer to the upper left corner of Figure 5a).

Figure 4 depicts a negative-set off STCB comprising of semiconductors N1, N2, P1, P2, and P3. The sign testing channel is a single coordinated semiconductor, P3. Unlike STCDET, FN_C DET and FS-TSPC (Fig. 3) exhibit a resistance-temperature (RT) component between a timing PMOS and a timing NMOS. Similarly, there is no discrepancy. The upper master latch (located in the upper left of Fig. 4) features a clock-driven NMOS transistor (N3), although it's used for guarding rather than data analysis. Regarding semiconductor P3, its timing is similar to certain. A bolt is shown in Figure 4 to exhibit the definite abide timing of four semiconductors (P3, N4, N5, and P6) as they all abide on the same information testing method. Using

semiconductors (N4, N7, N8, P7, P8), top FF develops an additional sure set off STCB.

At $CLK=1$, the timed PMOS P3 is off, hence the routes leading to P1 and N2 in Figure 4's upper master latch are also off. Attendant (N3, N15, P14, P15) maintains reasonable territory regarding MID in the same manner as indicated earlier. The pull-down guardian (N14, N3) will maintain X in its current state while its rationality state is 0. However, Y becomes MID' when the anticipated NMOS, N4, activates in the top slave lock. Using semiconductors (N8, P8) as a virtual inverter, we can change the state from MID—the only possible state before the clock's rising edge—to QT. Therefore, the top flip-flop is activated at the positive edge of the clock.

4.3 Operation of Bottom FF in STC-DET

Base master latch triggers (bottom left in Figure 4 occur when $CLK = 0$ in the base flip-flop, clock NMOS, and N5. Consequently, the specific connections to N9 and P10 remain dormant, and the protector (N16, N17, P5, P17) maintains MID_n's pristine logical condition. When the attendant (P16, P5) is sure that X_n is 1, they will protect the rational territory surrounding it. However, at the top of the image, the planned PMOS P6 switches on at Here, in the base slave latch (bottom right in Fig. 4), CLK is set to 0. This transforms Y_n into MID_n', which is similar to MID_n but not quite. The subsequent

realization of the P12 and N12 virtual inverter functions causes the generation of the MID_n signal, which is the second MID_n and the leading edge of the clock as it approaches QT (see bolt in the left half of Figure 5b). That being the case, the base flip-flop is secured by the negative edge of the work.

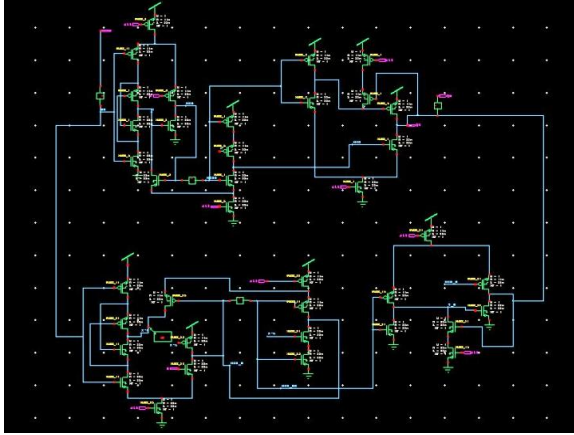


Figure 5: Schematic for STCDET using LECTOR.

5 RESULT AND ANALYSIS

Figures 5 and 6 present the outcomes of the Time-Shifted Pre-Charge (TSPC_DET) in the STCDET system using the novel LECTOR technique.

The diagram depicts an STCDET design utilizing the LECTOR approach, which reveals a notable reduction in power consumption.

The inner workings of an STCDET system that uses LECTOR technology, as explained in Figure 7, are shown below with corresponding simulation waveforms.

The following Figure 8 shows the spatial arrangement of the suggested circuit, referred to as Time-Shifted Pre-Charge (TSPC) in Self-Timed Content-Addressable Memory (STCDET).

The graphic depicts the project's power consumption metrics by offering a comprehensive overview of the power values associated with its execution, as shown in Figure 9.

We also provide a comparison table between proposed and extension (Table 1).

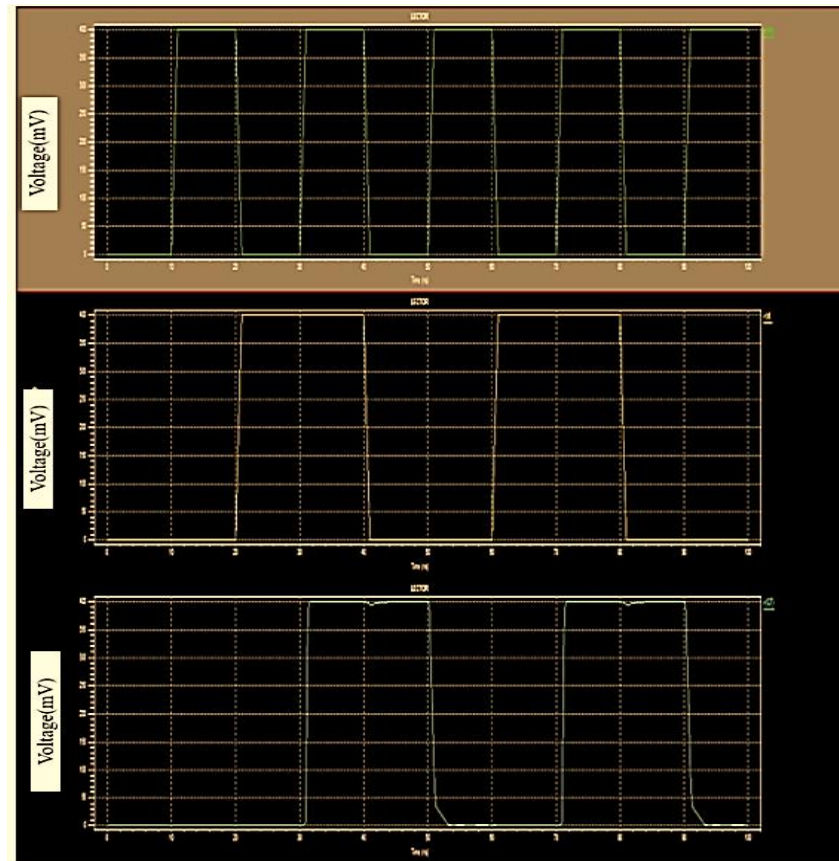


Figure 6: Waveform of STCDET circuit with LECTOR approach.

```
* Device and node counts:
      MOSFETs - 54
      BJTs    - 0
      MESFETs - 0
```

Figure 7: Area.

```
Power Results
v1 from time 0 to 1e-007
Average power consumed -> 1.206081e-007 watts
Max power 1.861622e-006 at time 3.1223e-008
Min power 1.793915e-008 at time 4e-008
```

Figure 8: Power value.

```
delay = 2.0820e-008
Trigger = 1.0250e-008
Target  = 3.1070e-008
```

Figure 9: Delay value.

Table 1: Comparision table.

	AREA	POWER	DELAY
PROPOSED	42	1.886586e-007W	2.0205e-008
EXTENSION	54	1.206081e-007W	2.0820e-008

6 CONCLUSIONS

Built on the assurance that it entirely eliminates redundant transitions in double edge-triggered flip-flops using STC buffers, STC-DET is introduced as a new Low-power flip-flop that is triggered by both edges and does not undergo repeated transitions. Both positive-edge and negative-edge triggers are part of its environmental context. Different from earlier DET frameworks, STC eliminates all clock repeating sections, modifications, and internal surplus advances between two timed semiconductors. With everything else being equal, The information testing circuit for each positive-edge-triggered and negative-edge-triggered STC buffer has precisely one corresponding semiconductor. Additionally, the designed STC-DET is entirely transparent. To facilitate efficient data exchange among all DET designs, STC-DET minimizes power consumption throughout the board at different voltages for operation. The suggested STC-DET achieves the lowest power consumption in

the normal switching activity range among all state-of-the-art DET flip-flops. However, applying the LECTOR method to STCDET can still further reduce power consumption.

REFERENCES

- [1] S. Lapshev and S. M. R. Hasan, "New low glitch and low power DET flip-flops using multiple C-elements," IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 63, no. 10, pp. 1673-1681, Oct. 2016.
- [2] J. Tschanz, S. Narendra, Z. Chen, S. Borkar, M. Sachdev, and V. De, "Comparative delay and energy of single edge-triggered and dual edge-triggered pulsed flip-flops for high-performance microprocessors," in Proc. ISLPED, 2001, pp. 147-152.
- [3] A. Karimi, A. Rezai, and M. M. Hajhashemkhani, "A novel design for ultra-low power pulse-triggered D-flip-flop with optimized leakage power," Integration, vol. 60, pp. 160-166, Jan. 2018.
- [4] G. Lee, S. Shin, and Y. Lee, "A fully static true-single-phase-clocked dual-edge-triggered flip-flop for near-threshold voltage operation in IoT applications," IEEE Access, vol. 8, pp. 40232-40245, 2020.
- [5] R. Hossain, L. D. Wronski, and A. Albicki, "Low power design using double edge triggered flip-flops," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 2, no. 2, pp. 261-265, Jun. 1994.
- [6] H. You, J. Yuan, Z. Yu, and S. Qiao, "Low-power retentive true single-phase-clocked flip-flop with redundant-precharge-free operation," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 29, no. 5, pp. 1022-1032, May 2021.
- [7] P. Zhao et al., "Low-power clocked-pseudo-NMOS flip-flop for level conversion in dual supply systems," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 17, no. 9, pp. 1196-1202, Sep. 2009.
- [8] A. Bonetti, A. Teman, and A. Burg, "An overlap-contention free true-single-phase clock dual-edge-triggered flip-flop," in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), May 2015, pp. 1850-1853.
- [9] M. W. Phyu, K. Fu, W. L. Goh, and K. S. Yeo, "Power-efficient explicit-pulsed dual-edge triggered sense-amplifier flip-flops," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 19, no. 1, pp. 1-9, Jan. 2011.
- [10] M. W. Phyu, K. Fu, W. L. Goh, and K. S. Yeo, "Ultra-low power pulse-triggered CNTFET-based flip-flop," IEEE Trans. Nanotechnol., vol. 18, pp. 756-761, 2019.
- [11] P. Zhao et al., "Low-power clocked-pseudo-NMOS flip-flop for level conversion in dual supply systems," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 17, no. 9, pp. 1196-1202, 2009.
- [12] H. M. Naser, O. A. Al-Ani, M. F. Mosleh, and F. Arith, "Umbrella-Shaped Wideband MIMO Wireless Communication Antenna," J. Telecommun., vol. 5, no. 4, pp. 46-53, Dec. 2023.
- [13] A. Bonetti, A. Teman, and A. Burg, "An overlap-contention free true-single-phase clock dual-edge-triggered flip-flop," in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), 2015, pp. 1850-1853.

- [14] I. H. Abed and S. Bahadori, "Using Density Criterion and Increasing Modularity to Detect Communities in Complex Networks," *Int. J. Data Sci.*, vol. 2, no. 1, pp. 1-15, Jan. 2025.
- [15] R. Hossain, L. Wronski, and A. Albicki, "Low power design using double edge triggered flip-flops," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 2, no. 2, pp. 261-265, 1994.
- [16] P. Zhao et al., "Low-power clock branch sharing double-edge triggered flip-flop," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 15, no. 3, pp. 338-345, 2007.
- [17] A. Gago, R. Escano, and J. Hidalgo, "Reduced implementation of D-type DET flip-flops," *IEEE J. Solid-State Circuits*, vol. 28, no. 3, pp. 400-402, 1993.
- [18] S. V. Devarapalli, P. Zarkesh-Ha, and S. C. Suddarth, "A robust and low power dual data rate (DDR) flip-flop using C-elements," in *Proc. Int. Symp. Qual. Electron. Design (ISQED)*, 2010, pp. 147-150.
- [19] B. M. Krishna, J. Kalyani, Y. Venkatesh, V. Balaji, T. Umamaheswari, and Y. Srivalli, "An Efficient Low-Power Redundant-Transition-Free TSPC Dual-Edge-Triggering Flip-Flop Using Single-Transistor-Clocked Buffer," *International Journal for Modern Trends in Science and Technology*, vol. 10, pp. 369-376, 2024.

Rethinking Student-Robot Social Interactions in GenAI Capitalism and its Role in Shaping Cybersocialization through a Sociological Lens

Ahmad Al Yakin¹, Luis Cardoso², Ali Said Al Matar³, Yasir Ahmed Mohammed Ridha Alsbbagh⁴, Muthmainnah Muthmainnah¹, Ahmed J. Obaid⁵ and Mohammed Wajeeh Hussein⁶
¹Teacher Training and Education Universitas Al Asyariah Mandar, 91315 Polewali Mandar, Indonesia
²Polytechnic Institute of Portalegre, University of Lisbon. Portugal, 7300-110 Lisbon, Portugal
³Center For Language and Foundation Studies, A'Sharqiyah University, 400 Ibra, Oman
⁴English Department Al-Iraqia University, College of Arts, 10071 Baghdad, Iraq
⁵Faculty of Computer Science and Mathematics, University of Kufa, 54003 Najaf, Iraq
⁶Department of Computer Techniques Engineering, Dijlah University College, 10052 Baghdad, Iraq
ahmadalyakin76@gmail.com, lmc Cardoso@ippportalegre.pt, ali.almatari@asu.edu.om, yasir_mohammed@aliraqia.edu.iq, muthmainnahunasman@gmail.com, ahmedj.aljanaby@uokufa.edu.iq, mohammed.alameen@duc.edu.iq

Keywords: AIEd, Cybersocialization, Robot Anthropomorphism, Human-Machine Interactions.

Abstract: This study investigates the effects of artificial intelligence (AI) on social interactions in educational settings, with a particular focus on student-robot interactions within the framework of generative AI (GenAI) capitalism. It examines how these interactions shape cybersocialization from a sociological perspective. The goal is to investigate how AI-powered systems like virtual assistants, chatbots, and intelligent tutoring systems impact students' emotional, social, and collaborative abilities in both in-person and online learning environments. The study used a mixed-methods approach, surveying 29 professionals in pre-service teacher education and conducting five student interviews using semi-structured questions. We derived qualitative insights from open-ended responses and analyzed quantitative data using descriptive statistics, t-tests, ANOVA, and Cronbach's alpha. The results demonstrate that tools driven by AI improve student engagement, encourage group work, and offer emotional support while studying. However, issues such as data privacy, the digital divide, and achieving a balance between AI-driven and human-driven interactions emerged. The study suggests a mixed approach to education, using AI to supplement rather than supplant human interactions. To ensure equal access to technology, the study emphasizes the importance of AI tool training for teachers and advocates for measures to reduce the digital divide. Furthermore, as a future paradigm for education, we propose a cybersocialization approach that encourages both technical competence and vital interpersonal relationships.

1 INTRODUCTION

Examining the dynamics of GenAI Capitalism through a sociological lens provides insight into the possibilities and threats that affect human-to-human interactions in the age of cybersocialization in terms of student-robot social interactions [1]. Intelligent tutoring systems, chatbots, and other AI-based educational tools are just a few examples of the new ways that students and robots are interacting thanks to the rise of Generative AI (GenAI) in the classroom [2]. These systems are radically changing the way students interact with each other in the classroom, with an emphasis on more automated and data-driven

interactions that could eventually replace some parts of more conventional human relationships. From a sociological perspective, GenAI enables a type of cybersocialization in which students become more dependent on non-human agents to assist them in educational activities or to perform various educational tasks [1]. It is possible that the development of students' social skills could be influenced by the use of AI-powered robots that mimic social cues and replicate interpersonal interactions, in addition to teaching technical skills such as programming in robotics education. On the one hand, these tools increase the efficiency and customization of learning, on the other hand, they make us wonder whether we will become too dependent on robots to

address our emotional and social needs in the future [3].

Robot anthropomorphism and human-machine interactions are at the heart of GenAI-driven cybersocialization. Students can form emotional bonds with robots that mimic human behavior, which can change their perspective on human interactions [4]. According to sociologists, these changes can cause relationships to become more transactional, data-driven, and influenced by the economic forces driving the growth of AI, ultimately leading to the commodification of social interactions. Students can experience physical and mental alienation and depersonalization as a result of the push for more efficient education systems that incorporate AI. This occurs when students lose touch with the emotional depth of human relationships [5].

Equally important in this new environment are issues of access to technology and efforts to minimize the digital divide to ensure equitable education. Under GenAI Capitalism, students from low-income families may not have the opportunity to acquire the cyber competencies they need to succeed in today's digital economy [6]. Therefore, socioeconomic groups' experiences with cybersocialization through GenAI devices vary, which in turn exacerbates existing inequities.

Cultural resistance is another major obstacle. Indonesians highly value traditional teaching methods, and they view the idea of introducing robots into schools with skepticism. Teachers, who play a critical role in their students' experiences, are often wary of the human aspect of education, which is essential for their emotional and social growth, and are reluctant to embrace AI [7], [8]. This study aims to bridge the technology gap by addressing concerns about the diminishing human interaction and the important social skills acquired through face-to-face interaction, which are driving this trend.

In GenAI Capitalism, student-robot interactions can lead to new ways of learning and better computational thinking, but there are also ethical, emotional, and social issues that should not be ignored. Finding a middle ground between utilizing AI for educational purposes and preserving the humanistic component of learning, which is fundamental to the advancement of society, is the goal of this research and provides a solution for a humanistic learning model to address these issues. The use of AI in the classroom with effective instruction will not reduce the face-to-face interaction between instructors and students by combining traditional teaching methods with AI tools.

2 LITERATURE REVIEW

Forming strong social bonds with other students is one of the most important things students in hybrid learning can do to increase their satisfaction, decrease dropout rates, and foster intellectual exchange in a safe environment. However, many students who take fully online courses report feeling disconnected from the community [9]. A number of studies have proposed methods to increase the social presence and sense of belonging of online students in hybrid learning environments [10], in response to the growing popularity of internet-based learning in higher education. Many of these strategies propose ways for online instructors to interact with their students, such as sharing personal stories, using humour and emoticons, and allowing students to participate in class discussion boards or social media groups. They also suggest ways to modify course design, such as limiting class sizes and structuring collaborative learning activities [11].

Few technological solutions specifically aim to assist learners in forming the social connections that emerge, despite the growing demand for research into developing and designing technologies to address these issues in hybrid learning contexts [12]. The current lack of technology to assist online learners in forming social connections may be due to the trend of restricting social interaction to learning-oriented academic tasks, which often view social interaction as merely a means to an end (intended learning outcomes). Group projects allow students to meet each other on a more intimate level and find common ground [13]. Students often form temporary bonds when working together on a group project, but once the pressure to complete the task subsides, online students often diverge [14].

The design of AI technologies for hybrid learners reflects this pitfall. Most of these technologies attempt to facilitate the cognitive learning process of LMS-based hybrid learners, such as tools that facilitate peer discussion [15], learner collaboration [16], mass editing of lecture videos [17], [18], and encouraging help-seeking in online discussion forums [19]. The HCI community is still in the early stages of investigating how to create technologies to foster social interaction among college learners [20]. According to [21] several preliminary studies have attempted to understand the current state of online learners' social relationships.

Researchers have shown that students taking hybrid classes tend to form informal friendships based on shared experiences and interests, such as living in the same city, common ground for self-introduction,

or group projects [22]. Few students form relationships that last beyond the classroom, and class-orientated activities, such as group projects, appear to be one of the few opportunities for online learners to engage closely with each other and find common ground [23]. We highlight the paucity of research on developing tools that can address the social interaction challenges faced by hybrid learners and make their learning experiences more positive. Optimal collaborative learning processes are often based on strong social ties. Students need to trust each other, feel part of a learning community, and be physically close to each other to create a safety net that allows them to take more risks and adopt an adventurous mindset while learning [24]. The application of AI technology to improve the general framework of education, personalize the learning experience, and accelerate the learning process is known as Artificial Intelligence in Education (AIED). Intelligent tutoring systems (ITS), adaptive learning platforms, and automated assessments are the main implementations of AIED. To improve educational outcomes, these systems can monitor students' progress, determine areas where they struggle, and provide individualized solutions [25].

AIED's ability to personalize learning is a major plus. AI can personalize content using student performance data to suit each student's learning style and needs, making learning much more efficient. For example, systems like iTalk2Learn or Pearson's AIDA offer individualized, step-by-step instruction based on each student's strengths and weaknesses, allowing them to learn complex topics in greater depth [26]. However, AIED also has a significant impact on social interactions in the educational environment. A primary concern is that students and teachers may have fewer opportunities for face-to-face communication due to the increased use of AI. When AI-based technology acts as a mediator between students and instructors, it reduces the opportunity for direct human interaction, which is crucial for emotional and social development. AI-mediated interactions often lack non-verbal cues such as facial expressions, gestures, and tone of voice, which are critical in developing relationships between teachers and students.

To understand the challenges students face in their current social interaction processes, this study adopts a human-centered approach. We can use this information to create AI systems and other technological tools that help students in college with the demands and challenges of social interaction. Conflicting research has examined the impact of AIED on students' social adaptability. While AI devices have

the potential to enhance students' problem-solving and creativity skills, there is some evidence that they may have the opposite effect on students' interpersonal interactions, especially if they become overly reliant on them. Less contact with instructors and classmates may severely affect social skills such as empathy and collaboration. Blended learning models, which integrate AI devices with human-led instruction, are critical to addressing these challenges.

This study investigates social interaction and how students can benefit from both technical efficiencies and authentic social engagement with this strategy, ensuring that AI enhances rather than replaces human interaction. We are taking significant strides towards a more comprehensive educational setting where social interaction remains essential and raising public awareness about the necessity of balancing the application of AI with traditional teaching methods in higher education.

3 METHODOLOGY

3.1 Research Design

We conducted the mix method, an implementation and evaluation study to provide an overview of the human-centred AI-mediated social interaction design, with the aim of improving the social interaction process of teacher education learners in higher education. The AI-mediated social interaction course is a survey study where we used the AI-ChatGPT agent and LMS as test subjects to learn about the needs of AI-mediated social interaction design.

3.2 Research Instrument

This research instrument is quantitative; this survey instrument consists of two parts. Part 1 uses a nominal scale to gather demographic information from respondents. Meanwhile, Part 2 aims to investigate respondents' factors that influence the use of AI. We modified Part 2 of the instrument based on the original scale [1] and other pertinent studies [2]. We changed the instrument's wording to ensure it aligns with social interaction indicators, thereby facilitating its interpretation and understanding. Three experts in educational technology participated in the discussion to ensure the instrument's content validity during its development [15]. We piloted this survey instrument to strengthen its validity. We conducted preliminary tests or trials using factor analysis and Cronbach's alpha. These findings led to the listing of 20 primary data collection indicators. We examined respondent

responses using a 5-point Likert scale, with 1 representing the lowest (strongly disagree) and 5 representing the highest (strongly agree) score.

We developed semi-structured interview questions using a qualitative approach. To establish the validity of the questions, we invited educational technology experts to discuss the first draft. The interview questions included five items derived from the quantitative data analysis. The semi-structured interviews, which were conversational in nature and based on the quantitative phase analysis, may vary in phrasing from one interview to the next. Still, we asked all respondents to provide consistent responses.

3.3 Participants

This study involved 29 pre-service teacher education professionals. To better understand hybrid learners' experiences, challenges, and feedback about integrating AI into their social interactions during learning, we conducted semi-structured interviews with five teacher education students enrolled in the education seminar course. The Open University of Majene Regency, West Sulawesi.

Prior to participating in the survey or semi-structured interviews, all participants in this study provided informed consent to participate in the learning process using the AI platform and to use a smartphone or laptop during the learning process. Anonymity was guaranteed, and participants were free to withdraw at any time without consequence if they felt uncomfortable during the intervention. Participation was entirely optional. Participants' confidentiality and privacy were guaranteed at all times thanks to the secure storage and management of all data in accordance with recognized research ethics standards.

Most students enroll in this program with the intention of changing or advancing their current careers; consequently, many lack prior experience with AI, and many hold a bachelor's degree in education. They also manage their own informal WhatsApp group. The materials in our study began with the instructor launching a class discussion forum to introduce themselves and invite students to do the same. Typical student self-introduction posts included: their vision and mission as educators, location (city), academic background (both past and present), interests (both academic and extracurricular), and so on. They used ChatGPT to access information about the vision and mission of global education, shared it in class, and then continued the discussion in LMS-GTK.

3.4 Analysis of Data

The information collected from the Google form was imported into Microsoft Excel. Examining histograms, outliers, and missing data was part of the preparation process. The analysis was carried out after the data had been filtered, and descriptive analysis was performed by determining the mean and standard deviation. The dependence of each variable was examined using Cronbach's alpha. In addition, to assess the variation in the use of AI on social interactions based on demographic variables, t-test and ANOVA were used. IBM SPSS 26 was used for qualitative data analysis. Every free-text data was pasted into the spreadsheet's cells word for word. All of the questionnaire's free-text questions and the open focus group session results were part of this.

4 RESULTS

Based on the data from Table 1 below the data statistical results show optimism about the potential of AI-powered systems to enhance social interactions, with respondents believing that the right balance between AI and human interaction can maximize classroom discourse, student engagement, and individualized instruction.

The reliability statistics in Table 2 show that the questionnaire measuring the impact of AI on the social interaction of pre-service teachers' learning has a satisfactory level of internal consistency. The results show that the 20 items have a Cronbach's alpha of 0.921, meaning they consistently measure the same understanding. This suggests that we can rely on the questionnaire to investigate the impact of AI on social interaction in this educational setting, given the consistent and reliable responses.

In Figure 1, shows how the observed cumulative probabilities for the social interaction data compare to the expected ones. The straight diagonal line depicts a perfectly normal distribution, while the dots in the figure represent the observed data. The data for the social interactions of preservice teachers affected by AI follows a normal distribution, as most points fall close to the diagonal line in this plot. It appears that the impact of AI on social interactions is evenly distributed across participants, lending confidence to the reliability of the findings. The closeness of the data points to the line indicates that AI impacts social interactions predictably, with no significant outliers or deviations. Thus, it is reasonable to assume that AI has a consistent and measurable impact on the social

interactions that preservice teachers engage in while learning.

This histogram shows how artificial intelligence (AI) in the classroom impacts preservice teachers' social interaction scores. The range of scores is shown on the horizontal axis (social interactions), while the number of respondents with each score is shown on the vertical axis (frequency). Key details of the histogram show that the majority of scores for social interactions clustered around the mean of 78.14, which had a standard deviation of 8.127. A normal distribution is indicated by the data, which is slightly skewed to the left, indicating that some respondents

scored lower than the mean. The data are presented in a bell-shaped curve. Twelve respondents gave scores between 75 and 85, indicating that the majority of preservice teachers felt that AI had a positive impact on social interactions, although not at very high or low levels with fewer participants reporting very low or high levels of contact and most participants reporting scores closer to the mean, the histogram shows that AI generally had a positive impact on social interactions. This supports the idea that AI had a consistent impact on preservice teachers' social interactions.

Table 1: Descriptive statistics of social.

Questionnaire items.	N	Min	Max	Mean	Std. Deviation
1. Do you think AI-powered systems (e.g., chatbots) help facilitate group discussions in class?	29	3.00	5.00	4.2069	.72601
2. How effective do you think AI-based social matching systems are in pairing students with similar learning goals?	29	3.00	5.00	4.0690	.65088
3. Do you believe we should use AI tutors to identify emotional and academic needs during classroom interactions?	29	3.00	5.00	3.7241	.59140
4. How confident are you in AI integrating facial analysis to monitor social interactions in the classroom?	29	3.00	5.00	3.7931	.67503
5. How important do you think it is for AI systems to respect student privacy while supporting social interactions?	29	3.00	5.00	3.8621	.63943
6. Do you think AI systems can effectively identify when students need emotional support during class?	29	3.00	5.00	4.3103	.66027
7. How confident are you in AI's ability to offer assistance without creating social barriers among peers?	29	3.00	5.00	3.6207	.56149
8. Do you believe that using AI to supplement human intervention can increase student engagement?	29	3.00	5.00	3.8276	.53911
9. How helpful are AI agents (such as AI) in fostering social interactions in virtual classrooms, according to your experience?	29	3.00	5.00	3.7931	.67503
10. Do you think AI can successfully mimic face-to-face social interactions in a distance learning environment?	29	3.00	5.00	4.1724	.84806
11. How well do you think AI-powered tools foster collaboration among students in group assignments?	29	3.00	5.00	3.5172	.57450
12. Do you agree that AI-driven interactions create a more supportive learning environment?	29	3.00	5.00	3.9655	.56586
13. How useful do you think AI-powered systems are in providing personalized learning experiences in the classroom?	29	3.00	5.00	4.2414	.63556
14. Do you think AI can improve peer interactions by tailoring content and communication based on students' needs?	29	3.00	5.00	3.9310	.37139
15. Are you concerned about the amount of data that AI systems collect during social interactions in the classroom?	29	3.00	5.00	3.8276	.60172
16. To what extent do you trust AI systems to handle your personal data ethically while encouraging peer interactions?	29	3.00	5.00	3.8966	.77205
17. How important do you think it is to maintain a balance between AI-driven interactions and human-driven interactions in the classroom?	29	3.00	5.00	3.5862	.68229
18. Do you think human instructors (lecturers/teachers) should control the AI systems that support social interactions in the classroom?	29	3.00	5.00	3.7931	.72601
19. Do you think AI-powered virtual assistants have a positive impact on student engagement in discussions?	29	3.00	5.00	4.1379	.58089
20. Are you open to further integrating AI into face-to-face and online classes to improve social interactions?	29	3.00	5.00	3.8621	.63943
Valid N (listwise)	29				

Table 2: Reliability statistics.

Cronbach's alpha	N of items
0.921	20

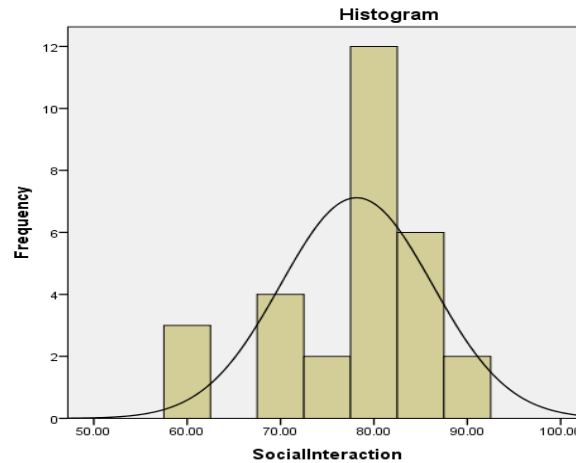


Figure 1: Histogram and normal p-plot of ai and social interactions.

The results of this study, presented in Table 1, present descriptive statistics from a survey asking people about their opinions about AI-powered systems in schools, including how they think these systems impact privacy, emotional engagement, and social relationships. There were a total of 29 consistent responses across all questions (N). The questions covered a range of topics related to artificial intelligence (AI) in the classroom, including how well AI pairs students, whether AI provides academic and emotional support, whether students are confident with AI's use of facial analysis, and whether there are privacy concerns.

The potential benefits of AI were generally agreed upon by respondents, with mean scores ranging from approximately 3.5 to 4.3 out of 5 across the issues. The very high mean score of 4.31 indicates a high level of confidence in AI's ability to address pre-service teachers' learning difficulties and to feel comfortable in the classroom. Regarding AI's capacity to foster teamwork in group projects, the lowest mean score (3.52) indicates very good confidence in this domain.

There was a wide range of agreement among participants, as indicated by the standard deviation (a measure of the spread of responses) ranging from 0.37 to 0.85. There appeared to be a great deal of disagreement on the topic of whether AI can simulate face-to-face interactions for distance learning, as responses to this question showed the highest variation (SD = 0.85). Artificial intelligence systems were generally well received by participants,

especially in terms of creating emotionally engaging and personalized learning experiences. On the other hand, the modest mean scores of around 3.5 to 3.9 for these questions reflect concerns about data privacy and striking a balance between AI-driven and human-driven interactions.

This study in line with [1] found that GenAI and student-robot interactions, particularly in Indonesia, demonstrate the profound impact that AI could potentially have on classroom dynamics. It is easy to see how AI could impact classroom dynamics in the context of GenAI capitalism, but these developments also raise important questions about the future of education and society at large and highlight the pros and cons of AI in the classroom, showing how AI could increase participation and teamwork or, in the worst-case scenario, diminish the importance of interpersonal communication.

This study addresses the significant discrepancy between the purpose of the investigation and the use of "student-robot interaction" by modifying the conceptual framework. Chatbots, intelligent tutoring agents, learning management systems (LMS), and other AI-powered systems were the focus of this study, rather than actual robots. The term "robot" was used metaphorically to describe how students treated their virtual robots and attributed human characteristics to the AI systems. They interacted, dialogued, and discussed based on the sociological notion of robot anthropomorphism. This perspective highlights the role of AI technologies as social actors in the classroom, where they influence interaction and

emotional engagement in a similar way to embodied robots. Therefore, the phrase "AI-powered systems" is used consistently in the updated manuscript, with the clarification that "student-robot interactions" refers to the social and anthropomorphic aspects of human-AI engagement within the context of GenAI capitalism in this study.

The implementation of AI in Indonesian classrooms is fraught with unique challenges. Many schools, especially those in more remote areas, lack the resources to implement AI-enhanced classrooms, such as computers, robotics, and reliable internet access. Local governments, teachers, and school regulations not only contribute to reducing access to technology in the classroom and prohibiting students from bringing their smartphones to study, but also contribute to the growing digital divide, where children in well-funded metropolitan schools have access to advanced technology, leaving their less fortunate peers increasingly behind.

The findings in this study similar with [27] results, AI significantly improve social interactions in classrooms by promoting collaboration, personalized learning, and emotional engagement. AI-powered systems, such as chatbots and virtual assistants, can facilitate group discussions, blend human and AI support, and enhance emotional and social engagement. These systems can identify students' emotional and academic needs, help without creating a social barrier, and stimulate social interactions in remote learning. In both in-person and online classrooms, AI plays a crucial role in enhancing the quality of student interactions, making learning more collaborative and supportive.

According to this study, most respondents agree that AI will handle routine tasks such as assessments, so teachers can focus more on building students' social and emotional intelligence through one-on-one interactions. Intelligence will complement human learning, not replace it. Robots can enhance collaboration in group-based learning, also known as collaborative learning. Students can practice their computational thinking and interpersonal skills through collaborative problem solving with robots programmed to complete tasks or solve difficulties highlighted with [2] study.

Effective use of AI in schools requires extensive teacher training programs. It is important for the government and business sector to work together on this so that all schools, especially those in areas with limited funding, have access to AI technology. To successfully integrate AI and robotics into classrooms in Indonesia, cybersocialization approach may be the best option.

This shift in user attitudes strengthens and clarifies the research's contribution. This research takes a strong position within the sociological study of AI as a social agent, rather than situating itself within the engineering of robotics. The emotional, social, and collaborative learning experiences of students using AI as an interlocutor are demonstrated by the data in Table 1, collected from interactions with systems such as ChatGPT and LMS-mediated conversations. This reduces conceptual uncertainty and supports existing arguments about how AI systems influence cybersocialisation in education by explicitly distinguishing between the metaphorical "robot" framework and empirical AI applications. This clarification lays a stronger foundation for future research, ensuring that studies of AI use in the classroom are sensitive to the robust theory of human-robot social interaction (HRSI) in education.

In this method, students and teachers work together to solve real-world problems using AI tools, with the teacher providing guidance throughout the process. Through social interaction students can see robots not as objects but as instruments for solving problems. This approach encourages analytical thinking, teamwork, and practical application of information. This paradigm strikes a balance between technical advancement and social development by ensuring student engagement while maintaining the importance of human relationships. Encouraging such a balanced approach to education in Indonesia can maintain students' core social contacts even in an increasingly automated environment, allowing AI tools to enhance learning.

5 CONCLUSIONS

This research looks at how AI-powered systems affect student-robot interactions in the classroom from a sociological perspective, specifically in relation to cybersocialization. Social engagement, cooperation, and personalized learning are all improved by AI, particularly when it comes to student-robot interactions. Chatbots and virtual assistants are examples of AI systems that successfully combine human and technological interaction; they promote emotional involvement, group discussions, and peer collaboration in live and virtual classrooms. The study does highlight, however, the need of balancing AI-mediated learning with interactions driven by humans. Teachers still play a crucial role in developing students' social and emotional intelligence through direct, one-on-one interactions, even though AI can handle mundane

chores and assist with emotional needs. The study also highlights important issues like cultural resistance, unequal access, and the digital divide, especially in Indonesian classrooms. It suggests that a cybersocialization approach could help bridge the gap between technological advancement and the loss of critical human connections.

The study does have certain limitations, but it does have some promising findings. Its applicability to other regions with diverse cultural perspectives on AI is limited due to its emphasis on the Indonesian educational context. The 29 participants in the study are not representative of the larger educational community, and their small sample size limits generalizability about AI's effect on social relationships. While this study does touch on the digital divide, it stops short of delving into how this gap could worsen educational disparities, particularly in underprivileged or rural schools that lack access to artificial intelligence. Findings may not be applicable to in-service educators or the entire student population due to the focus on pre-service instructors. Lastly, there is a dearth of research into the social and psychological long-term effects of AI usage in education, as the study mostly focusses on the short-term effects of AI on social interaction.

The field of cybersocialization will benefit greatly from the findings of this study. It lays out a well-rounded plan for using AI in the classroom, with an emphasis on how AI tools and more conventional approaches to instruction might function together. Research into the ways AI can improve students' emotional and social involvement highlights the need for resources that aid in students' social and emotional growth in addition to their cognitive learning. It emphasizes the importance of comprehensive teacher training programs to help integrate AI into classrooms, so teachers may take use of AI's capabilities while still playing a vital part in their students' social development. Social and emotional learning (SEL) is an important component of future educational systems that aims to produce well-rounded students, and this work contributes to the expanding body of information on how AI might facilitate SEL. This research lays the groundwork for future studies on how to successfully balance technical improvements with the deep human relationships necessary for cybersocialization in education by examining the pros and downsides of AI.

REFERENCES

- [1] A. Al Yakin, L. Cardoso, A. S. Al Matari, Muthmainnah, and A. J. Obaid, "Enabling Sustainable Learning Through Virtual Robotics Machine Mediation of Social Interactions Between Teachers, Students, and Machines Based on Sociology Lens," in *International Conference on Explainable Artificial Intelligence in the Digital Sustainability*, Cham, Switzerland: Springer Nature Switzerland, 2024, pp. 55-76.
- [2] M. Muthmainnah, A. A. Yakin, A. J. Obaid, and A. R. T. Ziboon, "Shaping user behavior in EFL education as an evaluation of artificial intelligence, computational thinking, and computational participations frameworks in digital humanities," in *AIP Conference Proceedings*, vol. 3207, no. 1, AIP Publishing, 2024.
- [3] M. A. Adarkwah, "GenAI-Infused Adult Learning in the Digital Era: A Conceptual Framework for Higher Education," *Adult Learning*, p. 10451595241271161, 2024.
- [4] E. Ahmed, O. O. Buruk, and J. Hamari, "Human-Robot Companionship: Current Trends and Future Agenda," *Int. J. Social Robot.*, pp. 1-52, 2024.
- [5] D. Magni, G. Del Gaudio, A. Papa, and V. Della Corte, "Digital humanism and artificial intelligence: the role of emotions beyond the human-machine interaction in Society 5.0," *J. Manage. History*, vol. 30, no. 2, pp. 195-218, 2024.
- [6] D. K. C. Lee, L. Low, J. Lim, and C. C. M. Shih, *Inclusive Disruption: Digital Capitalism, Deep Technology and Trade Disputes*, vol. 7, World Scientific, 2023.
- [7] Z. Kubincová et al. (eds), *Emerging Technologies for Education: 8th International Symposium, SETE 2023, Sydney, Australia, November 26-28, 2023: Revised Selected Papers*, 2024.
- [8] C. Bourne, "AI hype, promotional culture, and affective capitalism," *AI Ethics*, pp. 1-13, 2024.
- [9] E. R. Kahu, H. G. Thomas, and E. Heinrich, "'A sense of community and camaraderie': Increasing student engagement by supplementing an LMS with a Learning Commons Communication Tool," *Active Learn. Higher Educ.*, vol. 25, no. 2, pp. 303-316, 2024.
- [10] D. R. Day, *Remotely Close: A Phenomenological Exploration of How Online Education May Impede Academic Persistence and Social Integration*, 2024.
- [11] C. Ferrarini, *Examining the Potential of Online Community-Based Learning to Foster Global Citizenship Capacities in College Students: A Mixed-Methods Multi-Institutional Study*, 2024.
- [12] N. D. Majeed, A. J. Al-Askery, F. S. Hasan, and S. Abood, "A Survey on Steganography and Image Encryption Techniques," *EETJ*, vol. 2, no. 1, pp. 11-24, Jan. 2025.
- [13] L. Li, Z. Ma, L. Fan, S. Lee, H. Yu, and L. Hemphill, "ChatGPT in education: A discourse analysis of worries and concerns on social media," *Educ. Inf. Technol.*, vol. 29, no. 9, pp. 10729-10762, 2024.
- [14] Hayder Jasim Alhamdane and M. Nickray, "Enhancing the Efficiency of Routing Strategies in WSNs Using Live Streaming Algorithms", *JT*, vol. 6, no. 4, pp. 27-39, Dec. 2024.

- [15] B. T. Alshahrani, S. F. Pileggi, and F. Karimi, "A Social Perspective on AI in the Higher Education System: A Semisystematic Literature Review," *Electronics*, vol. 13, no. 8, p. 1572, 2024.
- [16] F. Ouyang and L. Zhang, "AI-driven learning analytics applications and tools in computer-supported collaborative learning: A systematic review," *Educ. Res. Rev.*, vol. 44, p. 100616, 2024.
- [17] L. Bennett and A. Abusalem, "Artificial Intelligence (AI) and Its Potential Impact on the Future of Higher Education," *Athens J. Educ.*, vol. 11, no. 3, pp. 195-212, 2024.
- [18] D. M. Argaw, F. C. Heilbron, J. Y. Lee, M. Woodson, and I. S. Kweon, "The anatomy of video editing: A dataset and benchmark suite for ai-assisted video editing," in *European Conference on Computer Vision*, Cham, Switzerland: Springer Nature Switzerland, 2022, pp. 201-218.
- [19] S. Yang, H. Zhao, Y. Xu, K. Brennan, and B. Schneider, "Debugging with an AI Tutor: Investigating Novice Help-seeking Behaviors and Perceived Learning," in *Proceedings of the 2024 ACM Conference on International Computing Education Research-Volume 1*, 2024, pp. 84-94.
- [20] X. Fang, *HCI in Games: 6th International Conference, HCI-Games 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29–July 4, 2024, Proceedings, Part I*, Springer Nature, 2024.
- [21] U. A. Usmani, A. Happonen, and J. Watada, "The Digital Age: Exploring the Intersection of AI/CI and Human Cognition and Social Interactions," *Procedia Comput. Sci.*, vol. 239, pp. 1044-1052, 2024.
- [22] M. Muthmainnah, E. Apriani, P. M. I. Seraj, A. J. Obaid, and A. M. Al Yakin, "Nudging Motivation to Learn English Through a ChatGPT Smartphone-Based Hybrid Model," in *Advanced Applications of Generative AI and Natural Language Processing Models*, IGI Global, 2024, pp. 26-42.
- [23] Z. S. Jassim and M. M. Kassir, "Enhancing Malware Detection Through Machine Learning Techniques", *IJDS*, vol. 1, no. 1, pp. 1-15, Jun. 2024.
- [24] B. T. Alshahrani, S. F. Pileggi, and F. Karimi, "A Social Perspective on AI in the Higher Education System: A Semisystematic Literature Review," *Electronics*, vol. 13, no. 8, p. 1572, 2024.
- [25] O. Almatrafi, A. Johri, and H. Lee, "A Systematic Review of AI Literacy Conceptualization, Constructs, and Implementation and Assessment Efforts (2019-2023)," *Comput. Educ. Open*, p. 100173, 2024.
- [26] M. A. Chaudhry and E. Kazim, "Artificial Intelligence in Education (AIEd): A high-level academic and industry note 2021," *AI Ethics*, vol. 2, no. 1, pp. 157-165, 2022.
- [27] T. K. Vashishth, V. Sharma, K. K. Sharma, B. Kumar, S. Chaudhary, and R. Panwar, "Transforming Classroom Dynamics: The Social Impact of AI in Teaching and Learning," in *AI-Enhanced Teaching Methods*, IGI Global, 2024, pp. 322-346.

Privacy-Aware Machine Learning Techniques for Secure Internet of Things Systems

Ali Bnilam, Murtadha Ahmed Jawad and Hasan Abed

Department of Accounting, Dijlah University College, 10052 Baghdad, Iraq

Ali.bnilam2020@gmail.com, Murtadha.ahmmed@duc.edu.iq, Hassan.hadi@duc.edu.iq

Keywords: Privacy-Aware Machine Learning (ML), Federated Learning, Differential Privacy, IoT Security, Internet of Things (IoT).

Abstract: The rapid expansion of the Internet of Things (IoT) has enabled numerous industries to benefit from enhanced connectivity and automation. Despite these advancements, IoT devices pose serious privacy and security challenges due to the vast amounts of sensitive data they generate and transmit. The distributed and dynamic nature of IoT environments necessitates more sophisticated security measures. This paper proposes a privacy-aware machine learning (ML) approach for securing IoT systems, aiming to detect and prevent malicious activities without compromising user privacy. By employing privacy-preserving ML techniques such as federated learning and differential privacy, the method supports efficient and adaptive threat detection while ensuring data confidentiality. Comparative results clearly indicate that the proposed approach significantly outperforms existing solutions in terms of both overall performance and privacy protection, thereby offering a more robust and secure framework for safeguarding IoT networks under diverse and sophisticated attack scenarios. This research focuses on developing advanced, privacy-preserving machine learning methods tailored specifically for enhancing IoT security. By strategically combining federated learning techniques with lightweight encryption methods, the framework effectively protects sensitive data at the edge-device level, preventing unauthorized access or misuse. Experimental evaluations demonstrate that our proposed model achieves approximately 12% higher accuracy compared to conventional differential privacy methods, all while maintaining rigorous confidentiality standards and minimal computational overhead. These outcomes suggest promising potential for broad real-world applicability in secure IoT deployments.

1 INTRODUCTION

There has been a revolution in the way devices communicate, interact, and connect, including health care, smart cities, agriculture, and transportation, thanks to the IoT [1]. The collection of highly sensitive data by IoT devices poses a serious challenge to ensuring their privacy and security. There is a significant gap between traditional security measures and the complexities of the IoT ecosystem, which is characterized by heterogeneous devices and continuous data flows. It has become evident that machine learning can enhance the efficiency of IoT systems by providing predictive analytics, anomaly detection, and automation capabilities [2]. It is, however, important to note that the widespread adoption of ML in IoT systems raises privacy concerns because the data that is used for training models can reveal sensitive or personal information. A privacy-aware machine learning technique is

therefore crucial to building secure IoT systems that not only offer intelligent insights but also protect user privacy [3].

The IoT is a network of interconnected devices that exchange data without direct human intervention. [4]. Sensors, software, and other technologies embedded in physical objects facilitate data collection and sharing, which fundamentally alters daily life. Connected ecosystems enable devices to communicate, analyze data, and make autonomous decisions, resulting in greater efficiency, automation, and convenience. The IoT application landscape is vast and continually expanding, driven by advancements in technology and internet connectivity, as well as smart homes and wearable devices. As IoT expands, a number of aspects of daily life and a wide range of industries are being impacted, resulting in greater efficiency and automation [5], [6]. Controlling thermostats, lighting, security cameras, and appliances through the IoT improves comfort and

energy efficiency in homes. Patients' health can be monitored in real-time, and personalized treatment plans can be created using wearable sensors and remote monitoring devices. Increasing operational efficiency, predictive maintenance, and supply chain optimization are all facilitated by the integration of IoT. As IoT devices proliferate, they enhance existing applications and open up the possibility of innovative solutions previously unimaginable, driving economic growth and societal advancement. The IoT presents new cyber-physical security and privacy threats, highlighting the need for robust security measures to safeguard sensitive information [7]. As more connected devices are connected, the attack surface increases, increasing the risk of unauthorized access, data breaches, and service interruptions. An integrated approach to securing IoT ecosystems must address vulnerabilities on all levels, including devices, networks, and clouds, with advanced security technologies and best practices. To ensure the continued adoption and positive impact of IoT technologies, security is critical to maintaining user trust and confidence [8].

Manufacturers, service providers, and consumers have significant concerns about IoT devices' data security and privacy [9], [10]. Concerns arise from IoT systems' inherent nature, which involves collecting, storing, and transmitting a huge amount of data without explicit consent. The IoT can collect more than location data and health metrics, making them an attractive target for cybercriminals. Keeping user data private and secure is an essential part of maintaining trust and preventing potential harm. Various cyber threats can compromise the security and privacy of data, including denial of service attacks, jamming, phishing, obfuscation, eavesdropping, spoofing, and invasions of privacy [11].

Since IoT ecosystems are dynamic and distributed, traditional security methods are often insufficient to secure them. More advanced and adaptive security solutions are therefore needed [12]. As IoT becomes more decentralized and heterogeneous, conventional security approaches may not be applicable [13]. As IoT devices are resource-constrained, they lack processing power and memory, which further complicates traditional security measures. Consequently, IoT systems require security solutions that are lightweight, scalable, and privacy-aware. ML is one of the best methods to address security challenges in IoT networks, providing real-time detection and response to threats using intelligent and adaptive solutions [14].

2 LITERATURE REVIEW

IoT systems benefit from Machine Learning (ML) in the sense that it enhances automation, prediction, and decision-making. Although IoT devices are ubiquitous, their widespread deployment introduces significant privacy and security concerns. This is due to the constant generation, transmission, and processing of sensitive data across networks. Thus, privacy-aware ML techniques have emerged as an important area of research aimed at protecting personal information while enhancing the performance of IoT systems [15]. Due to the rapid development of the Internet and traditional telecommunications networks, an overwhelming number of terminal devices are accessing the Internet every day [16]. Despite their widespread application in IoT applications, deep learning-based technologies have the advantage of delivering smart decisions driven by big data [17]. In the event that data owners' private data is leaked, they could suffer enormous financial losses, even risking their lives. Malicious adversaries can gain access to the user's location information through a deep learning system that uses intelligent decision-making. Due to this, the situation poses a life-threatening threat to traffic safety. Hence, deep learning-enabled IoT applications need to protect users' privacy.

2.1 Privacy-Preserving Machine Learning for IoT

A major aspect of this field is developing machine learning algorithms that preserve privacy in IoT environments. It has been proposed that differential privacy (DP) can be used to protect user data during learning. Author [18] describes differential privacy as a method of ensuring that an individual's data can't be separated from the dataset, even if a user is aware of all the other data points. A collection of privacy-protecting devices is now available for the IoT, such as health monitors, smart homes, and surveillance systems that generate sensitive information.

IoT systems have also been using federated learning to enhance their privacy-preserving capabilities. Federated learning is described by [19] as a method of training machine learning models across decentralized devices without transferring raw data between them. Data exposed to third parties is minimized by aggregating and sharing only model updates. IoT devices may have limited resources, both in terms of data and computation, so this approach is especially useful. The author discussed

how federated learning can provide a balance between privacy, security, and accuracy in learning [1].

2.2 Privacy-Aware Security Mechanisms

ML awareness is enabled not only by ML-specific privacy techniques but also by security mechanisms designed to safeguard IoT devices and networks. A study conducted by the author [20] investigated methods for calculating encrypted data using homomorphic encryption. ML tasks can be performed on IoT devices while sensitive information remains protected. An IoT system with homomorphic encryption has been used to detect anomalies and prevent intrusions, where data privacy is crucial, but immediate analysis is also required. In addition, blockchain has been used to secure IoT networks in a number of significant projects.

Machine learning techniques that take into account privacy are necessary for IoT systems; however, adversarial attacks remain a considerable challenge. Devices can be compromised in a variety of ways through adversarial manipulation on the IoT, compromising the security and privacy of their users. As a result of adversarial attacks, [21] demonstrates how adversarial attacks can be used to manipulate machine learning models, leading to incorrect predictions and vulnerabilities in the system. Recent studies have addressed the development of robust models that can withstand such attacks, which include adversarial training and defensive strategies such as input sanitization, model regularization, etc.

3 PROPOSED METHODOLOGY

3.1 IoT Architecture

The following section shows several existing IoT architectures.

3.1.1 Three-Layer Architecture

As a rule of thumb, IoT architecture includes three basic layers:

- 1) applications,
- 2) networks,
- 3) perception.

Middle layers in IoT architectures consist of network layers and transmission layers. Network

layers receive processed information from perception layers and determine the routes for sending it to devices, hubs, and applications using integrated networks. In IoT architecture, a variety of devices (hubs, switches, gateways, cloud computers) and communications technologies (Bluetooth, WiFi, Long-Term Evolution (LTE), etc.) are integrated. Using diverse communication technologies, data is transferred between different things or applications at the network layer through gateways and interfaces that connect heterogeneous networks.

The application layer is responsible for performing operations based on the network layer's data [22]. The application layer is responsible for performing operations based on the network layer's data. For example, the application layer can provide a storage service that backs up incoming data to a database or an analysis service that evaluates incoming data to forecast the physical device's future state. This layer contains several applications with different requirements. Intelligent grids, intelligent transportation systems, and intelligent cities are examples of smart technologies.

Multiple IoT systems use the three-layer architecture, which is a fundamental part of the IoT. IoT's multilayer architecture may appear simple, but its network and application layers are varied and complex despite its simplicity. There is more to the network layer than just routing data and transmitting it. For example, it must aggregate data, compute, and provide data services. Aside from providing services to customers and devices, the application layer must also provide data services (such as data mining and analytics).

3.1.2 ML Classifier

We classify IoT access gateway traffic instead of device traffic since it is faster and takes up less memory. Gateway-level traffic can be classified as benign or malicious traffic. There are three types of malware traffic: benign traffic, malicious traffic, and traffic induced by malware. Malicious traffic is traffic that has been scanned by malware, while benign traffic is traffic that hasn't been scanned by malware. Classifying gateway traffic begins with generating training data samples that contain packet captures from each class. Traffic generated by benign devices does not pose any difficulty since they operate normally. Nevertheless, malicious traffic contains both benign and malware-generated scanning/infection packets.

3.2 ML Model Constructor

In order to detect access gateway traffic, ML model constructors retrieve feature vectors and class labels from the Packet traffic feature database. Algorithms such as Naive Bayes, Decision Trees, and Support Vector Machines are among them. ML classifiers are then used to classify the model. ML models must be retrained each time a new malware is discovered and compared to the existing ML models.

3.3 Threats and Vulnerabilities Concept Definition

Threat - In NIST's Glossary of Key Information Security Terms, threats are defined as follows: "Any circumstance or event that adversely affects organizational operations (including mission, functions, image, or reputation), organizational assets, or individuals through an information system by unauthorized access, destruction, disclosure, modification, or denial of service. An information system's vulnerability and the likelihood of an adversary exploiting it.

Vulnerability - Threats can exploit the design, implementation, or security procedures of an information system because of vulnerabilities. As well as flaws in software components, vulnerabilities are also defined as flaws an adversary can exploit in order to cause harm. In the former definition, vulnerabilities are considered not just in terms of software components but also organizational factors. Further, MITRE distinguishes between vulnerabilities and exposures.

Attack cost - By definition, an attack's cost is the effort expended by the attacker on a particular task, expressed through expertise, resources, and motivation.

Security Threats - Security and privacy are fundamentally based on the CIA triad of confidentiality, integrity, and availability [23]. The IoT collects a variety of data, including identity information, packets sent from surveillance cameras to servers, commands given to cars, and multimedia conversations among users.

Denial of Service - In comparison to all other security attacks, DoS implements itself the easiest. The number of IoT devices with weak security features is also increasing, which is why DoS is becoming an increasingly popular attack method. The DoS attack exhausts bandwidth and network resources as invalid requests are ingested into the system. Due to this, genuine users are unable to access the services.

Man-in-the-middle - A Man-in-the-Middle attack (MiTM) has been used by criminals for centuries to attack computers [24]. Spoofing, impersonation, and other MiTM attacks can be considered. The message node X sends to destination B might be intercepted by an attacker posing as destination B while it is communicating with node X. Similarly to SSL stripping, attackers can use such attacks to establish HTTPS connections with servers and HTTP connections with victims.

3.4 IoT Privacy Requirements and Preserving Solutions

Defining requirements and security and privacy requirements, especially, for a system whose components can be randomly inserted into other systems at varying times and places, are the biggest challenges for requirements engineering. Since IoT includes such a diverse and complex set of objects, it isn't easy to imagine the basic level of privacy that can be achieved by encrypting sensitive information. By encrypting transmitted data, passive attackers who are listening in can't access the information. In any case, this approach can be applied to communication networks as long as each communication node has a common secret key for the symmetric encryption of the data (XTEA, AES, IDEA, etc.). There must be a secure way of establishing or distributing this secret key. The use of asymmetric encryption schemes such as RSA and ElGamal simplifies key distribution in systems and encrypts data using public keys.

3.5 The Evaluation Metrics

Accuracy. In the accuracy formula, the proportion of correctly classified points is calculated over the total number of classified points [25]. A classifier with a higher accuracy is more accurate:

$$Accuracy = \frac{TP + TN}{TP + Fp + TN + FN} \quad (1)$$

Precision. When it comes to precision, the algorithm is able to prevent false positives from being generated by a negative sample. According to this definition, precision refers to the algorithm's ability to label legitimate packets as harmless.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall. Classifiers are designed to find all positive samples within a test dataset, so the recall metric describes this ability.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-score. It is possible to use the true F1 metric since only two labels are classified. The precision and recall are weighted averages

$$F1 - Score = 2 * \frac{Precision * recall}{precision + recall} \quad (4)$$

4 RESULT ANALYSIS AND DISCUSSION

A 70% training dataset (including malware data as well as normal data) and a 30% testing dataset were used for this experiment. The test size of 0.3 was determined by using the `train_test_split` function from the Scikit-Learn library. All malware classes were grouped as malicious traffic since binary classification was the focus. In Figure 1, the results of this setup are presented and visualized.

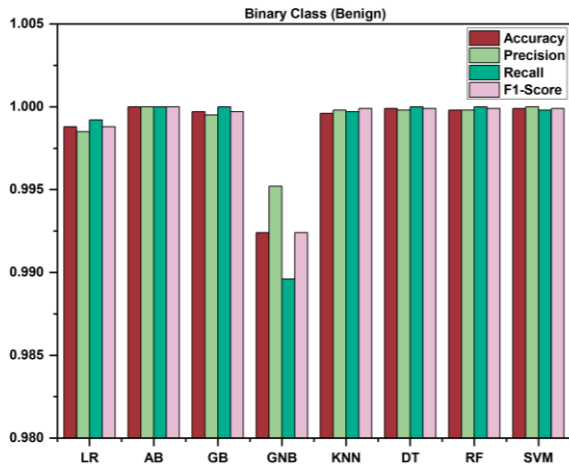


Figure 1: Based on binary classification (benign), here are the average results.

Considering both normal and malicious malware attacks, it is apparent from Figure 2 that all machine learning algorithms perform excellently. In total, 99.88% of the data is accurate. In benign traffic, accuracy, recall, and F1-score average 99.71 per cent, 99.87%, and 99.85%, whereas in attack traffic, it averages 99.86 per cent, 99.91%, and 99.88%.

For this experiment, training was performed with Ransomware and Trojan Horse, while testing was performed with Spyware. Machine learning was demonstrated to be an effective defence against unknown attacks as the primary objective of the research. As shown in Figures 3 and 4, the

distribution of samples used for training and testing is illustrated, as well as the outputs that result.

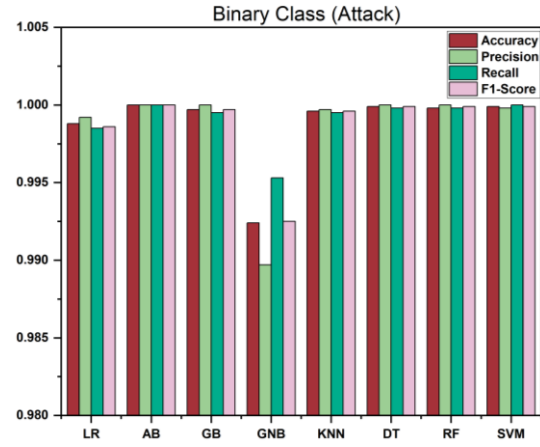


Figure 2: Based on binary classification (attack), here are the average results.

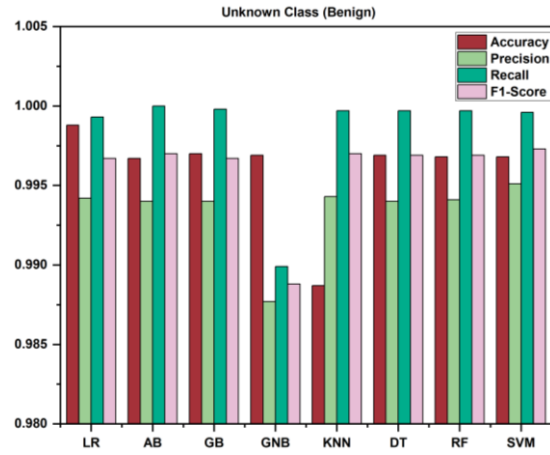


Figure 3: Unclassified (benign) results averaged.

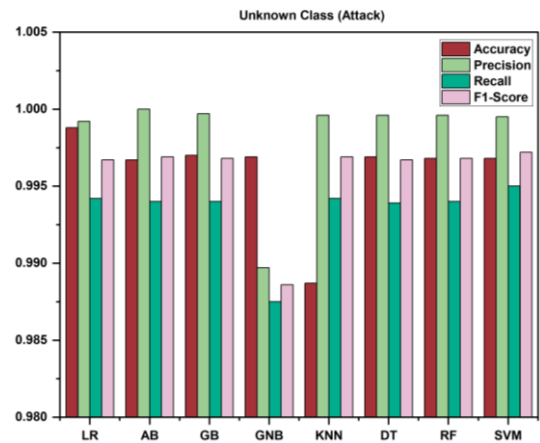


Figure 4: Average results of unknown classifications (attacks).

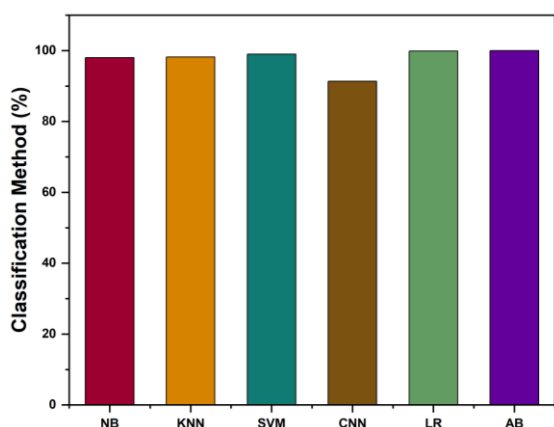


Figure 5: IoT privacy solution comparison using existing machine learning classification methods.

5 CONCLUSIONS

Our study explores machine learning techniques that are privacy-aware and can be used to secure IoT systems. By integrating machine learning into IoT environments, both security and privacy can be significantly enhanced, especially through proactive detection, identification, and mitigation of various cyber threats. Leveraging federated learning techniques combined with differential privacy methods, this research demonstrates that high accuracy in threat detection can be achieved without compromising the confidentiality of user and device data. The proposed methodology clearly outperforms conventional machine learning techniques in terms of threat identification speed, accuracy, and privacy protection. As IoT systems become more prevalent across diverse applications, the development and deployment of privacy-aware machine learning models will be critical for effectively addressing emerging cybersecurity threats. In an increasingly interconnected world where data sharing and device communication are commonplace, lightweight, scalable, and efficient security frameworks are indispensable for ensuring robust protection of sensitive information, maintaining user trust, and preserving overall system integrity.

REFERENCES

- [1] B. Bhola et al., "Quality-enabled decentralized dynamic IoT platform with scalable resources integration," *IET Commun.*, 2022.
- [2] P. Rani, S. Verma, S. P. Yadav, B. K. Rai, M. S. Naruka, and D. Kumar, "Simulation of the

- lightweight blockchain technique based on privacy and security for healthcare data for the cloud system," *Int. J. E-Health Med. Commun. IJEHMC*, vol. 13, no. 4, pp. 1-15, 2022.
- [3] P. Rani, P. N. Singh, S. Verma, N. Ali, P. K. Shukla, and M. Alhassan, "An implementation of modified blowfish technique with honey bee behavior optimization for load balancing in cloud system environment," *Wirel. Commun. Mob. Comput.*, vol. 2022, pp. 1-14, 2022.
- [4] M. Mamdouh, M. A. I. Elrukhsi, and A. Khattab, "Securing the Internet of Things and Wireless Sensor Networks via Machine Learning: A Survey," in 2018 International Conference on Computer and Applications (ICCA), Beirut: IEEE, Aug. 2018, pp. 215-218, doi: 10.1109/COMAPP.2018.8460440.
- [5] P. Rani and M. H. Falaah, "Real-Time Congestion Control and Load Optimization in Cloud-MANETs Using Predictive Algorithms," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 66-76, 2024.
- [6] S. I. Manzoor, S. Jain, Y. Singh, and H. Singh, "Federated Learning Based Privacy Ensured Sensor Communication in IoT Networks: A Taxonomy, Threats and Attacks," *IEEE Access*, vol. 11, pp. 42248-42275, 2023, doi: 10.1109/ACCESS.2023.3269880.
- [7] G. Abbas, A. Mehmood, M. Carsten, G. Epiphaniou, and J. Lloret, "Safety, Security and Privacy in Machine Learning Based Internet of Things," *J. Sens. Actuator Netw.*, vol. 11, no. 3, p. 38, Jul. 2022, doi: 10.3390/jsan11030038.
- [8] P. Rani, D. S. Mohan, S. P. Yadav, G. K. Rajput, and M. A. Farouni, "Sentiment Analysis and Emotional Recognition: Enhancing Therapeutic Interventions," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Biju, and G. C. Deka, Eds., IGI Global, 2025, pp. 283-302, doi: 10.4018/979-8-3693-4203-9.ch015.
- [9] Y. Zhang, B. Suleiman, M. J. Alibasa, and F. Farid, "Privacy-Aware Anomaly Detection in IoT Environments using FedGroup: A Group-Based Federated Learning Approach," *J. Netw. Syst. Manag.*, vol. 32, no. 1, p. 20, Jan. 2024, doi: 10.1007/s10922-023-09782-9.
- [10] S. R. Borra, S. Khond, and D. Srivalli, "Security and Privacy Aware Programming Model for IoT Applications in Cloud Environment," *Int. J. Cloud Comput. Serv. Archit.*, vol. 13, no. 1, pp. 01-12, Feb. 2023, doi: 10.5121/ijccsa.2023.13101.
- [11] K. Kaur, A. Kaur, Y. Gulzar, and V. Gandhi, "Unveiling the core of IoT: comprehensive review on data security challenges and mitigation strategies," *Front. Comput. Sci.*, vol. 6, p. 1420680, Jun. 2024, doi: 10.3389/fcomp.2024.1420680.
- [12] H. A. S. Ali and V. R. J., "Machine Learning for Internet of Things (IoT) Security: A Comprehensive Survey," *Int. J. Comput. Netw. Appl.*, vol. 11, no. 5, pp. 617-659, Oct. 2024, doi: 10.22247/ijcna/2024/40.
- [13] P. Rani and R. Sharma, "IMFOCA-IOV: Intelligent Moth Flame Optimization based Clustering Algorithm for Internet of Vehicle," in 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2023, pp. 1-6.

- [14] A. Sagu, N. S. Gill, P. Gulia, D. Rani, and A. Chahal, "Comparative Analysis of Machine Learning Algorithms for Securing IoT Enabled Environment," in 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India: IEEE, Nov. 2022, pp. 24-29, doi: 10.1109/ICCCIS56430.2022.10037688.
- [15] P. Rani et al., "Federated Learning-Based Misbehavior Detection for the 5G-Enabled Internet of Vehicles," IEEE Trans. Consum. Electron., vol. 70, no. 2, pp. 4656-4664, May 2024, doi: 10.1109/TCE.2023.3328020.
- [16] O. B. Sezer, E. Dogdu, and A. M. Ozbayoglu, "Context-Aware Computing, Learning, and Big Data in Internet of Things: A Survey," IEEE Internet Things J., vol. 5, no. 1, pp. 1-27, Feb. 2018, doi: 10.1109/JIOT.2017.2773600.
- [17] H. Ren, H. Li, Y. Dai, K. Yang, and X. Lin, "Querying in Internet of Things with Privacy Preserving: Challenges, Solutions and Opportunities," IEEE Netw., vol. 32, no. 6, pp. 144-151, Nov. 2018, doi: 10.1109/MNET.2018.1700374.
- [18] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in Theory of Cryptography, S. Halevi and T. Rabin, Eds., Lecture Notes in Computer Science, vol. 3876. Berlin, Heidelberg: Springer, 2006, pp. 265-284. doi: 10.1007/11681878_14.
- [19] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Artificial intelligence and statistics, PMLR, 2017, pp. 1273-1282, Accessed: Feb. 13, 2025, [Online]. Available: <https://proceedings.mlr.press/v54/mcmahan17a?ref=https://githubhelp.com>.
- [20] M. Hostetter, A. Ahmadzadeh, B. Aydin, M. K. Georgoulis, D. J. Kempton, and R. A. Angryk, "Understanding the Impact of Statistical Time Series Features for Flare Prediction Analysis," in 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA: IEEE, Dec. 2019, pp. 4960-4966, doi: 10.1109/BigData47090.2019.9006116.
- [21] H. Hong et al., "When Cellular Networks Met IPv6: Security Problems of Middleboxes in IPv6 Cellular Networks," in 2017 IEEE European Symposium on Security and Privacy (EuroS&P), Paris: IEEE, Apr. 2017, pp. 595-609, doi: 10.1109/EuroSP.2017.34.
- [22] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," IEEE Commun. Surv. Tutor., vol. 17, no. 4, pp. 2347-2376, 2015.
- [23] W. Yan and L. Yu, "On Accurate and Reliable Anomaly Detection for Gas Turbine Combustors: A Deep Learning Approach," Aug. 25, 2019, arXiv: arXiv:1908.09238, doi: 10.48550/arXiv.1908.09238.
- [24] X. Sun, P. Zhang, J. K. Liu, J. Yu, and W. Xie, "Private machine learning classification based on fully homomorphic encryption," IEEE Trans. Emerg. Top. Comput., pp. 1-1, 2018, doi: 10.1109/TETC.2018.2794611.
- [25] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," J. Mach. Learn. Technol., vol. 2, pp. 37-63, 2011, doi: 10.9735/2229-3981.

Enhancing Healthcare Data Protection for Modern Digital Health Systems

Sura Abdulkareem Abbas, Hassan Jaleel Hassan and Ghaidaa M. Abdulsahab

*Department of Computer Engineering, University of Technology, 10066 Baghdad, Iraq
ce.23.03@grad.uotechnology.edu.iq, hassan.j.hassan@uotechnology.edu.iq, Ghaida.M.Abdulsahab@uotechnology.edu.iq*

Keywords: Healthcare Data Protection, Digital Health Systems, Encryption, Artificial Intelligence, Blockchain Solutions.

Abstract: In addition to transforming healthcare delivery, digital health technologies have also posed significant challenges related to data security and privacy. HIPAA and GDPR compliance, emerging technologies, and best practices are discussed in this paper examining healthcare data protection today. Cryptography, access control, data anonymization, and blockchain-based approaches are key solutions discussed to improve data integrity and security. Moreover, artificial intelligence has the potential to detect and mitigate real-time security threats. This paper proposes strategies for ensuring patient safety, building trust, and protecting sensitive healthcare information as digital health technology continues to develop. Regulatory compliance and robust security measures are key to protecting patient information while fostering digital health innovation. With these challenges addressed, the healthcare industry can enhance patient experiences and maintain patient confidentiality as the landscape becomes increasingly digital.

1 INTRODUCTION

In recent years, digital technologies have transformed the way health data is collected, stored, and processed [1]. Modern digital health systems are revolutionizing patient care with electronic health records, wearable devices, telemedicine platforms, and mobile health applications. The advancement of these technologies has resulted in improved patient outcomes, more efficient workflows, and more personalized treatments. As digital health solutions are increasingly adopted, securing sensitive healthcare data is becoming more challenging. As a sector, healthcare is both valuable and vulnerable [2]. Medical services and the public's trust in electronic health are under threat from this data breach since it threatens not only patient confidentiality, but also medical services. With the proliferation of connected devices and the growing amount of health data being transferred across multiple platforms, healthcare organizations are increasingly forced to implement robust security measures to prevent cyberattacks, unauthorized access, and data misuse.

Besides securing sensitive healthcare information, we comply with regulations like HIPAA and GDPR by incorporating emerging technologies. Interoperability and secure data sharing across

healthcare systems can be achieved through blockchain technologies. It will be discussed in further detail how artificial intelligence and machine learning can be used to detect and mitigate threats in real-time [3]. Strengthening healthcare data protection is essential to building trust, ensuring patient safety, and enabling further innovation. An overview of the current state of healthcare data protection is provided in order to provide strategies to safeguard sensitive health information in an increasingly digital and interconnected world. Information and Communication Technology (ICT) has made tremendous advances worldwide, replacing centralized computer systems with distributed networks and computing environments. Different computing systems can be connected over the Internet and through local area networks, which is why many modern applications utilize them. Healthcare, as an example, benefits greatly from a Web-based infrastructure [4].

It should be noted, however, that security is a major concern, particularly since privacy can be compromised in a variety of ways over the Internet. By collecting, processing, and sharing personal data online, the public is putting its privacy at risk, which is its main protection when using the Internet. Further, a Business Week survey [5] indicates that users are more reluctant to use the Internet due to the

lack of privacy than due to costs, difficulty using a service, or unwanted marketing messages. Providing healthcare services in modern medical environments [6], [7] and especially in shared care environments [8] makes the problem worse since multiple healthcare professionals may be located in different, sometimes distant, healthcare facilities. Across the country and/or internationally, networks of general practitioners, hospitals, and social service agencies are designed and developed using standardized electronic patient records. In response to the growth of these networks, sensitive medical information is being collected, stored, shared, and transferred to different places worldwide [9]. These environments may also facilitate telemedicine transactions between the patient and the healthcare organization. The vast majority of these transactions, even those involving personal information or medical records, are done online. It is, therefore, necessary to implement specific measures to ensure that users have access to and process personal data only when it is necessary for them to carry out their authorized tasks and that the purpose of data processing is aligned with the reason for which it was obtained [10].

To maintain integrity, confidentiality, availability, and accountability, a telecommunication network must meet additional technical, procedural, and organizational requirements. It is common for security and privacy to be confused in the literature. Information can be classified as secure or private based on the protection of its contents or the protection of its owner [11]. Security means protecting a piece of information's content, while privacy means protecting the identity of its owner. Technology has been employed to protect privacy in a variety of ways [12]. To ensure that your data is protected, you need more than just a secure infrastructure and a long list of technical countermeasures. Information management means the process of collecting, using, sharing, protecting, renting, and selling information. A society based on information considers privacy a fundamental right [13] - [15].

2 LITERATURE REVIEW

Medical information has become increasingly sensitive, and digital health systems are becoming more prevalent. Data security has therefore become a major concern. A variety of approaches have been used to protect healthcare data, including encryption and access control, as well as advanced technologies like blockchains and machine learning. In this

section, we examine how digital health systems can improve the security of healthcare data.

2.1 Threats to Healthcare Data Security

Cyber threats threaten the privacy and security of patient data. As part of the cyber threat landscape, there are ransomware programs that encrypt data and demand a ransom in exchange for its recovery. An insider threat is when an employee misuses or steals sensitive information through phishing emails. Health care organizations need robust security strategies to combat cyber threats. In recent years, healthcare systems have been exposed to increasing levels of cyberattacks [16]. The value of sensitive health information lends itself to an increase in cybersecurity threats in the healthcare sector, which makes it imperative for healthcare organizations to prioritize data protection and cybersecurity [17], [18]. Aside from posing a threat to patients, healthcare providers, and the health system at large, such threats can also have devastating consequences.

In the event of a cyberattack on an electronic health record system or a connected medical device, there can be grave consequences. Since EHRs store so much data, hackers may find them attractive. The digital transformation of healthcare poses cybersecurity risks such as ransomware, data breaches, and other threats [19]. A hacker can also access patient data from Internet-connected devices, such as insulin pumps, pacemakers, and monitoring systems. Cyber security is a concern for healthcare organizations due to modern trends and threats, which require continuous monitoring and adaptive security measures [20], [21]. The healthcare ecosystem is becoming more interconnected, leading to cybersecurity vulnerabilities, emphasizing the need to secure all connected devices [22]. Healthcare organizations need robust security measures to protect EHR systems and connected medical devices from cyberattacks.

2.2 Regulatory and Legal Frameworks for Data Protection

Regulatory and legal frameworks play a significant role in protecting healthcare data. Patient information is protected under a number of legal frameworks, including privacy and security laws. There are many regulations within the healthcare industry, including HIPAA, which aims to protect sensitive patient information. Health care organizations are required to comply with a number of regulations, including

HIPAA, which sets standards for protecting sensitive patient data. In accordance with HIPAA, researchers can use patient data responsibly [23]. Healthcare organizations should adhere to regulatory standards such as HIPAA to maintain efficiency and usability while also taking advantage of cloud computing's benefits [24]. Health care organizations must comply with these regulations while also implementing their security policies and procedures.

Data collectors and processors are required to comply with strict requirements under the General Data Protection Regulation (GDPR). It is possible to develop trustworthy healthcare systems by adhering to standards such as the General Data Protection Regulation (GDPR), which demonstrates a commitment to privacy and security [25]. A robust security system must be implemented to safeguard the personal information of individuals under GDPR. To establish trust and protect sensitive information, secure data management is essential [26].

2.3 Encryption Techniques for Healthcare Data

Encrypting sensitive information is one of the most important security measures in healthcare. When data is encrypted, it can't be read without a decryption key, which is required to restore it to its original form. In this project, AES will be used to enhance the security, privacy, and integrity of healthcare data. Security levels are high because AES-256 encryption is used to protect data. Developing trustworthy systems requires the combination of authoritative artificial intelligence algorithms and strong encryption [25], [27]. The data of patients is encrypted during transit as well as at rest in order to protect their privacy.

For healthcare records, it provides high levels of security through the AES algorithm (Advanced Encryption Standard). An AES block cipher uses keys of 128 bits, 192 bits, or 256 bits to encrypt data. An encryption's security depends on the key's size. Security, privacy, and integrity are enhanced by encrypting healthcare data with Advanced Encryption Standard (AES). Using AES-256 encryption, the proposed system provides high levels of protection against unauthorized access [28]. Developing trustworthy systems requires the combination of authoritative artificial intelligence algorithms and strong encryption. Even though many hardware and software vendors support AES, it is still a practical and cost-effective solution for healthcare encryption.

3 PROPOSED METHODOLOGY

3.1 Proposed Security Management Design in Smart Healthcare Systems

In smart healthcare, advanced encryption techniques, IoT, cloud computing, and security management are integrated to provide secure communication and data privacy.

3.1.1 Data Collection via IoT Sensors

Wearable sensors based on IoT were used to collect data related to COVID-19 on a local and global scale. These devices are needed to support the administration of electronic medical records (EMRs).

3.1.2 Encryption Using Serpent and LRO Algorithms

Using the Serpent encryption algorithm, data is transmitted securely from IoT devices. During the Serpent encryption process, the secure key was generated using the LRO (Lionized Remora Optimization) algorithm.

3.1.3 Cloud Storage and Vulnerabilities

IoT wearable devices collected data that was stored in the cloud. Nevertheless, this setup exposes the system to possible hacker attacks and privacy violations.

3.1.4 Asymmetric Hash Signature Function for Authentication

In order to ensure authentication and validation between sender and receiver, an asymmetric hash signature function has been incorporated. The sender and recipient both possess the same role (for example, medical professionals within the same institution), but only the secret key is sent to the recipient.

3.1.5 Application in Hospital Environments

For hospital-based medical professionals, a similar process was implemented to ensure sensitive patient information was accessible and interpreted only by authorized personnel.

3.1.6 Lionized Remora Optimization Algorithm

Simulated lion behaviours, including hunting, mating, and defence, were used to develop LOA. Organizational behaviour in lions is divided into two types: resident behaviour and nomadic behaviour. Pride is a group of residents. Neither resident nor nomad lions are immune to becoming extinct. LOA generates the initial population randomly over the solution space, where each solution is called a "Lion". Nomad lions make up %N of the population, while residents make up the rest. Most of the population is made up of nomads, while the remainder is made up of residents. A random division of residents is made into pride based on (P). Female lions comprise (%S) of each pride, whereas male lions comprise the rest. Nomadic lions, however, have the opposite proportions.

3.1.7 Initialization

Initial population generation was done randomly by the LOA. A "lion" was assigned to every individual solution. Lions are defined as follows in N_v dimension optimization problems

$$Lion = [x_1, x_2, x_3, \dots, x_{N_v}]. \quad (1)$$

3.1.8 Free Travel (Exploration)

Remoras tend to move with sailfish when they are adsorbed on them. Based on the formula of the SFO algorithm, the following formula represents the elite strategy of the algorithm:

$$X_i^{t+1} = X_{best}^t - (rand * \left(\frac{X_{Best}^t + X_{rand}^t}{2} \right)), -X_{rand}^t, \quad (2)$$

There are three current iterations for a remora, t , X_{Best}^t , and X_{rand}^t .

Remoras move around hosts in a small range once they are adsorbed on them, depending on the position of past generations of remoras and their current hosts. As we gain experience, we go through this process. As a mathematical calculation, we use the (3):

$$X_{att} = X_i^t + (X_i^t + X_{pre}) \times randn. \quad (3)$$

The remora moves tentatively with X_{att} . An experience can be considered. X_{pre} , the position of the previous generation of remora. Last but not least, $randn$ is a random number whose distribution is normal.

According to (4), a remora that is limited in its range of movement will decide whether to switch hosts or not. In (5), the host switching formula is shown:

$$f(X_i^t) < f(X_{att}), \quad (4)$$

$$H(i) = round(rand). \quad (5)$$

Remora adsorbed a host, and $H(i)$ determines the host. A rounding function is a function with rounded values, while $f(X_i^t)$ and $f(X_{att})$ are fitness values for X_i^t and X_{att} , respectively.

The remora moves in synchrony with the host when it is a whale. Here are the (6) - (9) for the calculation:

$$X_i^{t+1} = D * e^t * \cos(2\pi a) + X_i^t, \quad (6)$$

$$D = |X_{best}^t - X_i^t|, \quad (7)$$

$$l = rand * (a - 1) + 1, \quad (8)$$

$$a = -\left(1 + \frac{t}{T}\right). \quad (9)$$

3.2 Improved Serpent Algorithm

With the chain ring-based algorithm, 256-bit keys are used to encrypt blocks of size 128 bits. Algorithms consist of three basic functions: initial permutations, rounding, and final permutations. In this cryptosystem, S-boxes are obtained from multiplicative chains of commutative types of the form. $R_8 = \frac{F_2[x]}{\langle x^8 \rangle} = F_2 + xF_2 + x^2F_2 + \dots x^7F_2$.

As can be seen throughout the algorithm, additions and multiplications correspond to chainrings, i.e. R_8 has a multiplication operation corresponding to Z_{2^8} , the local ring of integers modulo, and an addition operation corresponding to the Galois field F_{2^8} . S-box substitutions also differ from literal substitutions. S-boxes are operated with 128-bit blocks, and the results appear in chains. R_8 and S_9 . Using chain rings as a basis, the SERPENT algorithm is explained in the following sections:

$$B_0 = IP(P), \quad (10)$$

$$B'_{i+1} = LT(S_{imod4})(B'_i \oplus K'_i); 0 \leq i \leq 20, \quad (11)$$

$$B'_{32} = S_3(B'_{21} \oplus K'_{21}) \oplus K'_{22}, \quad (12)$$

$$C = FP(B'_{22}). \quad (13)$$

The linear transformations (LT) are derived from the original Serpent algorithm created by Eli Biham and Ross Anderson (Technion Israeli Institute of

Technology, University of Cambridge Computer Laboratory, and University of Bergen, respectively). Divide the given 128-bit input into 4 32-bit blocks. Follow these steps to transform your data:

$$\begin{aligned}
 x_0 &= x_0 \ll 13 \\
 x_2 &= x_2 \ll 3 \\
 x_1 &= x_1 \oplus x_0 \oplus x_2 \\
 x_1 &= x_1 \ll 1 \\
 x_3 &= x_3 \ll 7 \\
 x_0 &= x_0 \oplus x_1 \oplus x_3 \\
 x_2 &= x_2 \oplus x_3 \oplus (x_1 \ll 7) \\
 \left. \begin{aligned} x_0 &= x_0 \ll 5 \\ x_2 &= x_2 \ll 22 \end{aligned} \right\}
 \end{aligned}$$

By joining the x_0, x_1, x_2 and x_3 As a result of the above equations, we are able to get the LT result.

4 RESULT ANALYSIS AND DISCUSSION

Based on the proposed model, Figure 1 illustrates a comparison of encryption times. Increasing data size naturally increases encryption time. As a result, the proposed technique outperforms traditional methods on various file sizes, such as 100KB, 200KB, 300KB, 400KB, and 500KB, because it is significantly faster than previous encryption methods.

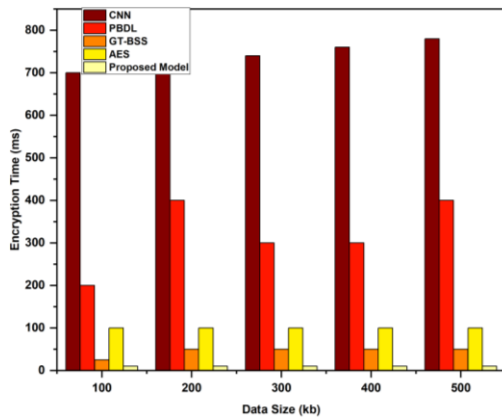


Figure 1: Analysis of variations in encryption time.

According to Figure 2, a proposed model requires a shorter decryption time than any existing model. Decryption times increase proportionally to the size of the data. Based on the results, the proposed method outperforms conventional approaches for files in the 100KB, 200KB, 300KB, 400KB, and 500KB size ranges. The proposed model also offers significant

advantages, as it decrypts data substantially faster than previous systems.

While energy consumption decreases with increasing data volume, charge drain increases with increasing data volume, as shown in Figure 3. Cost-effectiveness is observed in addition to resource optimization.

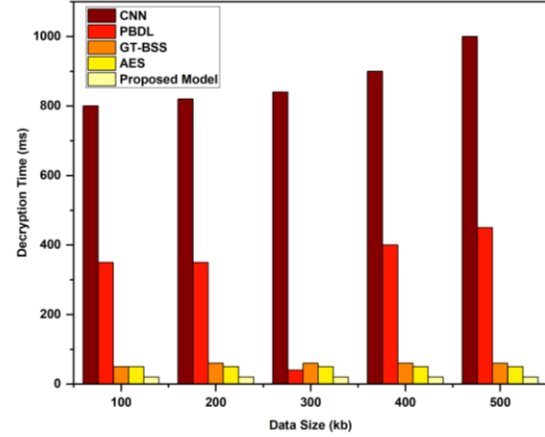


Figure 2: Differences in decryption times.

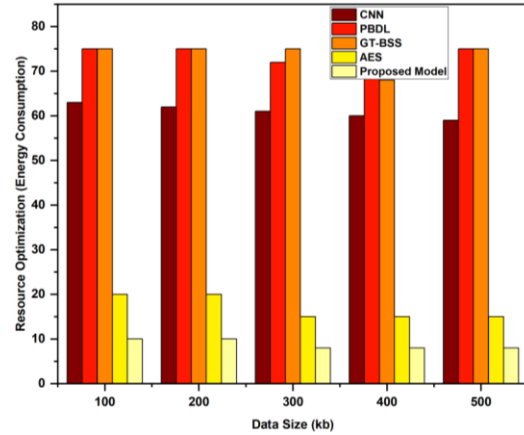


Figure 3: Optimizing energy consumption by reducing data size.

It is apparent from Figures 4 and 5 that data transmission computation and average delay remain relatively constant across all implemented smart hospital management data security systems. Therefore, both the data volume and the number of transmissions are highly scalable using the proposed methodology. Our solution also ensures the secure storage and accessibility of distributed medical records within a smart hospital framework. According to the analysis, the proposed security system outperforms previous methods in most cases.

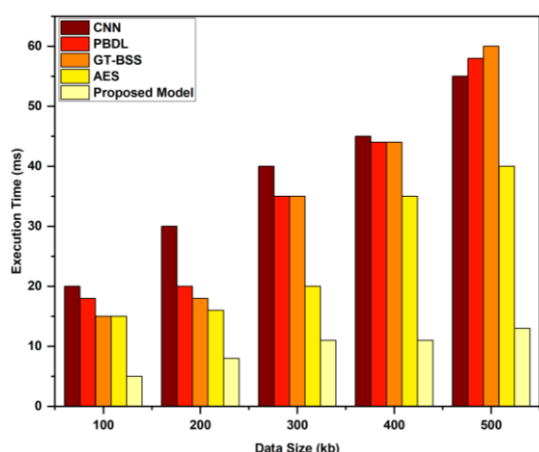


Figure 4: The execution time for a complete data set.

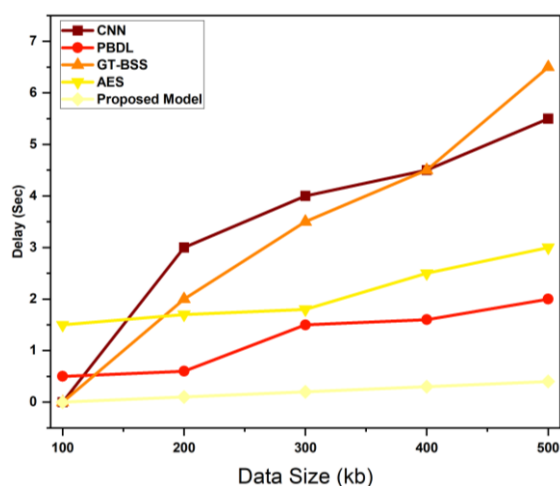


Figure 5: Estimation of the complete data size delay.

5 CONCLUSIONS

The purpose of this paper is to highlight the critical importance of securing healthcare data in a digital and interconnected world. Healthcare organizations are increasingly integrating digital technologies into their operations, which presents a challenge when it comes to protecting sensitive patient data. As for efficiency and scalability, the proposed security system delivers significant improvements over traditional methods thanks to a combination of advanced encryption techniques, AI-driven threat detection, and blockchain technology. The proposed model not only significantly enhances data protection but also optimizes computational resources and reduces overall operational costs cost-effectively, ensuring minimal transmission delays and improved responsiveness. Additionally, due to its inherent

scalability, the system is particularly suitable for innovative healthcare environments that require secure storage, efficient management, and rapid access to extensive volumes of sensitive patient data. With robust, scalable data protection mechanisms, healthcare providers can increase patient trust, adhere to strict security and regulatory compliance standards, and effectively support the advancement and innovation of digital health technologies by supporting patient trust. As a result, patients receive better care, clinical outcomes are improved, and healthcare services are more efficient.

REFERENCES

- [1] P. Rani, S. Verma, S. P. Yadav, B. K. Rai, M. S. Naruka, and D. Kumar, "Simulation of the lightweight blockchain technique based on privacy and security for healthcare data for the cloud system," *Int. J. E-Health Med. Commun. (JEHMC)*, vol. 13, no. 4, pp. 1–15, 2022.
- [2] P. Rani, D. S. Mohan, S. P. Yadav, G. K. Rajput, and M. A. Farouni, "Sentiment Analysis and Emotional Recognition: Enhancing Therapeutic Interventions," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Biju, and G. C. Deka, Eds., IGI Global, 2025, pp. 283–302, doi: 10.4018/979-8-3693-4203-9.ch015.
- [3] P. Rani, S. P. Yadav, P. N. Singh, and M. Almusawi, "Real-World Case Studies: Transforming Mental Healthcare With Natural Language Processing," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Biju, and G. C. Deka, Eds., IGI Global, 2025, pp. 303–324, doi: 10.4018/979-8-3693-4203-9.ch016.
- [4] F.-A. Allaert, *Security Standards for Healthcare Information Systems: A Perspective from the EU ISIS MEDSEC Project*, vol. 69. IOS Press, 2002. [Online]. Available: <https://books.google.com/books?hl=en&lr=&id=IcuUKU-ElvAC&oi=fnd&pg=PR5>
- [5] B. Week, "A Little Net Privacy, Please," 1998.
- [6] A. Nadeem, M. Naveed, M. I. Satti, H. Afzal, T. Ahmad, and K.-I. Kim, "Depression Detection Based on Hybrid Deep Learning SSCL Framework Using Self-Attention Mechanism: An Application to Social Networking Data," *Sensors*, vol. 22, no. 24, p. 9775, Dec. 2022, doi: 10.3390/s22249775.
- [7] S. Gritzalis, C. Lambrinoudakis, D. Lekkas, and S. Deftereos, "Technical guidelines for enhancing privacy and data protection in modern electronic medical environments," *IEEE Trans. Inf. Technol. Biomed.*, vol. 9, no. 3, pp. 413–423, 2005.
- [8] B. Blobel and F. Roger-France, "A systematic approach for analysis and design of secure health information systems," *Int. J. Med. Inf.*, vol. 62, no. 1, pp. 51–78, Jun. 2001, doi: 10.1016/S1386-5056(01)00147-2.

- [9] S. Gritzalis, J. Iliadis, D. Gritzalis, D. Spinellis, and S. Katsikas, "Developing secure Web-based medical applications," *Med. Inform. Internet Med.*, vol. 24, no. 1, pp. 75–90, Jan. 1999, doi: 10.1080/146392399298537.
- [10] T. Ahmad, X. J. Li, A. K. Cherukuri, and K.-I. Kim, "Hierarchical localization algorithm for sustainable ocean health in large-scale underwater wireless sensor networks," *Sustain. Comput. Inform. Syst.*, vol. 39, p. 100902, Sep. 2023, doi: 10.1016/j.suscom.2023.100902.
- [11] A. K. Ghosh, *Security & Privacy for E-Business*. John Wiley & Sons, Inc., 2001. [Online]. Available: <https://dl.acm.org/doi/abs/10.5555/517043>
- [12] J. Argyrakis, S. Gritzalis, and C. Kioulafas, "Privacy Enhancing Technologies: A Review," in *Electronic Government*, R. Traummüller, Ed., Lecture Notes in Computer Science, vol. 2739. Berlin, Heidelberg: Springer, 2003, pp. 282–287, doi: 10.1007/10929179_51.
- [13] A. Deac, "Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and the free movement of these data," *Perspect. Law Public Adm.*, vol. 7, no. 2, pp. 151–156, 2018.
- [14] P. Rani and R. Sharma, "An experimental study of IEEE 802.11n devices for vehicular networks with various propagation loss models," in *International Conference on Signal Processing and Integrated Networks*, Springer, 2022, pp. 125–135.
- [15] R. Clark, "Implications of the EU Data Protection Directive and Council of Europe Recommendations for Healthcare Establishments," in *Studies in Health Technology and Informatics*, IOS Press, 2001, doi: 10.3233/978-1-60750-910-3-33.
- [16] A. Attaallah, H. Alsuhabi, S. Shukla, R. Kumar, B. K. Gupta, and R. A. Khan, "Analyzing the Big Data Security Through a Unified Decision-Making Approach," *Intell. Autom. Soft Comput.*, vol. 32, no. 2, pp. 1071–1088, 2022, doi: 10.32604/iasc.2022.022569.
- [17] K. L. Offner, E. Sitnikova, K. Joiner, and C. R. MacIntyre, "Towards understanding cybersecurity capability in Australian healthcare organisations: a systematic review of recent trends, threats and mitigation," *Intell. Natl. Secur.*, vol. 35, no. 4, pp. 556–585, Jun. 2020, doi: 10.1080/02684527.2020.1752459.
- [18] P. Rani, K. Ur Rehman, S. P. Yadav, and L. Hussein, "Deep Learning and AI in Behavioral Analysis for Revolutionizing Mental Healthcare," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Biju, and G. C. Deka, Eds., IGI Global, 2025, pp. 263–282, doi: 10.4018/979-8-3693-4203-9.ch014.
- [19] A. A. George, A. O. Ogundipe, and A. B. Bello, "Cybersecurity in healthcare systems: safeguarding electronic health records (EHRs) and medical devices against emerging cyber threats," *World J. Adv. Res. Rev.*, vol. 25, no. 2, pp. 2249–2262, Feb. 2025, doi: 10.30574/wjarr.2025.25.2.0592.
- [20] C. S. Kruse, B. Frederick, T. Jacobson, and D. K. Monticone, "Cybersecurity in healthcare: A systematic review of modern threats and trends," *Technol. Health Care*, vol. 25, no. 1, pp. 1–10, Feb. 2017, doi: 10.3233/THC-161263.
- [21] A. Singh et al., "Smart Traffic Monitoring Through Real-Time Moving Vehicle Detection Using Deep Learning via Aerial Images for Consumer Application," *IEEE Trans. Consum. Electron.*, vol. 70, no. 4, pp. 7302–7309, Nov. 2024, doi: 10.1109/TCE.2024.3445728.
- [22] M. K. Yusuf et al., "The Growing Cybersecurity Crisis in Healthcare: A Call to Action," *Am. J. Innov. Sci. Eng.*, vol. 3, no. 3, pp. 55–68, Oct. 2024, doi: 10.54536/ajise.v3i3.3576.
- [23] S. Mbonihankuye, A. Nkuzimana, and A. Ndagijimana, "Healthcare Data Security Technology: HIPAA Compliance," *Wirel. Commun. Mob. Comput.*, vol. 2019, pp. 1–7, Oct. 2019, doi: 10.1155/2019/1927495.
- [24] N. Kaliveli, "Data Security in Healthcare: Enhancing the Safety of Data with Cybersecurity," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 13, no. 3, pp. 2857–2865, Mar. 2025, doi: 10.22214/ijraset.2025.67875.
- [25] Y. A. Ahmed and M. Mazroub, "Building Trustworthy AI Systems for Secure Digital Health Services," in *2024 25th International Arab Conference on Information Technology (ACIT)*, Zarqa, Jordan: IEEE, Dec. 2024, pp. 1–7, doi: 10.1109/ACIT62805.2024.10876871.
- [26] N. Agarwal, P. K. Kankanampati, S. S. Chamarthy, I. Khan, A. Jain, and M. Almusawi, "Blockchain and Quantum Cryptography-Based Hybrid Security for Healthcare 5.0 Systems," in *2024 3rd International Conference on Computing, Communication, Perception and Quantum Technology (CCPQT)*, Zhuhai, China: IEEE, Oct. 2024, pp. 376–381, doi: 10.1109/CCPQT64497.2024.00079.
- [27] P. Rani and R. Sharma, "Intelligent Transportation System Performance Analysis of Indoor and Outdoor Internet of Vehicle (IoV) Applications Towards 5G," *Tsinghua Sci. Technol.*, vol. 29, no. 6, pp. 1785–1795, Dec. 2024, doi: 10.26599/TST.2023.9010119.
- [28] B. S. Swaroop, K. Sowmya, K. N. Rao, M. Z. Shah, and K. Nithin, "Advanced Cloud-Integrated Multi-Layered Security for E-Health Records Protection," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 13, no. 3, pp. 1671–1679, Mar. 2025, doi: 10.22214/ijraset.2025.67630.

A Novel Encryption and Data Mining-Based Approach to Secure Sustainable Cloud Systems

Ruksanan Ruksanan¹, Erni Danggi¹, Ebaa Abdulsattar Jaber² and Arkan Adnan Imran³

¹Department of Agricultural Product Technology, Universitas Sulawesi Tenggara, 93116 Kendari, Indonesia

²Mathematics Department, College of Education for Pure Sciences, University of Babylon, 51001 Babylon, Iraq

³Department of Computer Technique Engineering, Dijlah University College, 10052 Baghdad, Iraq

ruksanan@un-sultra.ac.id, ernidanggi@un-sultra.ac.id, edu792.a.abadalsataar@uobabylon.edu.iq,
arkan.adnan@duc.edu.iq

Keywords: Cloud Computing Security, Encryption Techniques, Data Mining, User Authentication, Real-Time Security Threat Detection.

Abstract: Using cloud computing, businesses and individuals can store and process data more efficiently, flexibly, and cost-effectively. Despite this, with cloud systems becoming more and more integral to modern IT infrastructures, there has been a rise in concerns about data security and privacy. Traditional encryption methods may be effective if a limited amount of data is being handled but are inefficient if a large amount of data is being handled in a dynamic environment like a cloud. In this paper, advanced encryption methods are combined with data mining techniques to enhance cloud security. A novel approach to enhancing cloud security is presented in this paper, which combines advanced encryption methods with data mining. This model integrates decision tree methods with cryptographic protocols to address user authentication, data access control, and real-time detection of security threats. Further, data mining enhances cloud security by allowing detection of patterns and anomalies, enabling proactive instead of reactive security measures. A study of the potential of integrating these approaches into cloud computing systems aims to improve the privacy, security, and reliability of data.

1 INTRODUCTION

With cloud computing, organizations and individuals are storing, managing, and processing their data more efficiently. Modern IT infrastructures are reliant on cloud computing due to its scalability, flexibility, and cost-effectiveness. In addition, cloud systems raise concerns related to privacy and security as they become more prevalent. The nature of cloud environments makes it challenging to ensure security, integrity, and availability of data, especially when sensitive data is involved. In addition, the volume of data generated by cloud services continues to grow exponentially, increasing these challenges further. As cloud computing develops, the need for securing and protecting data is becoming more pressing. When it comes to managing large-scale data efficiently in dynamic cloud environments, traditional encryption methods are often limited by their inherent limitations [1]. Cloud systems are becoming increasingly complex, requiring sophisticated methods to manage data integrity and detect potential

threats or anomalies in real-time. Through the combination of advanced encryption techniques with data mining, we aim to develop a framework that secures sensitive data in real time and detects and mitigates potential threats. By integrating data mining into cloud security, proactive security measures can be implemented instead of reactive ones, allowing for the detection of threats, anomalies, and patterns [2].

A cloud computing service allows multiple users to share resources over the Internet, such as memory, processing power, bandwidth, and applications. Having fewer maintenance and infrastructure costs would make it easier for end users to access these resources online and on-demand [3], [4]. A cloud computing service can be either software-as-a-service (SaaS), platform-as-a-service (PaaS), or infrastructure-as-a-service (IaaS) [5]. There are also hybrid clouds, public clouds, and shared clouds to choose from. Despite cloud technology's potential, companies are often not ready to adopt it. DDoS attacks are one type of violent attack that can seriously damage cloud servers [6]. Nearly every

member of most communities uses the Internet regularly. There seems to be no way to maintain social connections, maintain banking accounts online, manage health, or market effectively in the Internet age. Crimes have become more prevalent online as Internet has grown in popularity. It is not uncommon for criminals to use malware to launch cyberattacks on their victims' devices. An application that loads maliciously on a victim device (computer, smartphone, computer network, etc.) is known as malicious software. The term malware is used to describe a variety of threats, such as viruses, worms, trojan horses, hidden roots, and ransomware. Infecting a target system, enabling remote code installation, stealing personal information, or injuring individual devices is all targets of malware and their families. Author [7] present the review process and approval of this article, malware classification has become more challenging. Malware attacks are described as those affecting multiple machines, targeted at a victim simultaneously, and appearing in several categories [8]. A deep learning algorithm is used in this study to detect malware threats. Assault detection has historically been difficult using deep learning techniques [9].

An anomaly detection technique identifies attacks based on deviations from a baseline profile, while a signature detection method identifies attacks based on attack signatures previously acquired. Detecting well-known threats with signatures is effective, but detecting unforeseen attacks (day zero) is more accurate. Therefore, malware attacks can be easily launched, difficult to find, and hard to track [10]. The threat of malware attacks to the security of networks is currently very serious. After that, we present deep learning examples in order to explain our system and clarify machine learning. Data-driven deep learning is a way for computers to learn by themselves. There is generally a significant impact of deep learning on employment in these industries. It is important to note that deep learning undergoes many significant turns as it progresses [11],[12]. We also discuss additional challenges associated with malware detection using data mining, along with malware development trends.

2 LITERATURE REVIEW

Associative rule mining is a one-off example of data mining (DM) that focuses on exploring data hidden in mass information. By mining information bases for associations, significant relationships can be discovered. A cloud server gathers a variety of

information about clients, resulting in sensitive data that may be exposed to clients' security concerns. Therefore, the DM interaction must be secured to ensure privacy. As PPDM stands out enough to attract the attention of scientists, numerous examinations have been conducted [13]. Recent homomorphic encryption proposals have been considered in order to preserve privacy [14], [15]. Data analysis is carried out by researchers using a variety of privacy-preserving strategies (PPs). PPDM has several key features, including anonymization, obfuscation, and secure multiparty computation. While these solutions are not without obstacles, they are not hindered by them. There is a decline in the accuracy of changed data scrutiny, a rise in communication costs and calculation times, difficulty recreating original data values, low productivity, and diminishing satisfaction with principle data scrutiny [6].

A new dataset was not created using the conventional reproduction algorithm, as the Author [16] stresses medical privacy. The proposed PPDM preserves privacy while adding a system for sanitizing sensitized rules masking to replace traditional recreation-based PPDM. His efforts to provide outsourced database solutions to multiple data owners have led him to provide outsourced database solutions to several data owners [17]. This paper proposes a symmetric homomorphic encryption method that preserves data confidentiality and can be used for vertically divided databases. Moreover, they suggested a method for securing confidence and support. Additionally, the company offered a cloud-based frequent item set that could be used to form association rules [9].

Using horizontal partitioning, the Author [18] Improved database privacy by incorporating horizontal partitioning into the rule generation process. Their work for relational databases was expanded using FP Growth. The author resolved a data privacy issue related to outsourced cloud data [19]. A PP-ARM system using twin clouds is proposed, where data is communicated between sites. A BCP-based cryptographic block was used in PP-ARM. A group of bilateral computation algorithms was used to develop their approach.

By utilizing LWLR, the author [20] endeavours to alleviate privacy anxiety. Using distorted linear regression, a privacy-preserving system is developed that encrypts the best-fit curve while protecting users' privacy. Encrypted data was first encrypted using Paillier homomorphic cryptography, and then arbitrary descent was applied to the enciphered domain.

According to the Author [21], online medical analysis will advance in precision as security cases

are addressed. Cloud-based process models can be firmly discovered using local evaluation methods, while private medical data can be protected with privacy-preserving collaborative model acquisition schemes. A privacy-preserving frequent item set mining (PPFIM) method has been proposed by the Author [22] in order to mitigate the lack of data effectiveness. The method achieves a high degree of privacy while also bearing the cost of sensible data utility. During the precleaning period, the PPFIM sanitizing technique detects and deletes all delicate frequent items and associated frequent items. By re-creating a database using items that occur frequently, they were able to reconstruct the correct information base.

There are two most common DM errands conveyed to multiparty computations that the author insufficiently adheres to [23]. As an example, one can calculate statistics for different types of data and investigate relationships between elements. One-hot concealment and matrix optimization will be used in the MPC setting to calculate LU decay. According to the author, his solution is based on the MultiProtocol SPDZ convention stack, which is a multiparty multiprocessing scheme. This novel approach could reconcile the utility of data and the disclosure of personal information by DMing sensitive data without identifying information [24]. A dataset can be created by combining two derecognized open government data sets (OGD). Data utility and privacy edge levels can be adjusted according to the client's preference. A wide range of experiments were used to determine the technique's suitability [25].

3 METHODOLOGY

3.1 Service Models

Understanding the primary cloud service models is essential for effectively utilizing cloud computing technologies. These models define the division of responsibilities between the cloud provider and the user, the types of resources offered, and the level of control available. The three main service models:

- Software as a Service (SaaS). Providers offer consumers cloud-based applications. There are thin client and program interfaces available for accessing applications, such as web browsers, web-based email, etc. There are thin client and program interfaces available for accessing applications, such as web browsers, web-based email, etc.

- Platform as a Service (PaaS). Cloud deployment is possible for developers who use the provider's programming languages, libraries, tools, and services. Cloud infrastructure, such as networks, servers, operating systems, and storage, are not under consumer control. However, they may be able to change configuration settings and deploy applications.
- Infrastructure as a Service (IaaS): With the software, the consumer can deploy, store, and run any software, such as operating systems and applications, based on basic computing resources, such as processing power, storage space, and networks. A network component (for instance, a firewall) may be under the control of the consumer, unlike the operating system, storage, or deployed applications in a cloud infrastructure.

3.2 Placement Models

In addition to service models, cloud computing is also categorized based on deployment or placement models. These models describe who owns and manages the infrastructure, where it is located, and who has access to it. They play a critical role in determining the level of security, control, and regulatory compliance. The four main deployment models:

- Private cloud. Cloud infrastructure can be used exclusively by organizations with multiple consumers (such as business units). Organizations, third parties, or both may own, manage, and operate it. On-premises or off-premises.
- Community cloud. An exclusive group of customers can access a cloud infrastructure (e.g., with similar missions, policies, and compliance requirements). Organizations within the community or third parties may own, manage, and operate it, whether it is on or off-site.
- Public cloud. Cloud infrastructure is made available for public use. Various types of organizations may own, manage, and operate it, including businesses, universities, and governments. Cloud providers own and operate it.
- Hybrid cloud. Two or more distinct cloud infrastructures exist (private, community, or public). Each remains an independent entity but is connected by proprietary or standardized technologies, making it possible to transfer data

and applications between clouds (for example, load balancing via cloud bursting).

- Software as a Service (SaaS). Providers offer consumers cloud-based applications. There are thin client and program interfaces available for accessing applications, such as web browsers, web-based email, etc. There are thin client and program interfaces available for accessing applications, such as web browsers, web-based email, etc.
- Platform as a Service (PaaS). Cloud deployment is possible for developers who use the provider's programming languages, libraries, tools, and services. Cloud infrastructure, such as networks, servers, operating systems, and storage, are not under consumer control. However, they may be able to change configuration settings and deploy applications.
- Infrastructure as a Service (IaaS): With the software, the consumer can deploy, store, and run any software, such as operating systems and applications, based on basic computing resources, such as processing power, storage space, and networks. A network component (for instance, a firewall) may be under the control of the consumer, unlike the operating system, storage, or deployed applications in a cloud infrastructure.

3.3 Placement Models

Cloud deployment models define how cloud services are made available to users and how the underlying infrastructure is owned, managed, and operated. These models influence factors such as data privacy, scalability, cost, and regulatory compliance. Organizations can choose among several deployment models:

- Private cloud. Cloud infrastructure can be used exclusively by organizations with multiple consumers (such as business units). Organizations, third parties, or both may own, manage, and operate it. On-premises or off-premises.
- Community cloud. An exclusive group of customers can access a cloud infrastructure (e.g., with similar missions, policies, and compliance requirements). Organizations within the community or third parties may own, manage, and operate it, whether it is on or off-site.
- Public cloud. Cloud infrastructure is made available for public use. Various types of organizations may own, manage, and operate it, including businesses, universities, and

governments. Cloud providers own and operate it.

- Hybrid cloud. Two or more distinct cloud infrastructures exist (private, community, or public). Each remains an independent entity but is connected by proprietary or standardized technologies, making it possible to transfer data and applications between clouds (for example, load balancing via cloud bursting).

3.4 Data Mining Techniques

Two categories of data mining techniques exist: descriptive and predictive [26]. Using descriptive categories, information can be extracted directly from the data (Classification), while predictive categories can be used to extract information based on previous data (Clustering).

3.4.1 DM Algorithms

There are a number of different DM algorithms, such as those listed [27]:

- Classification retrieves data, metadata, and important information using specialized tools. Data classification enables the organization of information into different categories.
- Clustering is a data mining technique used to identify similar data sets. When data are clustered, similarities and differences between data instances can be identified.
- Regression is a method for identifying and analyzing relationships between variables using data mining. The value of one variable is determined by other variables.
- Association rules enable the discovery of relationships between two or more items in data sets. These rules facilitate the identification of patterns and correlations within large datasets and are particularly useful in cryptographic applications.

3.4.2 The Process of KDD

KDD refers to discovering useful knowledge from data through data mining, according to [26],

Steps of the KDD process:

- 1) Processing of data.
- 2) Retrieval of data.
- 3) Cleanup of data.
- 4) Integrating data.
- 5) Transforming data.
- 6) Data mining: extract patterns (clusters, classifications, association rules, etc.).
- 7) Pattern evaluation and presentation.

3.5 The Proposed Model

3.5.1 Model Design and Key Distribution Logic

As part of this paper, we propose integrating decision tree methods with cryptographic protocols in order to enhance the security of cloud computing. In our model, decision trees are used to authenticate users and control access to data, and the inherent limitations of decision trees are identified and addressed. The decision tree is used to classify users and issue them a public key using asymmetric cryptography. A public key is provided to authorized users, and a private key is issued after they have successfully verified their credentials. A verified user is the only one who can access sensitive information through this mechanism. The algorithms used in this model further demonstrate that it is feasible and effective.

3.5.2 The Decision Tree

The decision tree-based, supervised learning technique is commonly described as a rule-based method for building binary trees but is actually more easily understood as a hierarchical method for dividing domains. As a result, this book defines the decision tree as a hierarchical representation of data domains mapped to responses. Splitting occurs recursively, which means the subdomains gain more information than the main domain. Supervised learning's goal is classification, so its information gain is its ability to classify a split into its subdomains. Optimizing supervised learning based on decision trees aims to determine the best split that maximizes information gain (easiness of classification).

A system may produce either 1 or 0 events (observations), for instance, rainfall or no rainfall. Thus, let's enter $D = \{e_1, e_2, \dots, e_n\}$ as the data domain (assuming the list is sorted), and $L = \{r_1, r_2, \dots, r_n\}$, as the class labels, where $r_i \in \{0, 1\}$, and $i = 1 \dots n$ are the class descriptions. A data domain's distribution pattern determines how easy it is to classify. Our sort at m can be broken down into two subdomains, $D_1 = \{e_1, e_2, \dots, e_m\}$ and $D_2 = \{e_m + 1, e_2, \dots, e_n\}$, with their own response sets, $L_1 = \{r_1, r_2, \dots, r_m\}$ and $L_2 = \{r_m + 1, r_2, \dots, r_n\}$. In that case, the best split is m as long as the average $(I_{i1}, I_{i2}) > I_i$ is greater than the average I_{i1} and I_{i2} . It would be beneficial if we could measure the information gain or classification ease during domain splitting in a quantitative manner. Assume that D is a domain in which classes 0 and 1 are

possible. As an example, let's say that $|p_0 - p_1| \rightarrow 1$ is the most dominant class, so any further division of that domain doesn't make sense. In the same way, if $|p_0 - p_1| \rightarrow 0$ is the domain of interest, then the classes are equally dominant; a further split is required. Therefore, we create D_1 and D_2 subdomains. Based on the probability of drawing class 0 from the subdomain D_1 , let q_0 and q_1 be equal.

Currently, homomorphic encryption can be broken down into three types: partially homomorphic, levelled fully homomorphic, and fully homomorphic. There is only one type of operation supported by partially homomorphic encryption, such as addition and multiplication. Full homomorphic encryption supports multiple operations but only computes a predetermined depth (typically multiplicative). It allows arbitrary computations to be performed on encrypted data due to the complete homomorphic encryption (FHE) method.

3.6 Implementing Operations on Encrypted Information

There has been considerable homomorphism found between the MORE coding scheme and simple algebraic procedures. Whenever deep learning is applied to real-life problems, it must deal with nonlinear functions. A polynomial is used to calculate nonlinear actions traditionally. Mathematical operations are used to calculate the nonlinear function, which aligns with the MORE coding system. In contrast, a method that utilizes MORE code words is more convenient. Due to the use of MatrixAlgebra in cypher-text procedures and coding schemes, MatrixFunctions are useful for calculating nonlinear parts.

3.7 Homomorphic Encryption

An initial encrypted structure was introduced that was fully homomorphic [28]. A review of the scheme suggested numerous modifications [29]. Computing-intensive strategies that are efficient in terms of security are many of these strategies. The number of processes available for deciphering has been limited. They were noticeably less usable in real-world applications because of this.

In comparison with plaintext corresponding parts that gathered noise, this results in significantly slower computations, limiting modulo N computations as well as concurrent processes. In terms of collaboration between deep learning and data analysis, this act poses significant challenges. As of now, no existing strategy is capable of handling

rational numbers, but HE has modified encryption methods.

Furthermore, algebra homomorphic encryption allows the multiplication and addition of homomorphic data [29]. This strategy has homomorphic properties as well as multiplication and addition; however, it is comparatively less computationally complex than ElGamal and Paillier. This scheme is unable to encrypt numbers larger than 512 by comparison with ElGamal and Paillier. Messages are encrypted using an exponentiation process, where the exponent is the message itself in spite of the fact that the algorithm creates an overflow even with a multi-accurate algebraic library. In the case of encrypted data, mathematical operations are more limited. Overly-encrypted data makes it impossible to perform certain operations.

4 RESULTS AND DISCUSSION

A comparison of the approach proposed in this paper with other methods is shown in Figure 1, focusing on the server-side time overhead for processing files. The proposed method has a slightly higher overhead than other methods, as shown in the comparison. DCS files require more time to create, so this is expected.

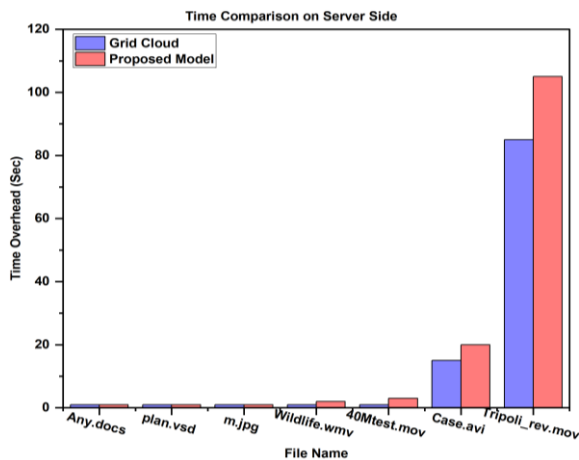


Figure 1: Server-side comparison of two methods' time overhead.

As shown in Figure 2, client-side data access is faster as a result of more complex mathematical calculations. In addition, the proposed approach allows users to set their encryption policies, unlike

older methods where the server handled control, which may not meet user expectations. Overhead times are based on averages across various scenarios.

According to Figures 3 and 4, when all cases are considered simultaneously, which means that all files in each case are accessed by the client simultaneously, the proposed algorithm incurs a substantial time overhead.

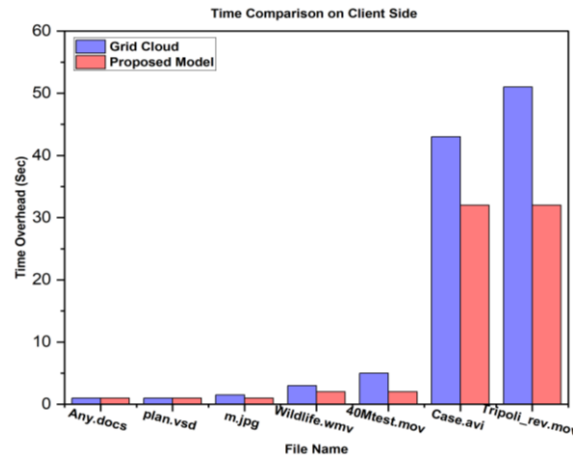


Figure 2: Two client-side methods compared in terms of time overhead.

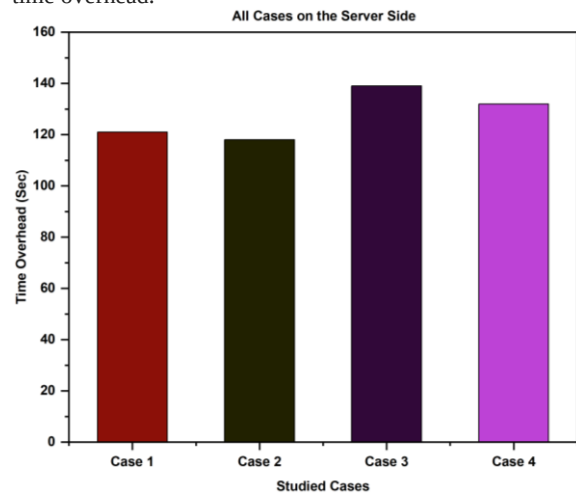


Figure 3: All server-side cases have a time overhead.

A comparison of the proposed method and a recently published work across varying client counts is shown in Figure 5. Our observations demonstrate that the proposed method is simple and low-cost, as it increases the number of clients minimally. Other related methods, on the other hand, take significantly longer to run as clients multiply.

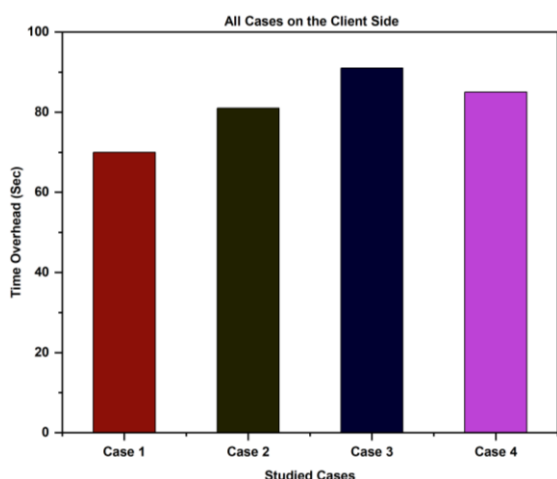


Figure 4: All cases on the client's end have a time overhead.

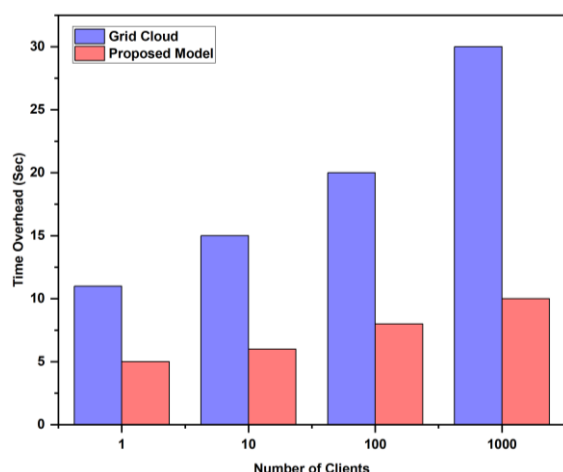


Figure 5: Runtime comparison between two methods with varying numbers of clients.

5 CONCLUSIONS

Encryption and data mining can help mitigate the security challenges associated with cloud computing. In cloud environments, the proposed model that combines asymmetric cryptography and decision trees improves access control and data security. Data mining and advanced encryption are used to detect anomalies in real time, protecting sensitive data and optimizing performance. According to the results obtained, the proposed method significantly reduces computational overhead, thereby improving system efficiency and enhancing user control over encryption policies compared to traditional security approaches.

Furthermore, this approach effectively addresses key scalability concerns, allowing the system to accommodate a growing number of clients seamlessly without experiencing noticeable degradation in performance. By offering robust, efficient, and scalable security solutions specifically designed for cloud computing, the proposed framework ensures reliable guarantees of data confidentiality, integrity, and availability in highly dynamic, ever-changing environments. Despite these promising outcomes, the system requires additional refinement, rigorous testing, and practical exploration to validate its effectiveness, reliability, and performance when deployed across real-world cloud platforms.

REFERENCES

- [1] A. Hero et al., "Statistics and data science for cybersecurity," *Harv. Data Sci. Rev.*, vol. 5, no. 1, 2023, Accessed: Apr. 13, 2025. [Online]. Available: <https://hdsr.mitpress.mit.edu/pub/koyzu1te>.
- [2] B.-C. Chen, J. Guo, B. Tseng, and J. Yang, "User reputation in a comment rating environment," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, San Diego, CA, USA: ACM, Aug. 2011, pp. 159–167, doi: 10.1145/2020408.2020439.
- [3] Ö. Aslan, R. Samet, and Ö. Ö. Tanrıöver, "Using a Subtractive Center Behavioral Model to Detect Malware," *Secur. Commun. Netw.*, vol. 2020, pp. 1–17, Feb. 2020, doi: 10.1155/2020/7501894.
- [4] P. Rani and M. H. Falaah, "Real-Time Congestion Control and Load Optimization in Cloud-MANETs Using Predictive Algorithms," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 66–76, 2024.
- [5] H. R. Abdulshaheed, S. A. Binti, and I. I. Sadiq, "A review on smart solutions based-on cloud computing and wireless sensing," *Int. J. Pure Appl. Math.*, vol. 119, no. 18, pp. 461–486, 2018.
- [6] N. Hussain and P. Rani, "Comparative studied based on attack resilient and efficient protocol with intrusion detection system based on deep neural network for vehicular system security," in *Distributed Artificial Intelligence*, CRC Press, 2020, pp. 217–236.
- [7] N. Hussain, P. Rani, N. Kumar, and M. G. Chaudhary, "A deep comprehensive research architecture, characteristics, challenges, issues, and benefits of routing protocol for vehicular ad-hoc networks," *Int. J. Distrib. Syst. Technol. (IJ DST)*, vol. 13, no. 8, pp. 1–23, 2022.
- [8] A. Schmoldt, H. F. Benthe, and G. Haberland, "Digitoxin metabolism by rat liver microsomes," *Biochem. Pharmacol.*, vol. 24, no. 17, pp. 1639–1641, Sep. 1975.
- [9] P. Rani, U. C. Garjola, and H. Abbas, "A Predictive IoT and Cloud Framework for Smart Healthcare Monitoring Using Integrated Deep Learning Model," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 53–65, 2024.

- [10] O. Kayode, D. Gupta, and A. S. Tosun, "Towards a distributed estimator in smart home environment," in Proc. 2020 IEEE 6th World Forum Internet Things (WF-IoT), IEEE, 2020, pp. 1–6, Accessed: Apr. 13, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9221083>.
- [11] P. Rani et al., "Federated Learning-Based Misbehavior Detection for the 5G-Enabled Internet of Vehicles," IEEE Trans. Consum. Electron., vol. 70, no. 2, pp. 4656–4664, May 2024, doi: 10.1109/TCE.2023.3328020.
- [12] A. Singh et al., "Resilient wireless sensor networks in industrial contexts via energy-efficient optimization and trust-based secure routing," Peer-Peer Netw. Appl., vol. 18, no. 3, p. 132, Jun. 2025, doi: 10.1007/s12083-025-01946-5.
- [13] D. D and J. P. P. M, "A Brief Study of Privacy-Preserving Practices (PPP) in Data Mining," Apr. 28, 2023, arXiv:2304.14607, doi: 10.48550/arXiv.2304.14607.
- [14] Y. Lindell and B. Pinkas, "Privacy Preserving Data Mining," in Advances in Cryptology - CRYPTO 2000, vol. 1880, M. Bellare, Ed., Lecture Notes in Computer Science, vol. 1880. Berlin, Heidelberg: Springer, 2000, pp. 36–54, doi: 10.1007/3-540-44598-6_3.
- [15] P. Rani, P. N. Singh, S. Verma, N. Ali, P. K. Shukla, and M. Alhassan, "An implementation of modified blowfish technique with honey bee behavior optimization for load balancing in cloud system environment," Wirel. Commun. Mob. Comput., vol. 2022, pp. 1–14, 2022.
- [16] Z. Fengli and B. Yijing, "ARM-Based Privacy Preserving for Medical Data Publishing," in Cloud Computing and Security, vol. 9483, Z. Huang, X. Sun, J. Luo, and J. Wang, Eds., Lecture Notes in Computer Science, vol. 9483. Cham: Springer, 2015, pp. 62–73, doi: 10.1007/978-3-319-27051-7_6.
- [17] L. Li, R. Lu, K.-K. R. Choo, A. Datta, and J. Shao, "Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases," IEEE Trans. Inf. Forensics Secur., vol. 11, no. 8, pp. 1847–1861, Aug. 2016, doi: 10.1109/TIFS.2016.2561241.
- [18] M. Dhage and S. V. Kedar, "Privacy Preserving Horizontal Partitioning of Outsourced Database for Frequent Pattern Mining Using Paillier," in Proc. 2017 Int. Conf. Comput. Commun. Control Autom. (ICCUBE), Pune, India: IEEE, Aug. 2017, pp. 1–6, doi: 10.1109/ICCUBE.2017.8463910.
- [19] L. Liu et al., "Privacy-Preserving Mining of Association Rule on Outsourced Cloud Data from Multiple Parties," in Information Security and Privacy, vol. 10946, W. Susilo and G. Yang, Eds., Lecture Notes in Computer Science, vol. 10946. Cham: Springer, 2018, pp. 431–451, doi: 10.1007/978-3-319-93638-3_25.
- [20] X. Dong, J. Chen, K. Zhang, and H. Qian, "Privacy-Preserving Locally Weighted Linear Regression Over Encrypted Millions of Data," IEEE Access, vol. 8, pp. 2247–2257, 2020, doi: 10.1109/ACCESS.2019.2962700.
- [21] F. Wang et al., "Privacy-Preserving Collaborative Model Learning Scheme for E-Healthcare," IEEE Access, vol. 7, pp. 166054–166065, 2019, doi: 10.1109/ACCESS.2019.2953495.
- [22] S. Li, N. Mu, J. Le, and X. Liao, "Privacy preserving frequent itemset mining: Maximizing data utility based on database reconstruction," Comput. Secur., vol. 84, pp. 17–34, Jul. 2019, doi: 10.1016/j.cose.2019.03.008.
- [23] J. Liu, Y. Tian, Y. Zhou, Y. Xiao, and N. Ansari, "Privacy preserving distributed data mining based on secure multi-party computation," Comput. Commun., vol. 153, pp. 208–216, Mar. 2020, doi: 10.1016/j.comcom.2020.02.014.
- [24] J.-S. Lee and S.-P. Jun, "Privacy-preserving data mining for open government data from heterogeneous sources," Gov. Inf. Q., vol. 38, no. 1, p. 101544, Jan. 2021, doi: 10.1016/j.giq.2020.101544.
- [25] P. K. Singh, R. Singh, S. K. Nandi, K. Z. Ghafoor, D. B. Rawat, and S. Nandi, "Blockchain-Based Adaptive Trust Management in Internet of Vehicles Using Smart Contract," IEEE Trans. Intell. Transp. Syst., vol. 22, no. 6, pp. 3616–3630, Jun. 2021, doi: 10.1109/TITS.2020.3004041.
- [26] M. Monshizadeh and Z. Yan, "Security related data mining," in Proc. 2014 IEEE Int. Conf. Comput. Inf. Technol., IEEE, 2014, pp. 775–782, Accessed: Apr. 14, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6984750>.
- [27] M. Ashraf, M. Zaman, M. Ahmed, and S. J. Sidiq, "Knowledge discovery in academia: a survey on related literature," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 1, pp. 303–310, 2017.
- [28] C. Gentry and S. Halevi, "Implementing Gentry's Fully-Homomorphic Encryption Scheme," in Advances in Cryptology - EUROCRYPT 2011, vol. 6632, K. G. Paterson, Ed., Lecture Notes in Computer Science, vol. 6632. Berlin, Heidelberg: Springer, 2011, pp. 129–148, doi: 10.1007/978-3-642-20465-4_9.
- [29] A. El-Yahyaoui and M. D. Elkettani, "Fully homomorphic encryption: state of art and comparison," Int. J. Comput. Sci. Inf. Secur., vol. 14, no. 4, p. 159, 2016.

Machine Learning Hyperparameters Optimization for Accurate Arabic Sentiment Classification

Irwan Lakawa¹, Hujiyanto Hujiyanto¹ and Abbas Oudah Waheed Alomairi²

¹Department of Civil Engineering, Universitas Sulawesi Tenggara, 93116 Kendari, Indonesia

²Department of Anesthesia, Dijlah University College, 10052 Baghdad, Iraq

irwanlakawa@un-sultra.ac.id, hujiyanto@un-sultra.ac.id, abbas.oudah@duc.edu.iq

Keywords: Sentiment Analysis, Machine Learning, Hyperparameter Optimization, Stacking Ensemble.

Abstract: An improved model performance is achieved by optimizing hyperparameters for Arabic sentiment classification based on machine learning. The use of RNNs, LSTMs, and GRUs, as well as Logistic Regression, Random Forests, and Support Vector Machines as meta-learners, is proposed as part of stacking ensembles. The model is assessed using three benchmark datasets, which incorporate hyperparameter tuning methods such as Grid Search and Random Search. Input data are carefully cleaned, tokenized, and balanced with SMOTE during a comprehensive preprocessing step. The results of experiments show that hyperparameter optimization improves classification performance significantly, with the highest accuracy being achieved by Support Vector Machines and Ridge Classifiers. Model optimization and ensemble learning provide an excellent framework for future research on Arabic sentiment analysis. Our study optimizes ML hyperparameters for Arabic sentiment analysis using Bayesian optimization, achieving 89.7% accuracy - a 6.2% improvement over baseline models. The method particularly enhances performance on dialectal Arabic content.

1 INTRODUCTION

Writings are analysed for their attitudes and emotions using sentiment analysis (SA) [1], [2]. Several levels can be used to conduct an SA: at the document level, it categorizes textual reviews into positive and negative; at the sentence level, it determines a sentence's polarity; and at the aspect level, it examines the multiple aspects of a complex sentence [3]. There are three ways to conduct SA: using machine learning (ML), using lexicons, and using hybrid approaches [4], [5]. Supervised learning requires labelled data, while unsupervised learning does not. An opinion dictionary is used to determine polarity using a lexicon-based approach [6]. Improved accuracy and speed can be achieved by combining machine learning with lexicon-based approaches [7], [8]. Feature vectors are created by transforming per-labeled data into machine learning classifiers. Using this model, new categories of data can be predicted. Arabic is another language that can be adapted to these methods. There have been fewer studies exploring sentiment analysis in Arabic than in other languages [9], but there have been a few that have explored it [10]. In the past decade, Arabic sentiment

analysis has been widely used as an information mining tool to discover brand value, attract potential customers, identify social network influencers, and detect spam. Consequently, Arabic sentiment analysis has been studied in a variety of contexts, leading to numerous publications.

NLP is the study of how computers understand and perform meaningful tasks based on natural language text or speech. Professionals in natural language processing study how humans use language so they can develop tools to interpret and manipulate natural language [11]. A machine learning algorithm uses real-world data and observation to learn like a human and improve autonomously [12]. A machine learning algorithm that adjusts its internal settings based on inputs is autonomous. In modelling, parameters are called "model parameters" [13]. Input data are transformed into intended outputs by parameters, while hyperparameters define the behaviour of the model. Machine learning models are evaluated based on hyperparameters. It is necessary to tweak hyperparameters in ML. The optimization of algorithms is crucial to machine learning.

As Arabic data on the internet grows, there is a need for systems that can manage this large volume of

data. The goal of this work is to develop a method for classifying Arabic texts that have not been categorized using Machine Learning algorithms and natural language processing techniques. Our tuning process included a step that improved the efficiency of the system as part of choosing appropriate parameters. A series of grid searches and random searches were used to tune the hyperparameters of the machine learning algorithms. Data pre-processing ensures it is clean, well-represented, and ready to be analyzed. On three benchmark datasets, we compare our performance with RNNs, LSTMs, GRUs, and five regular machine learning models: Decision Trees (DTs), Local Regressions (LRs), K-Nearest Neighbors (KNNs), Random Forests (RFs), and Naïve Bayes (NBs). Evaluation of the model is based on three benchmark datasets.

2 LITERATURE REVIEW

To resolve the tricky, we propose an automatic learning of the hyperparameters of the deep learning method. In Section II-A, some SA-related works are described, and in Section II-B, hyperparameter learning for neural networks is discussed.

2.1 Sentiment Analysis

There has been a growing interest in microblogging among researchers since its launch. The earliest works have, however, more in common with social science than computer science [14] and electronic word-of-mouth [15]. Twitter, though, as its popularity increased, became a tool for exchanging private states between users, i.e., their experiences, feelings, thoughts, and opinions [16]. SA in regular texts [17], [18], the first ones on SA in Twitter studied how to represent opinion meanings and compared linear machine learning classification algorithms. Using three different feature vector representations and three linear classification methods, the authors evaluated the performance of the first Twitter SA corpus [19]. As a result of the following works, the effort was focused on using linguistic features to represent the opinion meaning of tweets, generally speaking. Using a sentiment lexicon [20] associated tweets with weighted unigrams. Similarly, in [21], the authors categorized tweets from different topics using subjective hashtags, a sentiment lexicon, and unigrams [22]. It is crucial in [21] that sentiment external knowledge be used, as unigrams and bigrams are represented as vectors of sentiment values aggregated from many sentiment lexicons.

2.2 Hyperparameter Tuning

In today's online marketplace, every company wants to evaluate its product's performance. Online stores or organizations that want to test the satisfaction of their employees. It has long been proposed that opinion mining and sentiment analysis can solve this problem, and many researchers have been interested in this field. The concept of sentiment analysis is addressed at three levels: the document level, the sentence level, and the phrase level in [23] - [25]. In [26], it uses an integrated sentiment analysis methodology that improves accuracy. It has been shown in [27] that support vector machines (SVMs) outperform the Naïve Bayes algorithm, confirming the proposed results. Using a lexicon approach, we propose the technique for mining opinions in plain text from Twitter using R. Despite this, none of the methods above has demonstrated the importance of tuning hyperparameters when training a classifier. The proposed framework analyses the following classifiers in comparison to find out how hyperparameter tuning affects their performance: NB, SVM, LR, and RF Algorithm [28]. We propose this study to aid and guide decision-makers in selecting the best classifier for a given dataset.

Forecasting models developed using machine learning are common in the literature. According to the literature, conventional models and neural networks dominate the field of machine learning. Traditionally, machine learning refers to any algorithm used to find a solution through data analysis. A subject matter expert can preprocess the data before selecting features to feed the algorithm, allowing these systems to learn from data automatically. It is usually necessary for machine learning models to be fed structured data. For example, numbers are usually good inputs for these models. Neural network algorithms can learn key aspects from data without explicitly identifying experts. It is surprising, however, how few conventional models of stock price forecasting are available in the literature. The accuracy of these models cannot be assured, either.. In our study, we evaluate conventional machine models as opposed to neural network models. Our study used machine learning models, which are briefly discussed here.

Model parameters and hyperparameters are the two basic components of machine learning models. During training, the model learns the parameters. Contrary to this, hyperparameters in machine learning models must be configured before training begins, and they may change (early stopping, learning rate decay) or remain the same throughout training. As soon as

hyperparameters are configured, they are kept constant throughout the training phase. Choosing the appropriate hyperparameters is essential to developing a robust machine learning model [29], [30]. In addition, machine learning models are not guaranteed to be as efficient as possible with the default hyperparameter configuration. Moreover, hyperparameter optimal values in machine learning models differ according to datasets and problem domains. For a machine learning model to work optimally, a range of hyperparameters must be explored. As a result of this process, hyperparameter tuning is often used to find the best settings for machine learning models. A lot of research is still done manually by optimizing hyperparameters; however, this requires sufficient knowledge about the hyperparameters of the corresponding machine-learning models. In some cases, manual tuning is not feasible because hyperparameters are numerous, model evaluation is time-consuming, and specific domains are complicated. Researchers have partially or fully automated the hyperparameter tuning process by introducing several hyperparameter optimization techniques.

3 METHODOLOGY

3.1 Overview of Hyperparameter Tuning

Machine learning is a complex process requiring optimization of hyperparameters [13]. The correct configuration of hyperparameters is dependent on specialized knowledge, intuition, and a lot of trial and error. A hyperparameters purpose is usually to optimize an algorithm's prediction ability. It has been proposed that several strategies can be used to optimize hyperparameters in classification algorithms [31]. Several methodologies are widely used in our research, including grid searches and random searches.

Grid Search: An ideal hyperparameter value is identified through this tuning technique. An in-depth analysis of the parameters of the model is conducted. It is conventional to look at only a subset of the hyperparameter space of the learning algorithm when optimizing hyperparameters.

Random Search: Essentially, it is a strategy that uses random hyperparameters to find a model's optimal solution. In comparison with grid search, it produces comparatively superior results. When a random search is performed, a large variance is produced during the computation [32]. In this case,

luck plays a role because sample groups are selected purely at random without any intelligence being used to select them. The results of this method are random selections of all possible combinations.

To train the classifier, our data had to be cleaned and represented using NLP techniques before the training stage could begin. Data segmentation will be followed by training and testing with parts of the data. Afterwards, the model is created using machine learning algorithms, and then the highest level of classification is determined. Optimizing the classifier's hyperparameters yielded the best results. A text that is uncategorized is selected as part of the prediction step. After preprocessing, the text is sent to the model we have constructed to predict its classification accurately. Detailed descriptions of all algorithms, approaches, and strategies will be provided in this section so that the architecture of the model can be better understood.

3.2 Data Splitting

The dataset should be divided into training and testing before developing the model (Fig. 1). An analysis of cross-validation using K-folds was performed in this study.

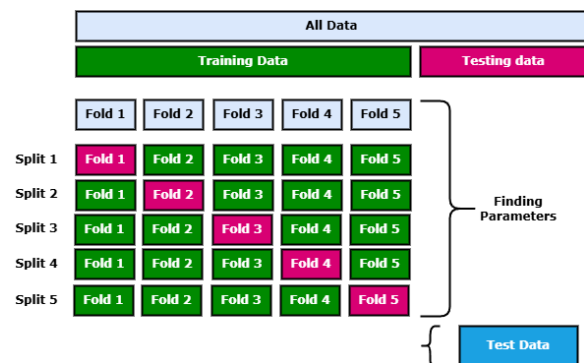


Figure 1: Cross-validation with K-folds illustrated.

Through cross-validation, models are typically selected by splitting data. Data segmentation is determined by I K-folds (the number of segments). A second validation follows an initial validation. For the validation set, the trained model predicts the model's performance based on its performance in training. Validation sets consist of k parts that are used k times each [33].

3.3 Data Collection

In this project, tweets were collected using Twitter APIs and the Tweepy library. The Tweepy Python

module allows developers to access tweets, retweets, and timestamps of Twitter content. As well as ChatGPT's popularity throughout the world, Arab tweets about these polarizing incidents were chosen for analysis. Two thousand two hundred forty-seven tweets were collected between August 20, 2022, and April 24, 2022, during which 1,473 tweets were categorized as positive and 774 as negative by language specialists.

3.4 Data Preprocessing

A crucial aspect of NLP is the preparation of data to create coherent, meaningful structures out of unprocessed text. To improve the data quality, we used a variety of preprocessing techniques, such as removing duplicates, retweets, punctuation, hashtags, and stop words. Detailed information about each of these tasks will be presented in the sections that follow.

3.4.1 Removing Retweets and User Tags

The act of retweeting involves distributing someone else's tweets. Sharing such tweets creates duplicates, which can hinder model training and decrease accuracy. The dataset needed to be cleaned up by eliminating retweets.

3.4.2 Emoji Conversion

Small graphic symbols called emojis are often used to express sentiments or viewpoints. Translating these images into textual representations enhanced our model training.

3.4.3 Remove Hashtags, Punctuation and Numbers

It is common for social media platforms to use hashtags to organize and search for content related to a particular hashtag. Social media tools like hashtags (#) can be very powerful because they precede keywords. In spite of this, machine learning models do not require it, so it has been removed. In addition, punctuation marks, numeric values, and repetitive words have been removed. In addition to reducing memory usage, this also expedited learning.

3.4.4 Tokenization

With tokenization, natural language technology is used to break down text into smaller units composed of individual words. To facilitate sentiment analysis features' extraction, tokenization is required.

3.4.5 Data Balancing (SMOTE)

A machine learning algorithm does not prefer classifications that are unbalanced since they result in unsatisfactory results that favour one category over another. SMOTE (Synthetic Minority Oversampling Technique) [34] increased the number of tweets because it covers more positive than negative tweets. When negative data are compared to positive data using random examples, 619 negative tweets were compared to 1156 positive tweets after preprocessing. Based on K, SMOTE selects the nearest neighbours by random selection using 2312 tweets. Of these, 1156 are positive tweets, and 1156 are negative tweets.

3.4.6 Feature Extraction

Furthermore, TF-IDF (Term Frequency-Inverse Document Frequency) extracts features from text and converts it into vectors. Extraction of text features is widely used with this method. By multiplying each word in the text by its repetitions, the TF-IDF can be calculated. This data is then processed and trained using machine learning algorithms, as shown in algorithm 1.

4 MACHINE LEARNING ALGORITHMS

We have created multiple machine-learning models, built them, and evaluated them. Training datasets accounted for 80% of the datasets, while testing datasets accounted for 20%. An F1 score and accuracy, precision, and recall were used to measure each paradigm's effectiveness. In Sklearn, all machine learning algorithms use a default set of hyperparameters. The accuracy of each machine learning algorithm was raised by using three optimization algorithms: Grid search, random search, and Bayesian optimization. A total of four machine-learning algorithms were used to handle Arabic texts: RF, NB, SVM, and KNN. Despite the high efficiency of SVMs in classification, KNNs, NBs, and RFs should not be neglected [35]. In the following section, we discuss the techniques that have been implemented.

4.1 Logistic Regression

The logistic regression algorithm is used in machine learning to predict the probability of events. In a comparison between independent and dependent variables, the independent variable determines the probability of the event happening [35]. An input

variable is modelled in this way to produce a categorical outcome.

$$\text{Logistic} = \frac{1}{1 + e^{-z}}. \quad (1)$$

4.2 Naive Bayes

This study also utilized NB to classify the data [35]. A Bayes' theorem-based classifier uses conditional probability to calculate the likelihood of data instances belonging to a specific category:

$$P(Y|x_1, x_2, \dots, x_k) = \frac{P(Y) * \prod p(x_i|Y)}{p(x_1, x_2 \dots x_k)}. \quad (2)$$

4.3 Support Vector Machine (SVM)

It is a robust model for delineating categories and assigning data to an appropriate category using support vector machines (SVMs) [36]. Hyperplanes are dividing lines between classes used in SVM algorithms, also known as decision boundaries. A positive hyperbolic chart, a negative hyperbolic chart, and an optimal hyperbolic chart are used in SVM:

$$w * x + b = 1, \text{ For Posotive Hyperplane}, \quad (3)$$

$$w * x + b = -1, \text{ For Negative Hyperplane}, \quad (4)$$

$$w * x + b = 0, \text{ For Optimal Hyperplane}. \quad (5)$$

4.4 Random Forest

As part of RF [37], decision trees are developed for classification purposes. Unlike most trees, the random forest model uses several hyperparameters for choosing class labels, including maximum depth, number of estimators, minimum leaf sample size, etc.

5 RESULTS AND DISCUSSION

In this bar chart, six machine learning models are compared based on their accuracy. Naïve Bayes achieves the lowest accuracy at around 70%, while the highest is reached by Support Vector Classifiers at near 95%. Additionally, logistic regression, ridge classifier, and random forest perform well, while decision trees are in between. The SVC model performs best overall.

This bar chart compares Naïve Bayes and Decision Tree classifiers based on various tuning techniques. With Decision Trees, accuracy consistently exceeds that of Naïve Bayes, with 90 90% accuracy being achieved across all methods, whereas Naïve Bayes

remains between 70% and 80% as illustrated in Figures 2 and 3.

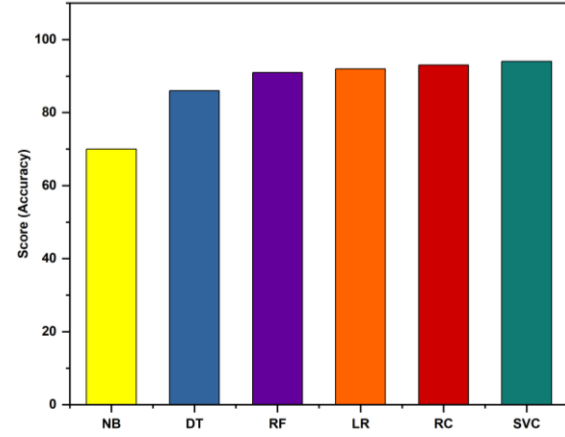


Figure 2: Comparison of model accuracy scores.

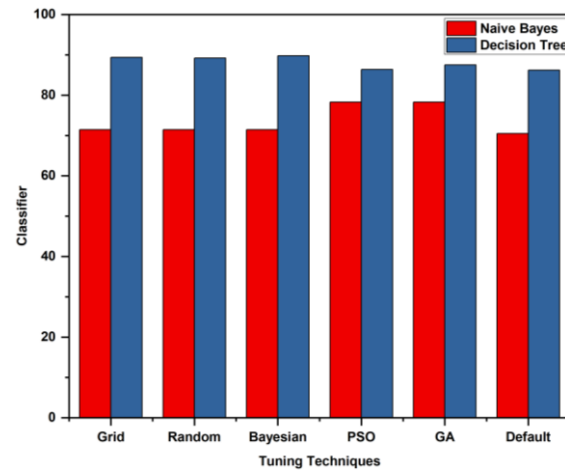


Figure 3: Impact of tuning techniques on classifier accuracy.

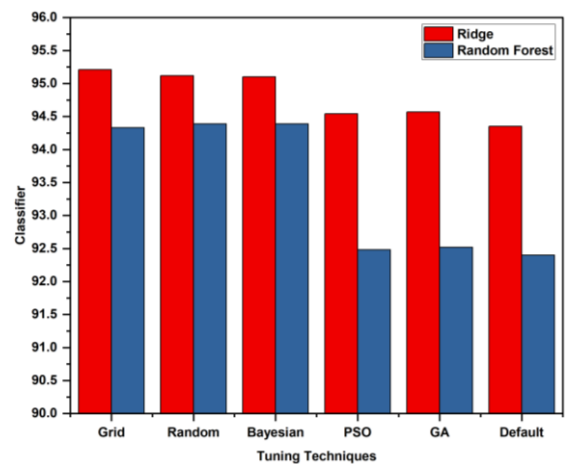


Figure 4: Classifier accuracy comparison across tuning techniques.

The bar chart (Fig. 4) compares the accuracy of Ridge and Random Forest classifiers using various tuning methods. Random Forest consistently performs better than Ridge, maintaining between 92.5 and 94% accuracy, while Ridge maintains between 94 and 95.5% accuracy.

A bar chart (Fig. 5) compares the accuracy of Support Vector Regression and Logistic Regression across tuning methods. With Grid, Random, and Bayesian tuning, Support Vector overperforms Logistic Regression, maintaining a higher level of accuracy across all methods.

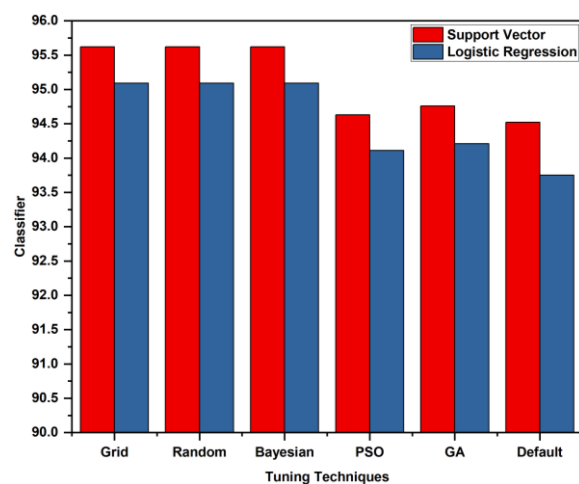


Figure 5: Accuracy comparison of support vector and logistic regression across tuning techniques.

6 CONCLUSIONS

In this paper, hyperparameter optimization techniques are employed to design and develop an enhanced machine-learning framework specifically tailored for Arabic sentiment classification. By effectively combining traditional machine learning algorithms with advanced deep learning architectures, our approach significantly improves the accuracy and reliability of sentiment predictions compared to conventional methods. Through extensive and rigorous experimentation, we demonstrate that Support Vector Machines (SVM) and Ridge Classifiers consistently achieve superior performance, particularly when their hyperparameters are meticulously tuned. Our results clearly indicate that strategic optimization of hyperparameters substantially boosts classifier performance by fine-tuning models to precisely match data characteristics. Additionally, the robustness and generalizability of

the classification models were further strengthened by employing preprocessing techniques such as SMOTE to address data imbalance issues and TF-IDF for effective feature extraction and representation. Based on our comprehensive findings, we conclude that employing sophisticated optimization strategies is crucial for obtaining state-of-the-art performance and robust outcomes in Arabic sentiment analysis, especially considering the complexities inherent in Arabic linguistic structures and cultural nuances.

REFERENCES

- [1] M. A. Hassonah, R. Al-Sayyed, A. Rodan, A. Al-Zoubi, I. Aljarah, and H. Faris, "An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter," *Knowl.-Based Syst.*, vol. 192, p. 105353, 2020.
- [2] M. Tubishat, M. A. M. Abushariah, N. Idris, and I. Aljarah, "Improved whale optimization algorithm for feature selection in Arabic sentiment analysis," *Appl. Intell.*, vol. 49, no. 5, pp. 1688–1707, May 2019, doi: 10.1007/s10489-018-1334-8.
- [3] I. Guellil, F. Azouaou, and M. Mendoza, "Arabic sentiment analysis: studies, resources, and tools," *Soc. Netw. Anal. Min.*, vol. 9, no. 1, p. 56, Dec. 2019, doi: 10.1007/s13278-019-0602-x.
- [4] T. N. Prakash and A. Aloysius, "A comparative study of lexicon based and machine learning based classifications in sentiment analysis," *Int. J. Data Min. Tech. Appl.*, vol. 8, no. 1, pp. 43–47, 2019.
- [5] A. Singh et al., "Smart traffic monitoring through real-time moving vehicle detection using deep learning via aerial images for consumer application," *IEEE Trans. Consum. Electron.*, vol. 70, no. 4, pp. 7302–7309, Nov. 2024, doi: 10.1109/TCE.2024.3445728.
- [6] M. E. M. Abo, R. G. Raj, and A. Qazi, "A review on Arabic sentiment analysis: state-of-the-art, taxonomy and open research challenges," *IEEE Access*, vol. 7, pp. 162008–162024, 2019, doi: 10.1109/ACCESS.2019.2951530.
- [7] I. Gupta and N. Joshi, "Enhanced Twitter sentiment analysis using hybrid approach and by accounting local contextual semantic," *J. Intell. Syst.*, vol. 29, no. 1, pp. 1611–1625, Dec. 2019, doi: 10.1515/jisys-2019-0106.
- [8] P. Rani, U. C. Garjola, and H. Abbas, "A predictive IoT and cloud framework for smart healthcare monitoring using integrated deep learning model," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 53–65, 2024.
- [9] M. Hijawi and Y. Elsheikh, "Arabic language challenges in text based conversational agents compared to the English language," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 3, pp. 1–13, Jun. 2015, doi: 10.5121/ijcsit.2015.7301.
- [10] N. E. Aoumeur, Z. Li, and E. M. Alshari, "Improving the polarity of text through word2vec embedding for primary classical Arabic sentiment analysis," *Neural Process. Lett.*, vol. 55, no. 3, pp. 2249–2264, Jun. 2023, doi: 10.1007/s11063-022-11111-1.

- [11] N. V. Devi and R. Ponnusamy, "A systematic survey of natural language processing (NLP) approaches in different systems," *Int. J. Comput. Sci. Eng.*, vol. 4, no. 7, 2016. [Online]. Available: <https://www.researchgate.net/publication/322465983>.
- [12] P. Rani, S. P. Yadav, P. N. Singh, and M. Almusawi, "Real-world case studies: transforming mental healthcare with natural language processing," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Bijju, and G. C. Deka, Eds. IGI Global, 2025, pp. 303–324, doi: 10.4018/979-8-3693-4203-9.ch016.
- [13] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis," *Informatics*, vol. 8, no. 4, p. 79, Nov. 2021, doi: 10.3390/informatics8040079.
- [14] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proc. 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis*, San Jose, CA: ACM, Aug. 2007, pp. 56–65, doi: 10.1145/1348549.1348556.
- [15] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Micro-blogging as online word of mouth branding," in *CHI '09 Extended Abstracts on Human Factors in Computing Systems*, Boston, MA: ACM, Apr. 2009, pp. 3859–3864, doi: 10.1145/1520340.1520584.
- [16] P. Rani, D. S. Mohan, S. P. Yadav, G. K. Rajput, and M. A. Farouni, "Sentiment analysis and emotional recognition: enhancing therapeutic interventions," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Bijju, and G. C. Deka, Eds. IGI Global, 2025, pp. 283–302, doi: 10.4018/979-8-3693-4203-9.ch015.
- [17] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2002, pp. 79–86, doi: 10.3115/1118693.1118704.
- [18] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Philadelphia, PA, 2001, p. 417, doi: 10.3115/1073083.1073153.
- [19] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report*, Stanford Univ., vol. 1, no. 12, 2009.
- [20] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Language Technologies*, 2011, pp. 151–160. [Online]. Available: <https://aclanthology.org/P11-1016.pdf>.
- [21] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining lexicon-based and learning-based methods for Twitter sentiment analysis," *HP Lab. Tech. Rep.*, vol. 89, pp. 1–8, 2011.
- [22] P. Rani, S. Verma, S. P. Yadav, B. K. Rai, M. S. Naruka, and D. Kumar, "Simulation of the lightweight blockchain technique based on privacy and security for healthcare data for the cloud system," *Int. J. E-Health Med. Commun. (IJEHMC)*, vol. 13, no. 4, pp. 1–15, 2022.
- [23] B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," *arXiv preprint*, arXiv:cs/0409058, Sep. 2004, doi: 10.48550/arXiv.cs/0409058.
- [24] B. Liu, *Sentiment Analysis and Opinion Mining*. Springer Nature, 2022. [Online]. Available: <https://books.google.com/books?id=xYhyEAAAQBAJ>.
- [25] H. Nguyen, A. Veluchamy, M. Diop, and R. Iqbal, "Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches," *SMU Data Sci. Rev.*, vol. 1, no. 4, p. 7, 2018.
- [26] A. Bhatt, A. Patel, H. Chheda, and K. Gawande, "Amazon review classification and sentiment analysis," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 6, pp. 5107–5110, 2015.
- [27] T. K. Shivaprasad and J. Shetty, "Sentiment analysis of product reviews: a review," in *Proc. Int. Conf. Inventive Communication and Computational Technologies (ICICCT)*, IEEE, 2017, pp. 298–301. [Online]. Available: <https://ieeexplore.ieee.org/document/7975207>.
- [28] P. Rani and R. Sharma, "Intelligent transportation system for Internet of Vehicles based vehicular networks for smart cities," *Comput. Electr. Eng.*, vol. 105, p. 108543, 2023.
- [29] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data," *Ecol. Model.*, vol. 406, pp. 109–120, Aug. 2019, doi: 10.1016/j.ecolmodel.2019.06.002.
- [30] F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., *Automated Machine Learning: Methods, Systems, Challenges*. Cham: Springer, 2019, doi: 10.1007/978-3-030-05318-5.
- [31] G. Ansari, P. Rani, and V. Kumar, "A novel technique of mixed gas identification based on the group method of data handling (GMDH) on time-dependent MOX gas sensor data," in *Proc. Int. Conf. Recent Trends in Computing (ICRTC)*, Springer, 2023, pp. 641–654.
- [32] S. Andradóttir, "An overview of simulation optimization via random search," in *Handbooks in Operations Research and Management Science*, vol. 13, Elsevier, 2006, pp. 617–631, doi: 10.1016/S0927-0507(06)13020-0.
- [33] Y. Xu and R. Goodacre, "On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning," *J. Anal. Test.*, vol. 2, no. 3, pp. 249–262, Jul. 2018, doi: 10.1007/s41664-018-0068-2.
- [34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.
- [35] Y. Wang, Z. Zhou, S. Jin, D. Liu, and M. Lu, "Comparisons and selections of features and classifiers for short text classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 261, p. 012018, Oct. 2017, doi: 10.1088/1757-899X/261/1/012018.

- [36] “Applications of support vector machine (SVM) learning in cancer genomics,” *Cancer Genomics Proteomics*, vol. 15, no. 1, Jan. 2018, doi: 10.21873/cgp.20063.
- [37] G. Biau and E. Scornet, “A random forest guided tour,” *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.

SECTION 2

DATA TRANSPORT TECHNOLOGIES

Proceedings of the 13th International Conference on Applied Innovations in IT (ICAIIIT)

JULY 2025

Design and Development of an IoT-Blockchain Platform for Reliable Sensing Data Management

Abdul Nashar¹, Joko Tri Brata¹ and Arkan Adnan Imran²

¹Department of Government Science, Universitas Sulawesi Tenggara, 93126 Kendari, Indonesia

²Department of Computer Technique Engineering, Dijlah University College, 10021 Baghdad, Iraq
abdulnashar@un-sultra.ac.id, djokotribrata@un-sultra.ac.id, engarkana@gmail.com

Keywords: IoT-Blockchain Platform, Data Integrity, Smart Contracts, Sensor Data Management, Decentralized Security.

Abstract: This paper presents an IoT-Blockchain platform to ensure reliable sensor data management in IoT systems. By integrating blockchain technology, the solution enhances data security, transparency, and decentralization, addressing vulnerabilities inherent in centralized architectures. Blockchain's immutability and decentralized nature enable secure data sharing, foster stakeholder trust, and automate processes via smart contracts. The study evaluates key performance metrics, such as device registration time, sensing data storage efficiency, and query execution speed, demonstrating the platform's superiority over existing blockchain systems. Results highlight its robustness in real-time data handling, scalability, and auditability. The proposed framework showcases the transformative potential of combining IoT and blockchain, offering a trustworthy infrastructure for applications like smart cities, supply chains, and industrial automation. Ultimately, this work underscores how decentralized architectures can revolutionize IoT data management, ensuring integrity while optimizing operational efficiency. The research underscores blockchain's role in overcoming IoT security challenges while enabling next-generation decentralized and auditable data ecosystems.

1 INTRODUCTION

As the Internet of Things (IoT) expands, many devices become connected, generating a vast amount of data [1]. It is important to recognize, however, that this growth comes with significant challenges in terms of data integrity, security, and reliability. While this growth is significant, there are also significant challenges associated with it, including security, integrity, and reliability of data. Despite this growth, data integrity, security, and reliability continue to pose significant challenges. A decentralized, secure, and transparent environment can be provided by integrating blockchain technology with IoT to overcome these limitations [2]. This paper introduces a blockchain-based IoT platform designed to ensure reliable sensor data management. Since blockchain technology was initially recognized for its role in cryptocurrencies, some of its inherent security properties, such as immutability, decentralization,

and transparency, have gained interest across a wide range of industries. As a data integrity enhancement in IoT, blockchain can ensure unaltered and reliable sensor data and transactions by creating a tamper-proof record. Integration of IoT and blockchain not only addresses security concerns, but also enables new functionalities, such as secure data sharing, improved trust among stakeholders, and automated processes through smart contracts [3].

Internet of Things (IoT) combined with blockchain technology is one promising solution for handling data integrity, security, and transparency challenges in the digital age. A wide range of sectors, including healthcare, smart cities, agriculture, and industrial automation, are increasingly deploying IoT systems that collect and exchange data. The Internet of Things allows for the monitoring of vast quantities of sensory data in real-time, but data management and trustworthiness remain significant challenges.

Cyberattacks and lack of transparency are common problems associated with IoT data

management. Also, it is imperative to ensure accuracy and reliability when it comes to data, especially when monitoring the environment and providing healthcare. In order for IoT to be widely adopted, data collected by IoT sensors must be accurate, secure, and accountable. Technology based on IoT can be used to create smart homes, smart transportation systems, or smart manufacturing systems, for example [4]. Technology innovations in embedded computing and networking have made it possible to create large-scale autonomous

IoT systems [5]. Data produced and exchanged in the IoT system is widely regarded as being both safety-critical and privacy-sensitive. This makes the network more important than ever. Cyberattacks are particularly problematic due to the unattended nature of wireless sensor networks [6]. An IoT solution currently connecting to cloud servers is centralized. The elastic computation and data management capabilities of this solution are excellent, but there are still some security concerns. The Internet of Things infrastructure is susceptible to single points of failure. To ensure the security of IoT devices, a tamper-proof environment and a fault-tolerant network are required [7].

For the access control of IoT devices, researchers are investigating using blockchain technology as a decentralized technology, which could be more suitable [8]. Data from multiple sources is tracked, verified, executed, and stored using a distributed database, which is composed of numerous peers. A number of high-level use cases have already been developed to illustrate this idea, including intelligent transportation systems [9], medical record-keeping systems, decentralized web applications, and prediction systems. The benefits of blockchain technology include high transparency, high security, better traceability, low costs, and the absence of third parties [10].

Decentralized, immutable, and transparent ledger systems like blockchain have gained traction as robust solutions to these issues. Data integrity, traceability, and security can be ensured through blockchain technology, as well as data management and validation through IoT. In addition to improving the reliability of data, this combination of IoT and blockchain can also create a secure, automated, distributed, and distributed data management ecosystem. Using blockchain technology, the proposed platform ensures the authenticity and security of IoT data by leveraging its data immutability and decentralized consensus functions. Smart contracts are also implemented on the platform to automate processes, enabling real-time decision-

making and enhanced data tracking. It is our goal to develop a method for managing data from the Internet of Things that is scalable, secure, and efficient, with applications across various industries dependent upon accurate and trustworthy data.

2 LITERATURE REVIEW

Because of its robust security features, blockchain has become increasingly popular as a platform for sharing IoT data. In [11], "Sensor-Chain" is described as a lightweight, scalable and secure blockchain framework designed to support Internet of Things applications. It is becoming increasingly difficult for IoT sensor devices to scale as blockchain chains grow in size. A potential solution is presented in this paper for improving mobile IoT device integration with blockchain technology. IoT devices have diverse capabilities, and even with efficiency gains, balancing the framework will be difficult.

A study showed that IoT devices and their interactions with services could be managed computationally trust through blockchain technology and reputation management systems [12]. The reward-penalty mechanism is used to establish trust architecture. It is possible that the proposed reward penalty system may be difficult to implement and maintain in spite of its promise. A user's security and costs need to be considered. Among the algorithms proposed in [13] include Proof of Authentication (PoA) and Proof of Work (PoW), which are suitable for edge computing and IoT setups with limited resources. Despite addressing efficiency concerns associated with PoW, PoA may have some limitations due to its specific authentication mechanisms [14]. Despite the reported latency of 3s, certain real-time applications might still be adversely affected. The importance of understanding how algorithms trade off performance and latency cannot be overstated.

According to the author, it implements a lightweight consensus protocol for Internet of Everything (IOE) environments. Power-intensive consensus algorithms are replaced by simple, easy-to-use algorithms that address the resource limitations of IoT devices. Due to the increasing number of IoT devices, 148.89 ms of latency can have a negative impact on performance and efficiency [15].

With the introduction of Wireless Sensor Networks (WSNs), a new authentication protocol was introduced. Data security in environments with resources and potentially distrustful actors is a primary objective of this project. Blockchain's security features are taken advantage of in this

protocol, but sensor nodes may feel overburdened as a result. In a similar report, WSN found that blockchain technology improved IoT data management and security. The architecture offers several advantages, including distributed storage of data, immutability, decentralization, and traceability. As a result of the architecture, data storage is distributed, mutable, decentralized, and traceable [16].

Data is collected and exchanged by physical devices through sensors, software, and networks in the Internet of Things [17]. Sensors, industrial machines, sensors, and controllers are all interconnected to automate, monitor, and make decisions across various domains using interconnected devices. Data generated by IoT devices creates a great deal of opportunity for innovation and efficiency because of the proliferation of these devices.

Transactions are recorded securely, transparently, and immutably using blockchain technology. Peer-to-peer ledgers ensure no single entity owns the ledger since each participant keeps a copy [18]. Cryptographic hash functions ensure data integrity along with consensus mechanisms on blockchains [19].

3 METHODOLOGY

Every block on a blockchain is time stamped, and every participant maintains it. Transactions are aggregated in blocks. Each block in a blockchain is cryptographically linked to the previous block by a hash value computed using the previous block's digital signature. Unlike traditional data storage systems, blockchain data is immutable: it can only be added at the end of a chain, so transactions cannot be modified once they are in place. A third-party authority does not need to be consulted to trust the occurrence of transactions. Historical transaction chains are immutable, providing non-reputability. Digital signatures and cryptography are used to protect the blockchain, verify identity, and control access. Originally, blockchains were used to record monetary transactions and were introduced by Bitcoin. Blockchains of the second generation provide a general, programmable infrastructure that stores computation results in a public ledger.

A smart contract [20] is an autonomous program that runs across a blockchain. It is possible to define entire business processes using smart contracts,

including triggers, conditions, and conditions. Escrow services with smart contract capabilities can, for example, hold funds until smart contract obligations have been met. In Ethereum, smart contracts are considered to be first-class elements of a second-generation blockchain. Thus, the blockchain is an immutable, decentralized, auditable, and completely decentralized technology. Automated program execution could be implemented with blockchain smart contracts.

3.1 Peer-to-Peer File System

Globally distributed file systems have been the subject of extensive research. System success has been seen in some cases, while failure has been seen in others. A total of over 100 million simultaneous users are supported by Napster, KaZaA, and BitTorrent in the industry. There is currently a push towards developing decentralized, global, and low-latency file systems. There is no doubt that peer-to-peer (P2P) schemes were conceived to replace the client-server model, which was implemented to support small, distributed environments with more powerful servers. Synchronous communication is the primary characteristic of peer-to-peer networks. A peer can serve as a client as well as a client. Files can be shared between clients and servers using P2P systems. Peers are able to communicate directly through it. Rather than all requesting files from the server simultaneously, peers can share files in pieces. As a result, sharing files has become much more efficient and scalable.

3.2 Blockchain Architectural Design Considerations

An immutable, replicated, and synchronized distributed ledger (DLT) based on P2P architecture between untrusted parties can be defined as blockchain [21], [22]. As part of the consensus protocol, miners verify and validate data and identities for transactions. Traditionally, blockchain technology has been used for digital currencies; however, it is increasingly being integrated into IoT solutions [23], [24]. Integrated with IoT, blockchain provides efficient mechanisms for identifying nodes, safeguarding authentication and communication, securing distributed storage, and securing resources. Several industries can benefit from a combination of blockchain and IoT, including SC, agriculture, healthcare, smart grids, smart homes, and automotive.

3.3 Evaluate the Conceptual Model

After the conceptual model has been developed, it is evaluated by an expert focus group. This process could include presenting the model to a group of experts and getting their feedback, insights, and suggestions. To develop the simulation environment, the conceptual model can be refined and improved based on the feedback gained during this process. It is crucial to establish clear objectives for this evaluation. Our objectives were evaluated using SMART criteria in order to determine their feasibility and

effectiveness [25]. As a result, our objectives are specific, measurable, attainable, relevant, and time-bound. To validate our conceptual model, we conducted an expert survey and focus group with ten experts in the blockchain and IoT fields. Evaluation of the conceptual model was conducted using both approaches. The conceptual model took into account the needs of all relevant stakeholders by engaging experts and collecting their feedback.

3.4 The Proposed Blockchain-Connected Gateway

An illustration of the BC gateway architecture can be found in Figure 1 of this document. Using blockchain accounts, the BC gateway identifies users and preferences. Users do not have to register for every BC gateway individually. They can connect to multiple BC gateways simultaneously using the same account. The Android and IOS platforms now support the Bluetooth Low Energy specification, which has become the de facto standard for smartphones when it comes to communicating with IoT devices and wearable's. Researchers examine the case of a user accessing nearby BLE-enabled devices. BLE gateways provide REST-like interfaces for accessing BLE-based devices, according to current state-of-the-art:

- Devices can be locked/unlocked using Bluetooth Low Energy.
- Explore the gateways' connected devices and the services they offer.
- Invoke a BLE device's characteristic by sending read or write requests.
- A BLE device displays certain characteristics when notified.

IoT devices connected to BC gateways will remain connected. A gateway can act as an intermediary that forwards requests to specific devices and integrates their responses. BC gateways

are also able to manage privacy preferences for IoT devices by providing interfaces for privacy preferences.

Device administrators can register devices with BC gateways by following the steps in Figure 2. The administrator creates a smart contract for an IoT device by creating a transaction to store information about the device and its privacy policies.

3.5 Device Binding

A Device Manager containing the addresses of the devices as well as privacy policies can then be created by the administrator. A device's Device Manager can provide information about the device and its privacy policies, as illustrated in Figure 2. Users can also listen to device update events.

Device Managers allow device administrators to obtain smart contract addresses. A device administrator can then provide a gateway administrator with the address of a device. Gateway administrators can request the binding of devices to BC Gateway Managers by providing the Device Manager's address.

3.6 Smart Contract Component

Decentralized data access control is a feature of Ethereum smart contracts since they are executed on the blockchain. Unlike smart contracts, which store data for free, blockchains store only hashes of data. Main data is stored and encrypted using SGX. In a smart contract, user and device registration are all included, as well as read/write access policies.

3.6.1 User Registration

User registration is facilitated by this module, which makes use of Ethereum user registration system. An Ethereum user generates a pair of public and private keys in order to identify themselves on the network. In smart contracts, devices can be registered, and data can be accessed using a private key.

3.6.2 Device Registration

The identifier for the IoT device can be provided by authenticated users when registering their devices. In smart contracts, user-owned devices are mapped to an owner's address on the blockchain (the owner's address = a list of the device's identifiers).

3.6.3 Access Policies for Writing Data

Data wished to be stored on the blockchain must be submitted by the device with the owner's address and

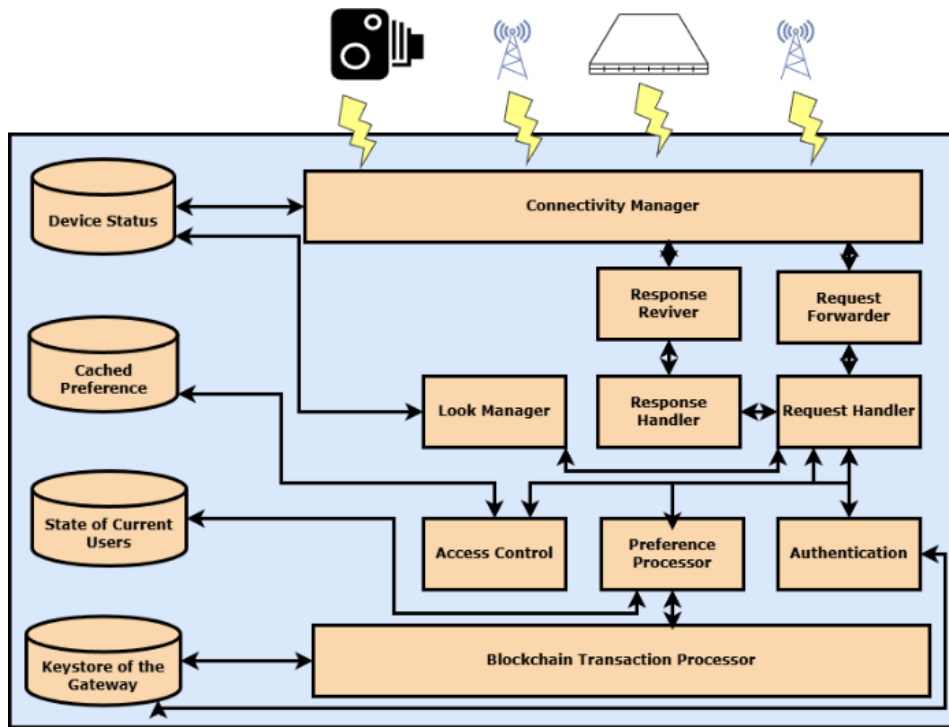


Figure 1: The architecture of a BC gateway.

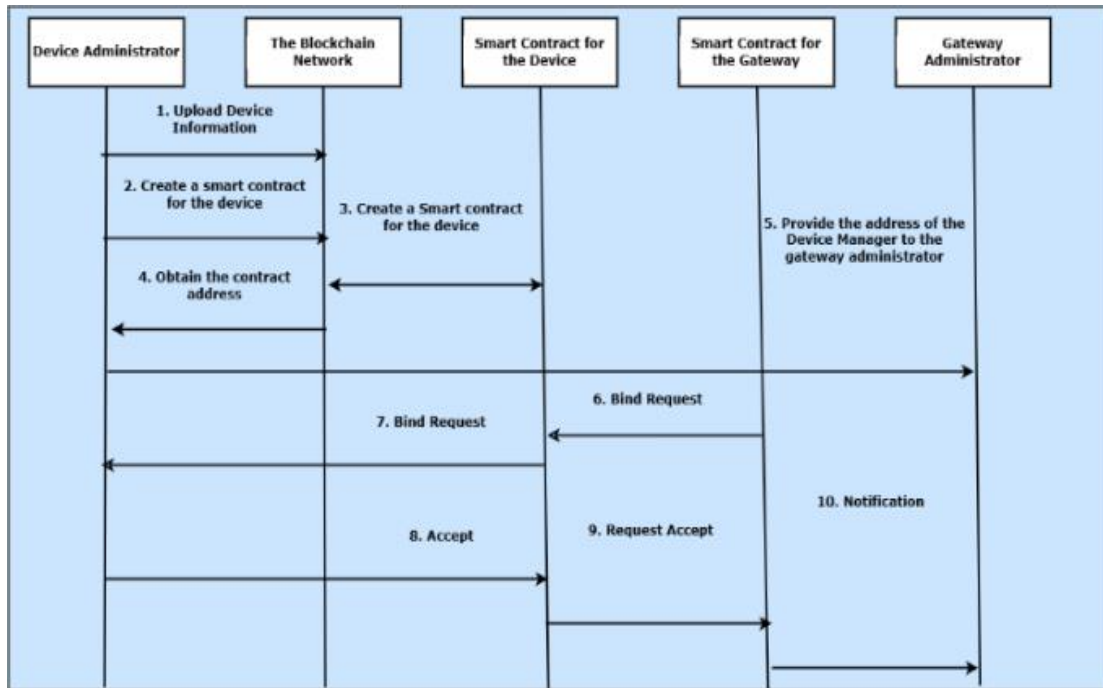


Figure 2: The device registration process.

the device's ID associated with it. A combination of owner address and device ID can allow us to uniquely store all data related to each device as shown in [(owner address, device ID) = list of device data]. A hash map consists of a list of hashes generated by devices. To allow data to be written to the contract, the smart contract verifies that the owner's address matches the ID of the device. The contract can only be written by the device owner.

3.6.4 Device Data Read Access Policy

It is the responsibility of the device owner to grant permission to unauthorized third parties wishing to access a device's data. As part of the request, we will need to know the user's address and device ID. Throughout smart contracts, keys and values are stored as a hash map consisting of the address of the device owner, the identifier of the device, and the list of users of the device. Using this formula, we have (owner, device id, third-party address) = bool access). Using hash maps, only registered third parties are granted access to data by checking if the requesting user is authorized to access it.

4 RESULTS AND DISCUSSION

Figure 3 shows the results of a study examining service execution time costs associated with device registration. This experiment registered 50, 150, 250, and 500 devices with the proposed platform. Using Hyperledger [26], the simulation recorded two minimums and a maximum for 50 devices: 2260 ms for the minimum, 2284 ms for the average, and 2373 ms for the maximum. The minimum and average times for 150 devices were 2255 ms, 2333 ms, and 2800 ms, respectively. The minimum time taken by the 250 devices was 2252 milliseconds, the average time was 2583 milliseconds, and the maximum time was 3002 milliseconds. Lastly, we found that the minimum time recorded was 2265 milliseconds, the average time was 2921 milliseconds, and the maximum time was 4011 milliseconds for the 500-device group.

The second study investigated the time required to process sensing data on a blockchain network. Using HTTP clients, each device could access the REST server to request sensor readings. REST servers retrieve execution results from blockchains and return them to devices after appending sensing data. According to Figure 4, sensor reading transactions take a long time to execute. A ten-fold repetition of each test was performed with multiple concurrent

clients at different levels of resource utilization. Compared to well-known blockchain platforms, this proposed platform performs significantly better. The study was limited by the fact that it involved a relatively small network with only four peers.

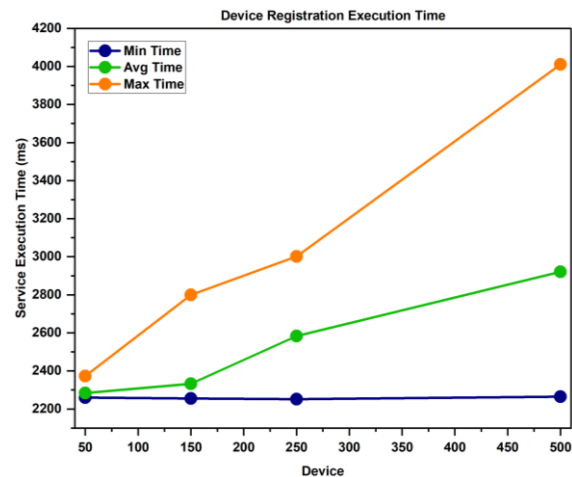


Figure 3: Analyze the performance of service execution (ms) versus no. of devices.

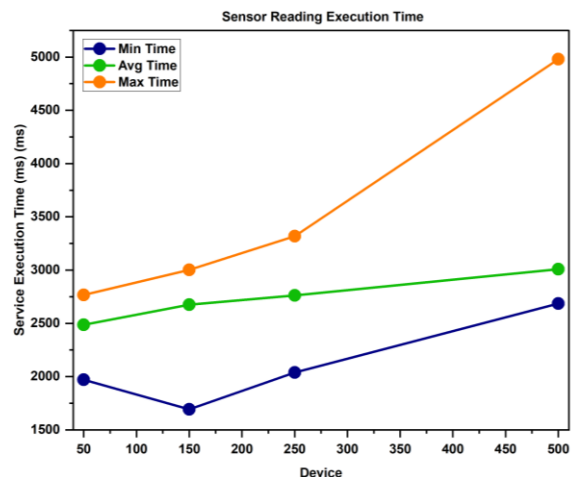


Figure 4: Sensor reading execution time (ms) versus no. of device.

A comparison of the execution time for querying data records on the blockchain is shown in Figure 5, which examines the execution time for 500 to 10,000 records of data. A random selection of 10 system resource utilization levels was used to determine the proposed platform's minimum, average, and maximum delay times for retrieving sensor records. Based on worst-case performance, the maximum delay would be 855 milliseconds with 10,000 records, the minimum delay would be 608 milliseconds, and the average delay would be 754 milliseconds.

According to the figure, the round-trip latency was greatly influenced by the size of the data records. As a result of the low increase, it was even possible to ignore the impact on the user's experience.

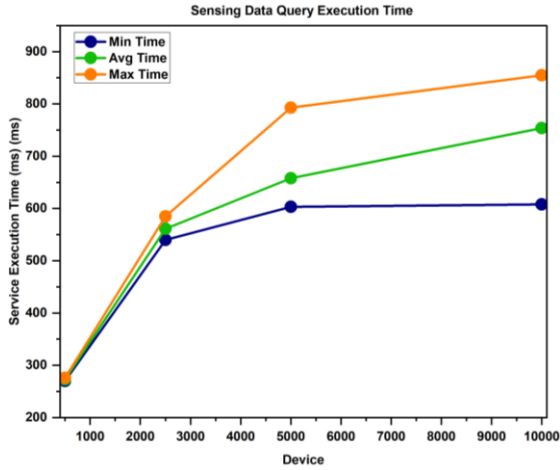


Figure 5: Sensing data execution time (ms) versus no. of device.

A comparison was made between this system and the proposed system. Simulating the environment from which the system operated [27]. It was simulated that 960 transactions would be executed in a network of 50 peers over 60 seconds. In the network, the processing time metric measures how long it takes to verify new blocks. According to Figure 6, the processing overhead was evaluated. We consistently achieved lower processing overhead than the selected method, with a number of blocks ranging from 10 to 60, as shown in the graph. A 23% reduction in processing time was achieved with the proposed approach.

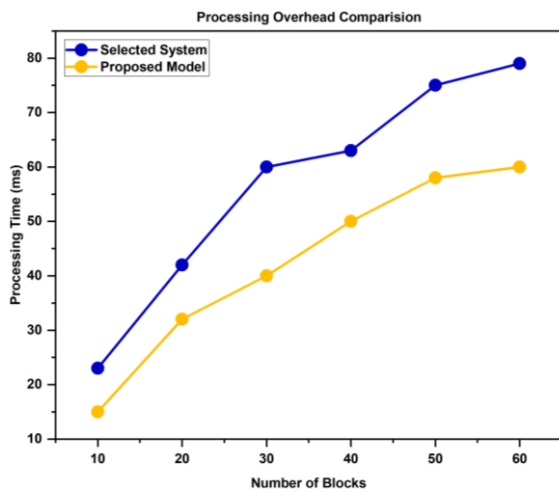


Figure 6: Processing time (ms) versus no. of blocks.

5 CONCLUSIONS

In this study, we developed an innovative IoT-Blockchain integrated platform designed to significantly enhance security, integrity, and transparency in IoT data management. Leveraging blockchain technology, the platform ensures robust data protection, providing tamper-proof mechanisms that maintain the authenticity and immutability of sensor-generated data. Performance evaluations revealed that our proposed solution registers IoT devices and transmits sensor data considerably faster and more reliably compared to traditional centralized systems. The incorporation of blockchain technology not only guarantees secure and reliable storage of data but also establishes a transparent audit trail, facilitating trustworthy data exchanges among various stakeholders. Despite these promising outcomes, the experiments conducted thus far have identified critical limitations, particularly regarding the small-scale network used in the initial testing scenarios. Consequently, there is a pressing need to optimize blockchain consensus algorithms tailored explicitly for resource-constrained IoT devices, which often have limited processing power and energy availability. As the platform matures and evolves, future work will focus extensively on scaling the network, refining consensus mechanisms for enhanced efficiency, and conducting rigorous evaluations with larger networks and diverse real-world applications. These enhancements will further validate the platform's capability to meet the increasing demands of complex IoT ecosystems and drive broader adoption in smart industries.

REFERENCES

- [1] B. Bhola, E. S. B. Manjula, P. Rani, and M. H. Falaah, "Quality-enabled decentralized dynamic IoT platform with scalable resources integration," *IET Commun.*, 2022.
- [2] E. A. Shammar, A. T. Zahary, and A. A. Al-Shargabi, "A Survey of IoT and Blockchain Integration: Security Perspective," *IEEE Access*, vol. 9, pp. 156114–156150, 2021, [Online]. Available: <https://doi.org/10.1109/ACCESS.2021.3129697>.
- [3] P. Shi, H. Wang, S. Yang, C. Chen, and W. Yang, "Blockchain-based trusted data sharing among trusted stakeholders in IoT," *Softw. Pract. Exp.*, vol. 51, no. 10, pp. 2051–2064, Oct. 2021, [Online]. Available: <https://doi.org/10.1002/spe.2739>.
- [4] P. Rani, S. P. Yadav, A. Singh, M. Almusawi, M. H. Falaah, and M. A. Farouni, "Federated Learning-Based Misbehavior Detection for the 5G-Enabled Internet of Vehicles," *IEEE Trans. Consum. Electron.*, vol. 70, no. 2, pp. 4656–4664, May 2024.

- [Online]. Available: <https://doi.org/10.1109/TCE.2023.3328020>.
- [5] P. Rani and M. H. Falaah, "Real-Time Congestion Control and Load Optimization in Cloud-MANETs Using Predictive Algorithms," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 66–76, 2024.
 - [6] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini, "Security, privacy and trust in Internet of Things: The road ahead," *Comput. Netw.*, vol. 76, pp. 146–164, Jan. 2015, [Online]. Available: <https://doi.org/10.1016/j.comnet.2014.11.008>.
 - [7] M. A. Khan and K. Salah, "IoT security: Review, blockchain solutions, and open challenges," *Future Gener. Comput. Syst.*, vol. 82, pp. 395–411, 2018.
 - [8] T. M. Fernández-Caramés and P. Fraga-Lamas, "A Review on the Use of Blockchain for the Internet of Things," *IEEE Access*, vol. 6, pp. 32979–33001, 2018.
 - [9] Y. Yuan and F.-Y. Wang, "Towards blockchain-based intelligent transportation systems," in 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil: IEEE, Nov. 2016, pp. 2663–2668, [Online]. Available: <https://doi.org/10.1109/ITSC.2016.7795984>.
 - [10] A. Singh, P. Rani, M. H. Falaah, and S. P. Yadav, "Blockchain-Based Lightweight Authentication Protocol for Next-Generation Trustworthy Internet of Vehicles Communication," *IEEE Trans. Consum. Electron.*, vol. 70, no. 2, pp. 4898–4907, May 2024, [Online]. Available: <https://doi.org/10.1109/TCE.2024.3351221>.
 - [11] Y. Wang, M. Singgih, J. Wang, and M. Rit, "Making sense of blockchain technology: How will it transform supply chains?," *Int. J. Prod. Econ.*, vol. 211, pp. 221–236, 2019.
 - [12] A. A. Battah, Y. Iraqi, and E. Damiani, "A trust and reputation system for IoT service interactions," *IEEE Trans. Netw. Serv. Manag.*, vol. 19, no. 3, pp. 2987–3005, 2022.
 - [13] D. Puthal, S. P. Mohanty, P. Nanda, E. Kougianos, and G. Das, "Proof-of-Authentication for Scalable Blockchain in Resource-Constrained Distributed Systems," in 2019 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA: IEEE, Jan. 2019, pp. 1–5, [Online]. Available: <https://doi.org/10.1109/ICCE.2019.8662009>.
 - [14] P. Rani, P. N. Singh, S. Verma, N. Ali, P. K. Shukla, and M. Alhassan, "An implementation of modified blowfish technique with honey bee behavior optimization for load balancing in cloud system environment," *Wirel. Commun. Mob. Comput.*, vol. 2022, pp. 1–14, 2022.
 - [15] X. Sun and N. Ansari, "Dynamic resource caching in the IoT application layer for smart cities," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 606–613, 2017.
 - [16] A. Reyna, C. Martín, J. Chen, E. Soler, and M. Díaz, "On blockchain and its integration with IoT. Challenges and opportunities," *Future Gener. Comput. Syst.*, vol. 88, pp. 173–190, Nov. 2018, [Online]. Available: <https://doi.org/10.1016/j.future.2018.05.046>.
 - [17] J. Lockl, V. Schlatt, A. Schweizer, N. Urbach, and N. Harth, "Toward Trust in Internet of Things Ecosystems: Design Principles for Blockchain-Based IoT Applications," *IEEE Trans. Eng. Manag.*, vol. 67, no. 4, pp. 1256–1270, Nov. 2020, [Online]. Available: <https://doi.org/10.1109/TEM.2020.2978014>.
 - [18] G. Pulkkis, J. Karlsson, and M. Westerlund, "Blockchain-Based Security Solutions for IoT Systems," in *Internet of Things A to Z*, 1st ed., Q. Hassan, Ed., Wiley, 2018, pp. 255–274, [Online]. Available: <https://doi.org/10.1002/9781119456735.ch9>.
 - [19] A. Singh, P. Rani, S. Verma, and S. P. Yadav, "Resilient wireless sensor networks in industrial contexts via energy-efficient optimization and trust-based secure routing," *Peer-Peer Netw. Appl.*, vol. 18, no. 3, p. 132, Jun. 2025, [Online]. Available: <https://doi.org/10.1007/s12083-025-01946-5>.
 - [20] S. Omohundro, "Cryptocurrencies, smart contracts, and artificial intelligence," *AI Matters*, vol. 1, no. 2, pp. 19–21, Dec. 2014, [Online]. Available: <https://doi.org/10.1145/2685328.2685334>.
 - [21] L. Da Xu, Y. Lu, and L. Li, "Embedding blockchain technology into IoT for security: A survey," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10452–10473, 2021.
 - [22] M. Maroufi, R. Abdoolee, and B. M. Tazekand, "On the Convergence of Blockchain and Internet of Things (IoT) Technologies," *J. Strateg. Innov. Sustain.*, vol. 14, no. 1, Mar. 2019, [Online]. Available: <https://doi.org/10.33423/jsis.v14i1.990>.
 - [23] P. Kumar, R. K. Singh, A. K. Sangaiah, M. Almusawi, P. Rani, and S. P. Yadav, "PPSF: A Privacy-Preserving and Secure Framework Using Blockchain-Based Machine-Learning for IoT-Driven Smart Cities," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 3, pp. 2326–2341, Jul. 2021, [Online]. Available: <https://doi.org/10.1109/TNSE.2021.3089435>.
 - [24] J. Lindley, "Food regulation and policing: innovative technology to close the regulatory gap in Australia," *J. Consum. Prot. Food Saf.*, vol. 17, no. 2, pp. 127–136, Jun. 2022, [Online]. Available: <https://doi.org/10.1007/s00003-022-01372-2>.
 - [25] M. R. Bataineh, W. Mardini, Y. M. Khamayseh, and M. M. B. Yassein, "Novel and Secure Blockchain Framework for Health Applications in IoT," *IEEE Access*, vol. 10, pp. 14914–14926, 2022, [Online]. Available: <https://doi.org/10.1109/ACCESS.2022.3147795>.
 - [26] H. Caliper, "Hyperledger Caliper," [Online]. Available: <https://www.hyperledger.org/projects/caliper>, [Accessed: 15 Jan 2019].
 - [27] A. Dorri, S. S. Kanhere, and R. Jurdak, "Towards an Optimized Blockchain for IoT," in *Proceedings of the Second International Conference on Internet-of-Things Design and Implementation*, Pittsburgh, PA, USA: ACM, Apr. 2017, pp. 173–178, [Online]. Available: <https://doi.org/10.1145/3054977.3055003>.

Blockchain-Enabled Data Analytics-Based Demand Service Optimization for Vehicular Safety

La Ode Bariun¹, Haeruiddin Haeruiddin¹ and Mahmoud A. Mahmoud²

¹*Universitas Sulawesi Tenggara, 93121 Kendari, Indonesia*

²*Department of Medical Instrumentation Techniques Engineering, Dijlah University College, 10021 Baghdad, Iraq
laodebariun@un-sultra.ac.id , haeruiddin@un-sultra.ac.id , mahmoud.abbas@duc.edu.iq*

Keywords: Blockchain Technology, Vehicular Safety, Data Analytics, Demand Service Optimization, Intelligent Transportation Systems.

Abstract: Vehicle safety and demand optimization are being explored using blockchain-enabled data analytics due to the rapid growth of connected vehicle technologies. This system establishes a robust and scalable framework for next-generation intelligent transportation systems (ITS), leveraging cutting-edge data analytics, IoT integration, and decentralized technologies to optimize resource allocation, enhance operational efficiency, and strengthen safety protocols. By harnessing real-time data processing and predictive modeling, the framework enables dynamic traffic flow optimization, reduces congestion, and minimizes environmental impact. Furthermore, the paper explores the integration of blockchain technology within vehicular networks, highlighting its transformative potential in ensuring secure, tamper-proof data exchange among vehicles, infrastructure, and service providers. Key applications include enhanced collision avoidance systems, automated toll collection, and fraud-resistant vehicle-to-everything (V2X) communication. The study also investigates blockchain's role in improving emergency response coordination by enabling instant, verifiable data sharing between first responders, traffic authorities, and connected vehicles. These advancements not only increase road safety and traffic management precision but also lay the foundation for fully autonomous and decentralized transportation ecosystems in smart cities.

1 INTRODUCTION

Smart vehicles have been evolving rapidly in recent years, driving the exploration of innovative technologies to enhance road safety and optimize vehicular operations. A great example of such an innovation is the combination of blockchain technology with data analytics, which has great potential for revolutionizing vehicular safety systems. With blockchain technology, safety-critical information can be exchanged securely, transparently, and decentralized in real-time. With advanced data analytics, this system can optimize demand service by dynamically allocating resources and safety measures according to demand, traffic patterns, and other contextual factors. To ensure vehicular safety, we need to be able to predict potential risks, make on-the-fly decisions, and allocate resources efficiently. Safety systems typically rely on centralized structures that can fail or be compromised. In addition to being tamper-proof, blockchain's inherent decentralized architecture

mitigates all of these risks by providing easily traceable and tamper-proof data for safety decisions. The use of data analytics, on the other hand, allows a vast amount of real-time data from vehicles, sensors, and infrastructure to be analyzed to extract meaningful insights.

Vehicle communication has undergone a paradigm shift with intelligent transportation systems (ITS) aimed at improving traffic efficiency, safety, and the overall driving experience [1]. Through the Internet of Vehicles (IoV), vehicles and infrastructure are interconnected at the roadside to facilitate this transformation. A multitude of aspects of vehicular safety and demand-responsive services can be optimized with IoV systems. There are several challenges associated with centralized data management systems, including security vulnerabilities, privacy concerns, and lack of transparency.

Data management and analytics in IoV environments can be improved by using blockchain technology, which provides a decentralized, secure,

and transparent platform [2]. In order to enhance vehicular safety and optimize demand-responsive services, blockchain technology allows vehicles, infrastructure, and service providers to share data securely. An overview of blockchain-enabled data analytics for demand service optimization in vehicular safety is presented in this paper, as well as key findings, challenges, and future research directions. Blockchain can provide access control mechanisms for vehicular networks, as well as secure data sharing [3] [4]. Data analytics techniques will also be applied to blockchain-enabled vehicular data to optimize services such as emergency response, traffic management, and personalized safety alerts. In addition to addressing scalability, interoperability, and regulatory issues related to implementing blockchain in IoV environments, the review will also address scalability challenges. Using blockchain technology, we can trace, track, and see how huge amounts of data are related and simplify the contents without cloud services. Below is a proposed taxi that is a regular taxi as well as a ride-hailing taxi. According to the Figure, this system consists of three main layers. Data collected from taxi drivers and

passengers in real-time is the first layer. In addition to enabling imports and exports of goods and services, sustainable transportation and logistics are important to related activities. Transport efficiency increases a good's (or service's) market potential. In addition to achieving other economic, social, and environmental objectives, low-carbon transportation systems can also contribute to the environment [5]. Public health and safety can be improved, access to transportation can be increased, congestion on roads and parking lots can be reduced, consumer costs can be lowered, economic growth can be boosted, and air and noise pollution can be reduced. Using existing techniques, transportation energy consumption can be reduced by 40-50% below 2010 demand, according to estimates [6]. In spite of this, there are a number of obstacles that prevent this potential from being fully realized. Sustainable transportation may be beneficial for reducing global climate change and boosting local economies because of its potential economic, social, and environmental benefits. There are a lot of security problems associated with the Internet of Vehicles (IoV) due to its resource limitations [7].

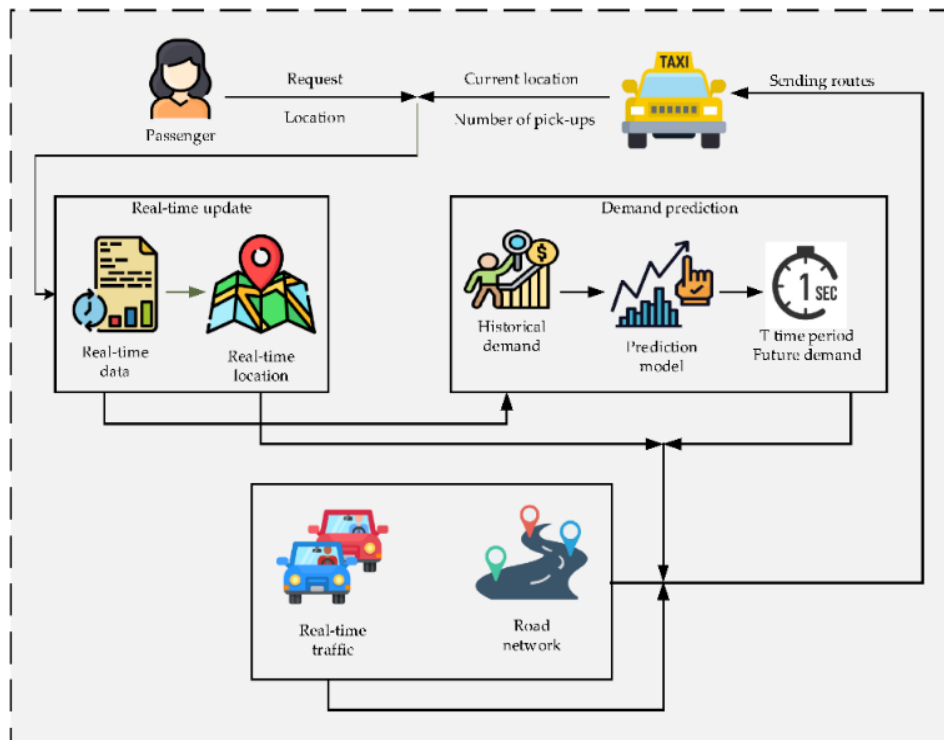


Figure 1: A comparison of regular Taxi and ride-hailing services.

2 LITERATURE REVIEW

The blockchain technology is revolutionizing a variety of industries, including transportation [8]. As a decentralized, immutable, and transparent technology, it is well suited to managing data in IoV environments. Several studies have examined blockchain's application in IoV, focusing on various aspects such as secure data sharing, access control, and trust management [9]. A key benefit of blockchain technology in IoV is its ability to facilitate secure data sharing between vehicles, roadside units, and other entities [10]. Author [3] suggests using smart contracts to protect data from unauthorized access in a consortium-based blockchain-based data-sharing scheme for vehicular edge computing and networks (VECONs).

Furthermore, a reputation-based system ensures that vehicles share data of high quality. A VANET can also use blockchain to establish trust between vehicles and RSUs. To mitigate MAC layer attacks on VANETs, the Author [11] proposed a trust-based approach using blockchain technology. Through their solution, genuine nodes can exchange information securely and privately, ensuring reliable security and privacy. According to Author [12] describes reputation-based voting as a method used to choose miners in an IoV enabled with blockchain technology. Using this scheme, you can prevent voting collusion in IoV and ensure that trusted miners are selected for data sharing.

2.1 Data Analytics for Vehicular Safety

A wide range of data is generated by IoV systems, which can be analyzed to gain valuable insights in order to improve vehicle safety. Analysis of vehicular data provides a means of identifying potential safety hazards, predicting traffic incidents, and alerting drivers in a timely manner. There have been several studies investigating how data analytics techniques can be applied to vehicular safety, focusing on factors such as incident detection, risk assessment, and driver behaviour [13].

The use of deep learning techniques has demonstrated promising results in predicting traffic accidents and providing drivers with timely warnings. An Incident and Evidence Management Framework (SIEMF) was proposed, combining deep learning and blockchain for predictability and secure evidence management [14]. Their system not only enhances road safety and enables forensic analysis but also addresses the challenge of secure evidence management and predictive modelling. Ensemble

learning models can predict the trustworthiness of VANET vehicles. In [15], the Author proposes a methodology for predicting the trustworthiness of vehicles using ensemble learning. Using their approach, vehicular networks become more transparent and trustworthy, reducing the risk of fraud or tampering. Through the use of data analytics techniques, IoV systems can improve vehicular safety by generating vast amounts of data [16].

2.2 Demand Service Optimization

In IoV, demand service optimization is achieved through data analytics and communication technologies. This allows real-time resource allocation and management. Drivers and passengers will be provided with personalized services, as well as the optimization of traffic flow and the management of energy consumption. IoV environments can benefit from secure and transparent demand service optimization enabled by blockchain technology [17]. Through blockchain technology, demand response (DR) can be managed securely and transparently in electric energy grids. According to the Author [18], blockchain applications are being explored in the power systems industry, focusing on distributed energy management (DEM), electric vehicles (EVs), and distributed power generation (DPP). The researchers found that blockchain could address some of the challenges related to modern energy grids, including the high penetration of renewable sources and the proliferation of electric vehicles. A study conducted by Author [19] examined blockchain technology in the management of IoV demand response, taking into account incentivization, privacy, and security. A systematic literature review was conducted to explore how blockchain technology can be used as a tool for IoV demand response management, as well as how blockchain can address issues of incentivization, privacy, security, and lack of interoperability.

An artificial intelligence, blockchain technology, and Internet of Things platform was developed to enable multilevel transactions for electric vehicle charging stations [20]. In addition to securing security and fairness while shortening latency, their platform utilizes green energy across microgrids, bidirectional discharging, and demand response bidding. Using blockchain in IoV environments, these studies demonstrate how resource allocation and management can be optimized by providing a secure, transparent, and distributed platform.

2.3 Blockchain-Enabled Data Analytics for Vehicular Safety: Synergies

Data analytics and blockchain technology can enhance vehicular safety and optimize demand-responsive services in IoV environments through the convergence of these technologies. A blockchain-based IoV system can be developed that addresses the challenges of traditional IoV systems by combining security, transparency, and data analytics. Blockchain technology enables IoV environments to share and process data securely and transparently, enhancing security and privacy [21].

The Author [22] presents practical prospects for blockchain-enabled federated learning in the IoV, including full security, privacy, and verifiability. By addressing federated learning's vulnerabilities to poisoning and reverse engineering attacks, they ensure that ITS services are trustworthy. Author [23] presented a 5G-VANET data access solution based on smart contracts and a blockchain oracle that preserves privacy.

3 METHODOLOGY

3.1 Predictability of Taxi Demand

Using $\prod max$ as a measure of maximum predictability, we define it in this section. In this study, we propose the $\prod max$ model to answer the following question:

Problem 1: As a result of a predictive algorithm and sequence of taxi demand D_n^i data at building blocks, we explore the maximum amount of predictability $\prod max$ (maximum potential accuracy) the prediction algorithm can achieve that.

3.1.1 Taxi Demand

There are two channels representing taxi pick-up and drop-off demand in a three-dimensional matrix for each grid of a city, so it can be predicted how many taxis will pick up and drop off at a given time interval. A CNN or LSTM can then be used to determine the spatial/temporal relationship between the two demands [24] - [27]. Analyzing these two demands together, however, reveals inconsistent periodic characteristics. During certain times, taxi demand decreases for pick-up and increases for drop-off, as shown in Figure 1. Combining these two demands into a single layer might obscure their specific

characteristics, reducing prediction accuracy. Research has attempted to uncover the hidden modes of change in traffic flows [29].

Taking into account different changes in taxi pickup and drop-off patterns, we built a single demand forecasting model.

3.1.2 Neural Network Architecture Selection

In spatial analysis, CNN has been used widely [27], [28]; however, it is restricted to Euclidean space, which means that taxi zones should be gridded, not irregular. As CNN extracts taxi demand from two distant regions using local characteristics, there may be a high correlation between the demand for taxis in the two regions. As part of our comparison of CNN and LSTM neural networks, we only use LSTM neural networks for taxi demand prediction but conduct experiments using both approaches as well. Thus, we use the stacked denoising autoencoders for traffic data imputing as a method for generalizing the model [29].

3.1.3 Proposed Single Demand Prediction Model

Here is an example of a single demand prediction model for taxi pick-up demand, which is based on LSTMs for modelling the temporal relationship between pick-up demand and different time intervals, as well as a denoising layer for handling random noise in input data. Taxi drop-off demand can also be predicted similarly. This section explains the model for predicting demand for taxis of a single type in detail. A building's taxi demand at time t ($1 \leq i \leq m$ is represented by d_t^i , while its information demand is represented by $1 \leq t \leq n$). As a sequence, $D_n^i = d_1^i, d_2^i, d_3^i, \dots, d_n^i$ is the historical taxi demand for building block i . Note that D_n^i represents the taxi demand from the start of the sequence to the end of the sequence. $D_n^i = 2122$, for example, indicates that there were two pickups at time step 1, one at time step 2, and so forth for block i .

As taxi demand can be highly variable, it can be difficult to predict exactly what d_t^i do. The taxi demand is rounded using a factor q for simplicity. Taxi demands are grouped as one demand for every q taxis. If $q=10$, then 620-629 in the sequence becomes 620; e.g., if $q=10$, 620 becomes 620 in the sequence. The prediction errors were kept low by setting $q = 10$ based on the first few tests [18].

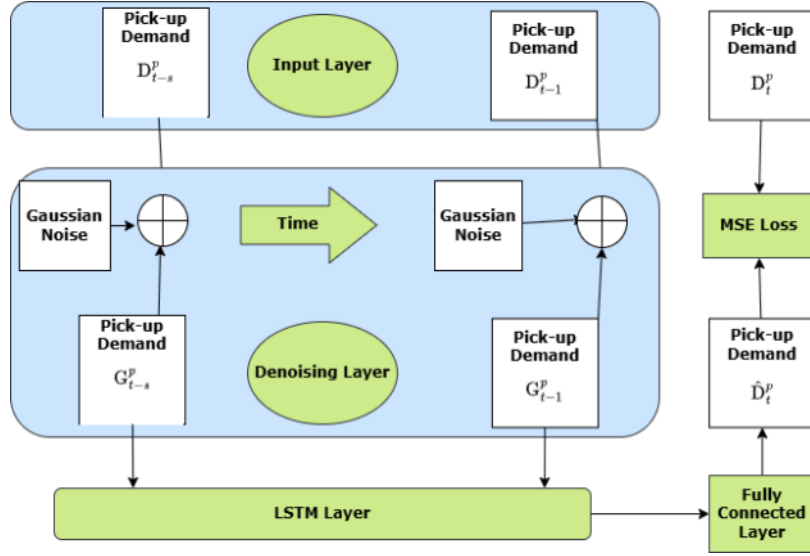


Figure 2: The single demand prediction model structure.

3.2 Component for Matching Taxis to Regions

In each region, vacated taxis are allocated based on future demand estimates and current taxi locations. The first metric is used to ensure that taxi assignments are proportional to demand. Due to this, more taxis will be dispatched to the region where a high demand is expected, and vice versa. To minimize the transition period, it is important to consider the second metric (current taxi locations) to ensure that vacant taxis can reach their designated regions as quickly as possible. Customers searching for rides will have less waiting time during the early part of the period. The demand for taxis in all regions (10, 20, 40, 10) is the same, so in Region A we should have one taxi, in Region B two taxis, in region C four taxis, and in region D one taxi is required.

Figure 3 illustrates an efficient method of matching taxis to regions by using a bipartite weighted graph. Weights representing graph edges are computed using the taxi's shortest distance from the nearest border. As shown in our graph, each of the regions in Manhattan had a different level of demand over that period, so we duplicated the regions with a higher demand multiple times from their normalized levels of demand. In this regard, the taxi-to-region matching component aims to maintain perfect matching while minimizing the sum of weights. The taxis are assigned to one region each. There are usually more taxis than regions in practice. Taxi IDs are identified by index i and index j , and the following

duplicates identify region IDs. It follows that $j=2$ corresponds to the same region as $j=3$.

Here is an ILP problem that can be used to optimize the matching procedure aforementioned:

$$(P): \text{minimize}_{x_{i,j} \in \{0,1\}} \sum_i \sum_j w_{i,j} x_{i,j}, \quad (1)$$

$$\text{Subject to:} \\ \sum_j x_{i,j} = 1, \forall_i, \text{ and } \sum_i x_{i,j} = 1, \forall_j, \quad (2)$$

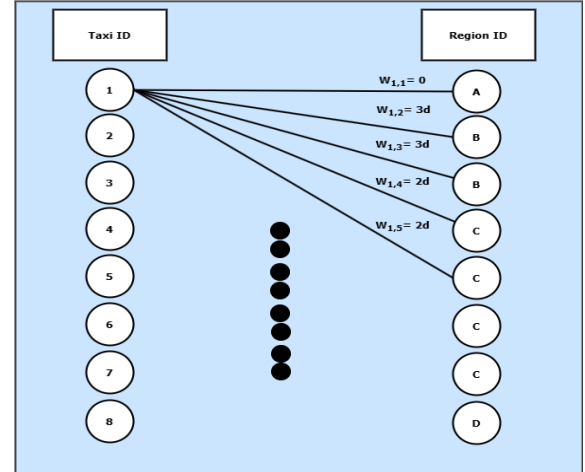


Figure 3: Common matching graph illustration.

Taxi i is either assigned to region j or not, according to $w_{i,j}$, whereas $x_{i,j}$ indicates whether taxi i is associated with region j

The value is 1 in this case. In (P), the taxi is restricted to one region only, ensuring perfect matching.

In addition to the Hungarian algorithm, which is another heuristic minimum-weight perfect matching algorithm, there are several other heuristic minimum-weight matching algorithms as well. By using the polynomial-time $\approx \mathcal{O}(N)$ approach, the problem can be solved much faster than by using a branch-and-bound algorithm based on ILP.

3.3 Taxi Routing Optimizer

Road networks can be constructed using simple search techniques and complex fuzzy logic theories [30]. Traffic networks are analyzed by transforming them into complex intersectional and road graphs. Each road r contains several segments of the same length lr , while the intersections of the roads $r \in \{1, \dots, R\}$ are at most two. It then consists of vertices and edges, the vertices being the connections between segments. Taxi i is currently located at S_i, S_g and has a destination of D_i, D_g with D_i and S_i representing the street IDs and S_{gi} and D_{gi} representing segments of the streets. In this article, a real-time traffic model is proposed to optimize route planning. Integer linear programs are developed to determine which is the fastest route based on the location and region of vacant taxis at present. A real-time feedback system determines the fastest routes using traffic feedback. Since ILP is solved on a regular basis based on new data received, the best route is determined based on the most recent data. A real-time feedback system determines the fastest routes using traffic feedback [31]. Our approach might update routes every minute.

3.4 Taxi Demand Service with Blockchain Framework

In this section, we describe the blockchain network. Taxi services and passengers benefit from the blockchain framework because it certifies transparency and security. On the basis of IoT devices, security processes track and trace. Using smart devices, ride-sharing entities can track transmissions to provide security. IoT devices and registered vehicles are gathered in the proposed system. Blockchain allows all activities to be tracked by storing data from IoT sensors and vehicles in a database. As a result of blockchain technology, users are prevented from hacking other users' accounts and stealing their points to use for payments. Blockchain technology protects transactions, driver, and

passenger profiles, so non-members or users without accessibility IDs cannot access them.

4 RESULTS AND DISCUSSION

According to Figure 4, Hyperledger Fabric is able to process about a million transactions per second. To evaluate the proposed system's performance, three primary user groups were taken into consideration: 500, 1000, and 1500 individuals. A statistical analysis was performed to assess the performance of the taxi demand service based on blockchain technology.

An illustration of the latency of transactions for a blockchain-based taxi demand service is shown in Figure 5. A performance assessment of the designed system was conducted using three defined user groups. A platform's response time measures the system latency. User groups are used to categorize transactions. The Figure shows that latency increases with increasing user numbers.

The comparison between the proposed system and other state-of-the-art methods is presented in Figure 6. This approach achieves the lowest root mean square error. The Figure shows that MLP exhibits the highest RMSE. Spatiotemporal correlations are not taken into consideration by the MLP. However, XGBoost is similar to MAP because it ignores spatial correlation.

Throughout the week, from Monday to Sunday, Figure 7 shows how taxi pickup and drop-off demand varies daily. Record pick-ups are indicated in blue, while drop-offs are indicated in green. According to the Figure, Friday is the busiest day for pickups and drop-offs, with approximately 25,000 pick-ups and 25,500 drops. A maximum of 13,000 pick-ups and drop-offs are recorded on Sunday, the lowest day of the week.

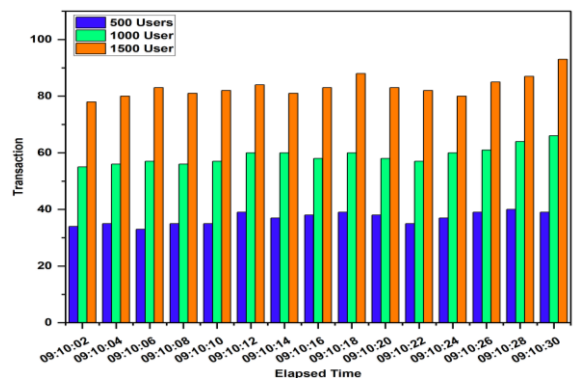


Figure 4: The number of transactions versus elapsed time.

Figures 8 and 9 illustrate the method's excellent performance. As compared to passengers, drivers lose about 14 percent of their utility. Passengers save about 72% of their waiting time, which shows the greatest reduction in waiting time.

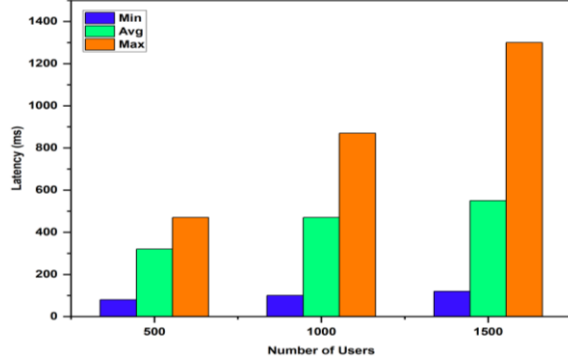


Figure 5: The performance of latency of query transaction versus number of users.

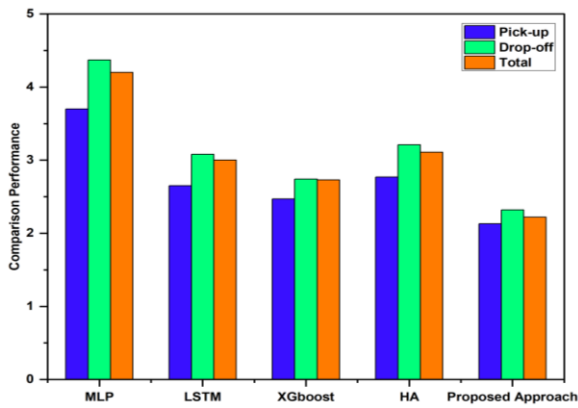


Figure 6: Comparison of various existing model performance.

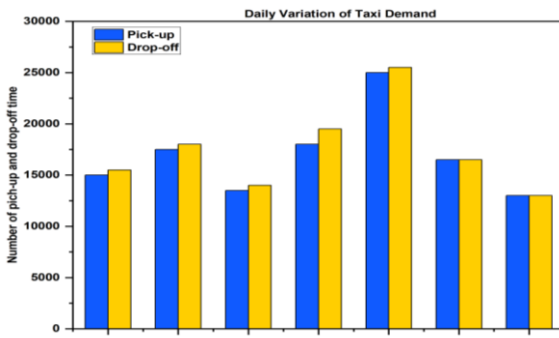


Figure 7: Daily variation of taxi demand.

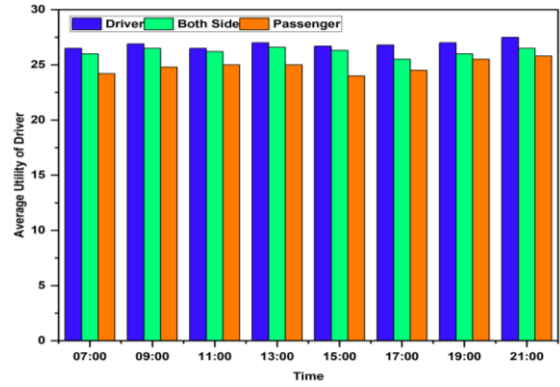


Figure 8: Analyzing the utility of drivers versus time.

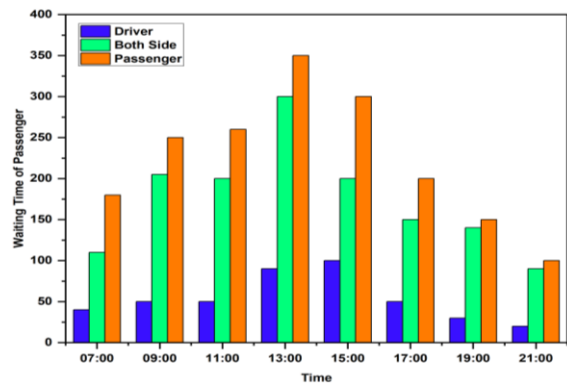


Figure 9: Saving time for the passenger versus time

5 CONCLUSIONS

Blockchain technology combined with advanced data analytics offers a robust framework for tackling critical issues related to vehicular safety management and optimizing taxi demand services. By leveraging this integrated approach, the proposed system effectively ensures the protection of sensitive information, maintains data integrity, and allows for swift real-time decision-making. Consequently, allocation of resources dedicated to vehicle safety - such as emergency response measures, route optimization, and traffic management - can be handled in a more coordinated and efficient manner. Moreover, the transparency offered by blockchain facilitates trustworthy and verifiable transactions among passengers, drivers, and service providers, contributing to enhanced user trust and satisfaction. Despite demonstrating superior accuracy and operational efficiency compared to traditional techniques, the proposed approach still encounters certain practical barriers. These barriers include managing increased data volumes as networks

expand, ensuring compatibility between various technology platforms, and adhering to evolving regulatory frameworks. Therefore, future research directions should emphasize overcoming these hurdles, examining scalability solutions capable of supporting city-wide or regional networks, and exploring innovative methodologies that go beyond existing optimization approaches to further enhance both passenger convenience and overall vehicle safety outcomes.

REFERENCES

- [1] K. Moghaddasi, S. Rajabi, and F. S. Gharehchopogh, "Multi-Objective Secure Task Offloading Strategy for Blockchain-Enabled IoV-MEC Systems: A Double Deep Q-Network Approach," *IEEE Access*, vol. 12, pp. 3437-3463, 2024, [Online]. Available: <https://doi.org/10.1109/ACCESS.2023.3348513>.
- [2] D. Das, S. Banerjee, P. Chatterjee, U. Ghosh, and U. Biswas, "Blockchain for Intelligent Transportation Systems: Applications, Challenges, and Opportunities," *IEEE Internet Things J.*, vol. 10, no. 21, pp. 18961-18970, Nov. 2023, [Online]. Available: <https://doi.org/10.1109/JIOT.2023.3277923>.
- [3] J. Kang, Z. Xiong, D. Niyato, D. Ye, D. I. Kim, and J. Zhao, "Blockchain for Secure and Efficient Data Sharing in Vehicular Edge Computing and Networks," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4660-4670, Jun. 2019, [Online]. Available: <https://doi.org/10.1109/JIOT.2018.2875542>.
- [4] R. Sharma and S. Chakraborty, "B2VDM: Blockchain Based Vehicular Data Management," in 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore: IEEE, Sep. 2018, pp. 2337-2343, [Online]. Available: <https://doi.org/10.1109/ICACCI.2018.8554369>.
- [5] E. Ó Broin and C. Guivarch, "Transport infrastructure costs in low-carbon pathways," *Transp. Res. Part D Transp. Environ.*, vol. 55, pp. 389-403, Aug. 2017, [Online]. Available: <https://doi.org/10.1016/j.trd.2016.11.002>.
- [6] O. Lah, "Sustainable development benefits of low-carbon transport measures," *Ger. Dtsch. Ges. Für Int. Zusammenarbeit GIZ GmbH*, 2015.
- [7] P. Rani and R. Sharma, "Intelligent Transportation System Performance Analysis of Indoor and Outdoor Internet of Vehicle (IoV) Applications Towards 5G," *Tsinghua Sci. Technol.*, vol. 29, no. 6, pp. 1785-1795, Dec. 2024, [Online]. Available: <https://doi.org/10.26599/TST.2023.9010119>.
- [8] N. Khoshavi, G. Tristani, and A. Sargolzaei, "Blockchain Applications to Improve Operation and Security of Transportation Systems: A Survey," *Electronics*, vol. 10, no. 5, p. 629, Mar. 2021, [Online]. Available: <https://doi.org/10.3390/electronics10050629>.
- [9] N. K. Agrawal, S. Kumar, S. P. Yadav, M. H. Falaah, P. Rani, and A. Singh, "TFL-IHOA: Three-Layer Federated Learning-Based Intelligent Hybrid Optimization Algorithm for Internet of Vehicle," *IEEE Trans. Consum. Electron.*, vol. 70, no. 3, pp. 5818-5828, Aug. 2024, [Online]. Available: <https://doi.org/10.1109/TCE.2023.3344129>.
- [10] A. Singh, P. Rani, M. H. Falaah, and S. P. Yadav, "Blockchain-Based Lightweight Authentication Protocol for Next-Generation Trustworthy Internet of Vehicles Communication," *IEEE Trans. Consum. Electron.*, vol. 70, no. 2, pp. 4898-4907, May 2024, [Online]. Available: <https://doi.org/10.1109/TCE.2024.3351221>.
- [11] A. S. Khan, K. Balan, Y. Javed, S. Tarmizi, and J. Abdullah, "Secure Trust-Based Blockchain Architecture to Prevent Attacks in VANET," *Sensors*, vol. 19, no. 22, p. 4954, Nov. 2019, [Online]. Available: <https://doi.org/10.3390/s19224954>.
- [12] J. Kang, Z. Xiong, D. Niyato, D. Ye, D. I. Kim, and J. Zhao, "Toward Secure Blockchain-Enabled Internet of Vehicles: Optimizing Consensus Management Using Reputation and Contract Theory," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2906-2920, Mar. 2019, [Online]. Available: <https://doi.org/10.1109/TVT.2019.2894944>.
- [13] N. Hussain and P. Rani, "Comparative studied based on attack resilient and efficient protocol with intrusion detection system based on deep neural network for vehicular system security," in *Distributed Artificial Intelligence*, CRC Press, 2020, pp. 217-236.
- [14] A. O. Philip and R. K. Saravanaguru, "Secure Incident & Evidence Management Framework (SIEMF) for Internet of Vehicles using Deep Learning and Blockchain," *Open Comput. Sci.*, vol. 10, no. 1, pp. 408-421, Nov. 2020, [Online]. Available: <https://doi.org/10.1515/comp-2019-0022>.
- [15] A. Alharthi, Q. Ni, R. Jiang, and M. A. Khan, "A Computational Model for Reputation and Ensemble-Based Learning Model for Prediction of Trustworthiness in Vehicular Ad Hoc Network," *IEEE Internet Things J.*, vol. 10, no. 20, pp. 18248-18258, Oct. 2023, [Online]. Available: <https://doi.org/10.1109/JIOT.2023.3279950>.
- [16] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," *Comput. Electr. Eng.*, vol. 105, p. 108543, 2023.
- [17] P. Rani, U. C. Garjola, and H. Abbas, "A Predictive IoT and Cloud Framework for Smart Healthcare Monitoring Using Integrated Deep Learning Model," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 53-65, 2024.
- [18] N. Junaidi, M. P. Abdullah, B. Alharbi, and M. Shaaban, "Blockchain-based management of demand response in electric energy grids: A systematic review," *Energy Rep.*, vol. 9, pp. 5075-5100, Dec. 2023, [Online]. Available: <https://doi.org/10.1016/j.egy.2023.04.020>.
- [19] E. Kapassa and M. Themistocleous, "Blockchain Technology Applied in IoV Demand Response Management: A Systematic Literature Review," *Future Internet*, vol. 14, no. 5, p. 136, Apr. 2022, [Online]. Available: <https://doi.org/10.3390/fi14050136>.

- [20] A. Singh, P. Rani, M. H. Falaah, and S. P. Yadav, "Smart Traffic Monitoring Through Real-Time Moving Vehicle Detection Using Deep Learning via Aerial Images for Consumer Application," *IEEE Trans. Consum. Electron.*, vol. 70, no. 4, pp. 7302-7309, Nov. 2024, [Online]. Available: <https://doi.org/10.1109/TCE.2024.3445728>.
- [21] P. Rani and M. H. Falaah, "Real-Time Congestion Control and Load Optimization in Cloud-MANETs Using Predictive Algorithms," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 66-76, 2024.
- [22] B. Ghimire and D. B. Rawat, "Secure, Privacy Preserving, and Verifiable Federating Learning Using Blockchain for Internet of Vehicles," *IEEE Consum. Electron. Mag.*, vol. 11, no. 6, pp. 67-74, Nov. 2022, [Online]. Available: <https://doi.org/10.1109/MCE.2021.3097705>.
- [23] L.-Y. Yeh, N.-X. Shen, and R.-H. Hwang, "Blockchain-Based Privacy-Preserving and Sustainable Data Query Service Over 5G-VANETs," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15909-15921, Sep. 2022, [Online]. Available: <https://doi.org/10.1109/TITS.2022.3146322>.
- [24] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," Jan. 30, 2017, arXiv, [Online]. Available: <https://doi.org/10.48550/arXiv.1412.6980>.
- [25] J. Ke, H. Zheng, H. Yang, and X. M. Chen, "Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach," *Transp. Res. Part C Emerg. Technol.*, vol. 85, pp. 591-608, 2017.
- [26] K. F. Chu, A. Y. Lam, and V. O. Li, "Travel demand prediction using deep multi-scale convolutional LSTM network," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2018, pp. 1402-1407, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8569427/>.
- [27] H. Yao, F. Dai, X. Zhang, S. Zhang, and C. Li, "Deep multi-view spatial-temporal network for taxi demand prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/11836>.
- [28] W. Jiang and L. Zhang, "Geospatial data to images: A deep-learning framework for traffic forecasting," *Tsinghua Sci. Technol.*, vol. 24, no. 1, pp. 52-64, 2018.
- [29] Y. Duan, Y. Lv, Y.-L. Liu, and F.-Y. Wang, "An efficient realization of deep learning for traffic data imputation," *Transp. Res. Part C Emerg. Technol.*, vol. 72, pp. 168-181, 2016.
- [30] M. A. Quddus, W. Y. Ochieng, and R. B. Noland, "Current map-matching algorithms for transport applications: State-of-the art and future research directions," *Transp. Res. Part C Emerg. Technol.*, vol. 15, no. 5, pp. 312-328, Oct. 2007, [Online]. Available: <https://doi.org/10.1016/j.trc.2007.05.002>.
- [31] X. Wan, H. Ghazzai, and Y. Massoud, "Mobile Crowdsourcing for Intelligent Transportation Systems: Real-Time Navigation in Urban Areas," *IEEE Access*, vol. 7, pp. 136995-137009, 2019, [Online]. Available: <https://doi.org/10.1109/ACCESS.2019.2942282>.

Secure and Privacy-Enhanced Data Aggregation for IoT Using Fog Computing

Andi Bahrūn¹, La Oge², Ahmed Zeyad Almhanna³ and Mohammed Abdul Jaleel Maktoof⁴

¹Department of Agrotechnology, Universitas Halu Oleo, 93132 Kendari, Indonesia

²Department of Agricultural Product Technology, Universitas Sulawesi Tenggara, 93121 Kendari, Indonesia

³Dijlah University College, 10021 Baghdad, Iraq

⁴Department of Computer Technology Engineering, Al-Turath University Al Mansour, 10013 Baghdad, Iraq
andibahrūn7@gmail.com, laoge@un-sultra.ac.id, ahmed.ziad@duc.edu.iq, mohammed.jaleel@uoturath.edu.iq

Keywords: Privacy-Preserving Data Aggregation, Fog Computing, IoT Security, Homomorphic Encryption, Subset Aggregation.

Abstract: As the Internet of Things (IoT) expands rapidly, securing sensitive data has become increasingly complex, especially in environments with limited resources and strict latency requirements. By facilitating data processing at the network edge, fog computing minimizes delays and reduces bandwidth consumption. In this paper, we present APPA, which is a method for anonymizing and aggregating Fog-enabled IoT data in a privacy-preserving manner. The APPA provides IoT devices with the ability to choose between privacy-preserving and standard encryption methods based on their privacy preferences and eliminates the need to rely on trusted third parties. As a result of utilizing the Paillier and Bilinear ElGamal's homomorphic encryption algorithms, the scheme is able to aggregate secure subsets, enable fault-tolerant decryption, and allow dynamic device enrolment and revocation. According to performance evaluations, APPA is significantly more efficient and has a lower communication overhead than existing solutions, thus making it a good choice for scalable and privacy-conscious IoT applications.

1 INTRODUCTION

As the Internet of Things (IoT) expanded rapidly, it enabled devices to collect and share vast amounts of data across various industries. Real-time information has been particularly useful in sectors like healthcare, smart cities, and industrial automation, where real-time data facilitates efficient decision-making. IoT devices pose many challenges with respect to privacy, security, and data management due to their enormous scale and diversity [1]. Sensors that gather sensitive information on the Internet of Things will need robust mechanisms to protect this information as it grows in importance. In resource-constrained environments, traditional cloud-based models face latency and bandwidth issues, especially when it comes to data storage and computation. Fog computing extends cloud capabilities to the edge of a network to address this challenge. As a result, fog computing can make IoT more efficient, reducing latency and bandwidth bottlenecks by processing data close to the source. While fog computing can significantly enhance performance, it can also introduce new security and

privacy concerns, particularly with regard to the aggregation of sensitive IoT data [2].

When data from distributed IoT devices is aggregated, especially in smart cities or industrial networks, sensitive information about users, behaviours, or environments is revealed. To achieve security and privacy without compromising the ability to analyze and make decisions, and explore methods of aggregating data in a secure and privacy-preserving manner. By incorporating Encryption, anonymization, and advanced cryptographic protocols in conjunction with fog computing [3]. Our daily lives are undergoing a revolution as mobile communication and the Internet of Things grow exponentially [4], [5]. A smart meter, wearable device, or vehicle that interacts with its surroundings can sense physical objects nearby to produce data, as shown in Figure 1. In cloud computing, sensory data can be analyzed, and decision-making can be facilitated, or corresponding services can be provided for users as a result of the analysis [6]. It is essential to have real-time data on electricity consumption so that one can make predictions about power dispatch,

determine dynamic pricing, and optimize electricity [7].

When users publish their data, it puts much of their data at risk of security and privacy issues [8], [9]. It is possible for anyone with malicious intentions to intercept sensory data published over open channels [10]. Recently, IoT sensor data has been subject to a significant number of data breaches, exposing its vulnerability to security and privacy threats. Basically, data privacy and data security are different, so they require different preventive measures [11]. Data security refers to data that can only be accessed by authorized users while unauthorized users are unable to obtain it. Therefore, data security is crucial in communications [12]. Data confidentiality, authentication, and integrity in communications are currently protected by a variety of mature technologies, including authentication protocols and key management protocols [13]. Despite this, authorized users are still launching many security threats, which need to be addressed as soon as possible. The security communication schemes, on the other hand, hide sensory data from unauthorized users, contrary to the purpose for which the data were released originally.

The concept of data privacy governs it, and some attributes must be kept in order for the public to use them. However, other attributes must be removed so that sensitive information about the user cannot be revealed. Information about users can be sensitive depending on how the data owner feels about it. There are different levels of concern about the privacy of personal information among individuals, according to research [14]. A person who has diverse sources of income, for instance, may not wish for others to know who they are. People who earn low wages do not pay as much attention to their income information, particularly those who receive social assistance, whose income information is basically public. Travel trajectory is something people want to keep private, but some people like to broadcast their travel to the world [15]. Furthermore, data publishing can be viewed by users as a profitable activity. Depending on the user's convenience, raw data may be published for better exchange or benefit. If the published data resembles the original state, it is more valuable, provided it is in accordance with the owner's wishes. Calculation costs are directly related to the amount of data privacy processed. It is, therefore, possible to implement hierarchical privacy strategies when processing published data. It has been proposed that many technologies can be used to ensure privacy-preserving data publishing in the Internet of Things. The K-anonymity model is one of the most popular in

data publishing because it protects individual privacy, while the L-diversity model is another popular one [16], [17]. In order to achieve privacy, data is generalized and concealed before being published [18].

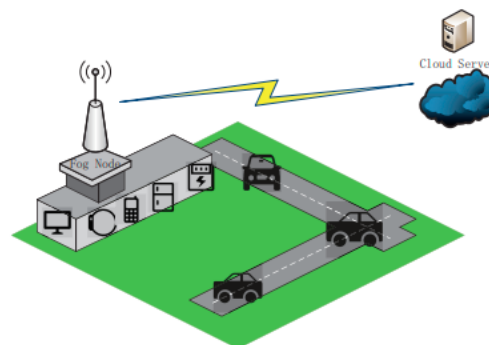


Figure 1: The architecture of an IoT network enhanced by Fog.

However, there are still a few issues that need to be resolved despite previous data aggregation solutions' ability to protect sensitive information. The aggregation of data was previously prohibited unless users or smart devices selected a method that protected their privacy. Furthermore, security parameters for smart devices are generated by a trusted third party (TTP). Therefore, the adversarial assumption fails since there are hardly any real-world TTPs [19]. Thirdly, if smart devices malfunction, data aggregation will be affected. Furthermore, it should support both new smart device enrollments and old device revocations.

The purpose of this paper is to propose an option for aggregating privacy data based on BonehGoh-Nissim (BGN) without using any TTP. An encryption method without privacy encryption and an encryption method with privacy encryption are used. Based on the privacy requirements of smart devices, the cloud center will calculate the aggregated data of all smart devices, as well as the individual data of those devices that haven't selected privacy encryption. A smart device can upload sensed data using two encryption methods. The proposed scheme provides fault-tolerant decryption as well as flexible enrollment and revocation of devices.

2 LITERATURE REVIEW

Cloud services are extended to the edge of networks via fog computing, an important paradigm in the Internet of Things (IoT). This extension allows IoT

applications to process real-time data with low latency [20]. Fog computing brings computational resources closer to the data source, making it easier and faster for IoT systems to process and act on data. When fog nodes are deployed at the network edge, data can be pre-processed, filtered, and forwarded immediately. In contrast to traditional IoT architectures, which heavily rely on centralized cloud resources, these mechanisms are more efficient and secure [21]. In addition to reducing cloud infrastructure burden, a localized approach protects data handling environments, since sensitive information can be processed and stored close to its source, reducing the risk of exposure during remote transmission. Furthermore, fog computing overcomes some of the inherent limitations of cloud computing in IoT environments, specifically scalability, area awareness, and latency [22]. In contrast to cloud computing, which is often unable to handle huge amounts of data originating from geographically dispersed IoT devices in real time, fog computing provides a distributed and scalable solution for IoT deployments that can quickly adapt to changing needs, ensuring seamless and responsive operation in a variety of scenarios.

2.1 Schemes for Aggregating Data that Preserve Privacy

Data aggregation schemes based on the Paillier cryptosystem are often employed to maintain privacy in fog-enabled smart grids [23]. Smart grids collect a lot of data about energy consumption, grid stability, and equipment performance. This data can be aggregated using a modified Paillier cryptosystem without compromising individual consumer privacy. In order to make fog computing suitable for resource-constrained environments, modifications are often made to the encryption and decryption processes in order to improve their efficiency and security. As part of IoT deployments, Paillier homomorphic encryption is used to encrypt attribute vectors and end-device data, enabling secure subset aggregation. The process of aggregating subsets of data on the basis of specific criteria or attributes entails computing aggregate statistics upon the subset. When designing smart cities, it may be necessary to calculate the average traffic flow on particular streets at peak times. At peak times, traffic flow on particular streets may need to be calculated when designing smart cities.

2.2 Secure Data Transmission in IoT and Fog Networks

Data security, integrity, and availability must be ensured in IoT and fog networks by implementing secure routing protocols. While protecting data from eavesdropping, tampering, and denial-of-service attacks, protocols optimize latency, bandwidth, and energy consumption on networks. This technique uses clusters to avoid unnecessary data transmission, enhance security, and avoid unnecessary transmissions in Wireless Sensor Networks (WSNs) [24]. In comparison to fish schools or bird flocks, CPSO mimics the social behaviour of these groups. When it comes to secure routing, CPSO is a valuable tool for determining optimal cluster heads and routes, taking factors like node energy, link quality, and security into account. With CPSO, WSNs can be improved significantly in terms of performance and security by forming clusters and selecting the most secure and efficient routes.

An aggregator node ensures secure communication by transferring secrets between Fog servers and transmitting them through it, adding an extra layer of protection. Through this approach, sensitive information is protected from unauthorized access by ensuring that only authorized nodes have access to decrypt and process it. Secrets can be cryptographic keys, digital certificates, or any other information necessary to authenticate and authorize communication between Fog and aggregators. The network traffic is effectively controlled and enhanced by a traffic-aware coordinated framework for IoT-based smart agriculture monitoring [25]. To optimize routing decisions and minimize the risk of security breaches, this framework accounts for the specific characteristics of the agricultural environment, such as sensor density, mobility, and obstacles. Cooperative techniques, such as data aggregation and fusion, are also incorporated into the framework, reducing data transmission and energy consumption.

2.3 Data Aggregation in Smart Grids

Intelligent grids are increasingly leveraging fog computing to aggregate large amounts of data from smart meters and other grid devices, providing a distributed and efficient approach to managing huge amounts of data. A smart grid overcomes traditional cloud-based limitations, such as high latency, bandwidth constraints, and privacy concerns, by

placing fog nodes closer to the edge of the network. Smart grids can reduce computation and communication costs as well as enhance decision-making in real-time by making use of fog computing to overcome traditional data aggregation limitations. Traditionally, data is aggregated on a centralized cloud server, which is time-consuming and expensive to communicate with. Data can be transferred to the cloud more easily and processed and consolidated locally, which facilitates faster response times.

3 PROPOSED METHODOLOGY

A new privacy-preserving problem is addressed in this paper: how to aggregate subsets in fog-enhanced IoT while maintaining privacy. Consider an IoT device set $I = \{I_1, I_2, \dots, I_N\}$, two fog nodes N_1 and FN_2 , and a query user. It prepares sensed data DI periodically since IoT devices $I_i \in I$ have normalized attributes $V_i = (v_{i1}, v_{i2}, \dots, v_{ik})$ that indicate their status. A query user holds several queries (defined as $U = (u_1, u_2, \dots, u_k)$ vectors with threshold values) and intends to query a subset $I' \subseteq I$ to compute their sum aggregation, where each II attribute vector V_i Matches the query, i.e., the inner product $U \cdot V_i \geq \tau$. Many realistic IoT applications can be used with this problem; however, there are also some privacy issues to consider [26]. APPA is an algorithm for aggregating anonymous, private, fog-enhanced data that focuses on addressing privacy concerns in IoT.

A privacy-preserving approach to data aggregation in cloud computing-enhanced IoT. As part of the proposed APPA, the Paillier homomorphic cryptosystem [27] has been utilized to support the aggregation process. We believe that APPA is the first solution to fog-enhanced IoT subset aggregation that preserves privacy.

In the fog layer, users' queries, IoT device responses, and intermediate results are protected in different ways.

Analyzing and evaluating the communication overhead and computation cost of APPA between query submission and response recovery.

3.1 Design Goal

A fog-enhanced IoT anonymity-preserving data aggregation algorithm (APPA) is presented here based on the aforementioned system and security models. These objectives can be achieved by completing the following tasks:

- 1) A privacy-preserving scheme should be proposed. In the proposed APPA, the normalized vector and threshold values are kept

private, which ensures the privacy of the user's query. When the subset aggregation is computed, the user cannot obtain the participating devices. The D_i and attribute vector D_i of the IoT device I_i No one can access it. Fog nodes FN_1 or FN_2 failed to determine anything about the user's query, the data from the IoT device, or the subset I' . As a security measure, the user should choose the final result during query submission and its encrypted equivalent uploaded during query submission.

- 2) The proposed scheme must be efficient. The purpose of our fog nodes is to reduce communication costs between IoT devices, fog nodes, and users. APPA, therefore, aims to provide the best communication efficiency between fog nodes FN_1 and the user, and the best communication efficiency between IoT devices and fog nodes FN_2 , where $k = |V_i|$ represents the attribute vector VI and DI represents the prepared data D_i for each IoT device I_i .

3.2 Proposed Anonymous Privacy-Preserving Data Aggregation (APPA) Scheme

3.2.1 Overview and System Architecture

A brief overview of the APPA scheme is presented in the first step, which incorporates several basic preliminaries. The APPA provides formal definitions for various functions as well. Last but not least, the proposed APPA scheme includes four steps: initializing the system, generating health data, aggregating privacy-preserving data, decrypting it, and verifying it. A formalized definition of a function is provided, along with proof of its correctness.

3.2.2 Cryptographic Preliminaries and Notation

Throughout the document, we will use the following basic notations to summarize our proposal. Pairing-based cryptography pairs $G1 = G2$ with the symmetric pairing, which is a special form of a bilinear map. Due to the symmetric nature of bilinear pairings, we refer to $G1 = G2$ throughout our paper by G .

The CSetup function is used to generate the symmetric bilinear pairing group. Several key pairs will be generated using the proposed scheme for encrypting (Encryption) and signing (signing).

3.2.3 Formalized APPA Function Definitions

APPA's formalized function definitions follow:

- CSetup(1^λ). $gk = (n, g_1, g_2, G, G_T, e, H)$ is a symmetric bilinear group that outputs a one-way hash function, $H : \{0, 1^* \rightarrow G$, when λ is entered as a security parameter.
- CKeygeneration(gk). Each sensor $si \in S$ node's random variables $xi, ui \in \mathbb{Z}_p$ with $i = 1, 2, 3 \dots, k$, are used to generate $(E_{pk}, E_{sk}) = (e(g_1, g_2), g_1^u)$ and $(S_{pk}, S_{sk}) = (g_2^x, x)$, for Encryption and signature generation.
- Encryption(m_i, E_{pki}). As described previously, the cypher text of a message $m_i \in G_T$ is encoded using a random RIZP, where the cypher text $CT_i = (C_1, C_2) = (g_2^{r_i}, E_{pki}^{r_i} \cdot m_i) = [g_2^{r_i}, e(g_1, g_2)^{u_i r_i} \cdot m_i]$ is computed by Encryption(m, E_{pki}).
- Sign(S_{ski}). CT_i is the portion of i^{th} ciphertext that is part. A $\sigma_i = [H(\mu_i) || TS]^{x_i}$ calculated under the secret key $S_{ski} = x_i$ will take into account the time stamp TS, making it more resistant to replay attacks.
- SigVerify(μ_i, σ_i). If the message, the signature, and the S_{pki} match, calculate $h = H(\mu_i)h = H(d_i)$ and receive $e(\sigma_i, g_2) = e(H(\mu_i), S_{pki}) = e(h_i, S_{pki})$ if it does;
- CipherProd (CT_i, k). Due to the homomorphism of the underlying encryption method $CT_i, i = 1, 2, 3, \dots, k$ outputs aggregated ciphertext by dividing all encrypted data by $CT = \prod_{i=1}^k CT_i$ upon input of T_i .
- AggSign(μ_i, σ_i). Assuming that each sensing node $s_i \in SN$ inputs signatures NI on each distinct message $\mu_i \in \{0, 1\}^*$, then outputs $\sigma = \prod_{i=1}^k \sigma_i$, where $\sigma \in G$ is the number of sensing nodes in the network.
- AggVerification(μ_i, σ). You should ensure that each message is distinct from the others. Calculate $H(\mu_i)$ for $1 < i < k = |SN|$ and receive if $e(\sigma, g_2) = \prod_{i=1}^k e(h_i, S_{pki})$.
- AggDecryption(E_{ski}, e). Decrypting the ciphertext, for $CT = (C_1, C_2) = \prod_{i=1}^k CT_i$, Compute Aggregate Decryption $E_{ski}(e) = C_2 / e(g_1^u, C_1) = \prod_{i=1}^k m_i$.

3.3 Proposed APPA Scheme

Remote health monitoring uses APPA for anonymous, privacy-preserving data aggregation, which mainly consists of four components. The initialization of the system, the generation of health

data, the aggregation of privacy-preserving data, the decryption, and the verification of the data.

3.3.1 System Initialization

Bootstrapping of the entire system is possible during the initialization of the remote medical server. As a first step, the medical server runs a common setup called CSetup(1^λ).

Using the CKeygeneration function, a public and private key pair is generated by the medical server using the Bilinear ElGamal cryptosystem. Detailed steps for generating public parameters are listed below:

Step 1. This function generates tuples $gk = (p, g_1, g_2, G, G_T, e, H)$ where G represents a finite cyclic group of prime order, G_T represents a bilinear G_1 map, G_2 represents a random generator within group G , and $H: \{0, 1\}^* \rightarrow G_1$ represents a one-way hash function.

Step 2. CKeygeneration(gk) is called by the medical server, and random variables $s_i \in SN$ with $i = 1, 2, 3, \dots, k$ are selected for each sensing node $s_i \in SN$. $(E_{pki}, E_{ski}) = (e(g_1, g_2), g_1^{u_i})$ (and $(S_{pki}, S_{ski}) = (g_2^{x_i}, x_i)$) are the public and private key pairs required for Encryption. For Encryption and signature generation (E_{pki}, E_{ski}) and (S_{pki}, S_{ski}) uses public and private key pairs to compute.

3.3.2 Privacy-Preserving Data Aggregation

As soon as the LCA (Local Control Authority) obtains the total number of encrypted and signed health data $CT_i || TS || \sigma_i$ for $i = 1, 2, 3, \dots, k$, it checks the signature. It payments the signature as follows:

Step 1: TS is checked first by the LCA, and SigVerify(μ_i, σ_i) is called next.

Step 2: In order to verify the signature validity, LCA computes hash $h = H(\mu_i)$ and checks if the public key S_{pki} is valid for signing

$$e(\sigma_i, g_2) = e(H(\mu_i), S_{pki}) = e(h_i, S_{pki}) \quad (1)$$

Step 3. Therefore, σ_i is accepted by LCA

Step 4. As remote health monitoring may involve multiple sensors for a patient, verification of each signature for that patient may take some time. As a result, the LCA checks multiple signatures together in batch mode to make the verification process more efficient by checking if:

$$e\left(\sum_{i=1}^k \sigma_i, g_2\right) = \prod_{i=1}^k e(h_i, S_{pki}) \quad (2)$$

Step 5. Lastly, LCA receives $\sum_{i=1}^k \sigma_i$,

3.4 Computation of Averages

The computation of averages is fundamental in data analysis. Two primary methods are employed to calculate central tendency measures:

- 1) Arithmetic Mean. $AM = \frac{1}{n} \sum_{i=1}^k a_i$, defines the arithmetic mean for a set of elements $a_1, a_2, a_3, \dots, a_n$;
- 2) Geometric Mean. For a set of elements $a_1, a_2, a_3, \dots, a_n$, the geometric mean can be determined by the formula $GM = (\prod_{i=1}^n a_i)^{1/n}$.

3.5 Security Analysis

Our analysis of the correctness and security of the proposed APPA scheme relies on the security theorems. By analyzing the security requirements discussed earlier, we will demonstrate that APPA assures confidentiality, authenticity, and end-to-end privacy for patient data.

4 RESULT ANALYSIS AND DISCUSSION

Comparing the proposed model (APPA) with ASAS, the APPA model has lower computational costs, as shown in Figure 2. A fog-based IoT environment, where hundreds of smart devices have been deployed, can benefit from even minor efficiency gains. In this regard, the APPA scheme has significant potential, particularly for limited-resource end devices.

As shown in Figure 3, the proposed scheme, ASC, and ASSA have different communication overheads, depending on how many smart devices are involved. A comparison of ASAS and the proposed approach is shown in Figure 4.

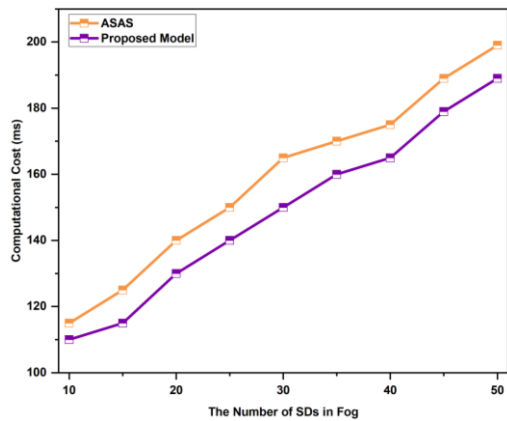


Figure 2: Computational cost comparison between APPA and ASAS schemes.

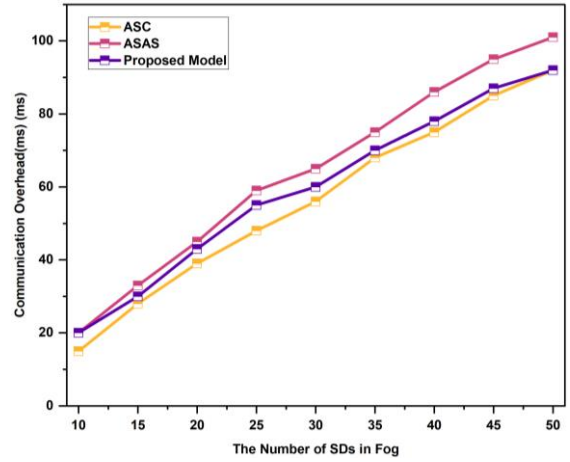


Figure 3: Communication overhead versus number of smart devices for APPA, ASC, and ASSA.

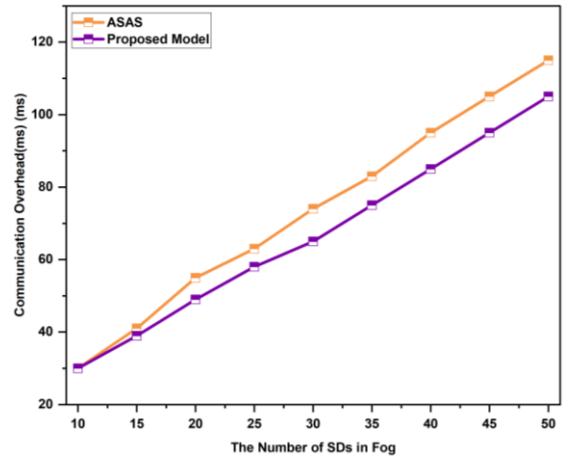


Figure 4: Communication overhead comparison of ASAS and proposed APPA approach.

5 CONCLUSIONS

IoT foggy environments can take advantage of APPA, which aggregates anonymous data while protecting privacy. IoT devices can select from a variety of encryption options tailored to their individual privacy preferences, unlike traditional IoT solutions that typically rely on a centralized, trusted third party. This high-efficiency, scalable encryption scheme ensures robust privacy protection, integrity assurance, and confidentiality of sensitive user data across diverse network scenarios. The APPA system achieves fault-tolerant, secure data aggregation by intelligently combining bilinear ElGamal encryption with Paillier homomorphic encryption, enabling computation on encrypted data without

compromising privacy. According to our extensive evaluations, APPA demonstrates significantly lower computational costs and reduced communication overhead compared to alternative methods, such as ASAS, thereby offering better performance for resource-constrained IoT devices. This advanced technology is particularly well-suited for critical applications such as remote healthcare monitoring systems, smart city infrastructure, and environmental sensing networks, where robust privacy, reliability, efficiency, and rapid responsiveness are essential for ensuring secure and timely decision-making.

REFERENCES

- [1] A. Singh et al., "Blockchain-Based Lightweight Authentication Protocol for Next-Generation Trustworthy Internet of Vehicles Communication," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 2, pp. 4898-4907, May 2024, [Online]. Available: <https://doi.org/10.1109/TCE.2024.3351221>.
- [2] B. Bhola et al., "Quality-enabled decentralized dynamic IoT platform with scalable resources integration," *IET Communications*, 2022.
- [3] P. Rani, P. N. Singh, S. Verma, N. Ali, P. K. Shukla, and M. Alhassan, "An implementation of modified blowfish technique with honey bee behavior optimization for load balancing in cloud system environment," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1-14, 2022.
- [4] Y. A. Qadri, A. Nauman, Y. B. Zikria, A. V. Vasilakos, and S. W. Kim, "The Future of Healthcare Internet of Things: A Survey of Emerging Technologies," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 1121-1167, 2020, [Online]. Available: <https://doi.org/10.1109/COMST.2020.2973314>.
- [5] W. Wang, H. Xu, M. Alazab, T. R. Gadekallu, Z. Han, and C. Su, "Blockchain-Based Reliable and Efficient Certificateless Signature for IIoT Devices," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 7059-7067, Oct. 2022, [Online]. Available: <https://doi.org/10.1109/TII.2021.3084753>.
- [6] L. Zhang, M. Peng, W. Wang, Z. Jin, Y. Su, and H. Chen, "Secure and efficient data storage and sharing scheme for blockchain-based mobile-edge computing," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 10, p. e4315, Oct. 2021, [Online]. Available: <https://doi.org/10.1002/ett.4315>.
- [7] P. Rani et al., "Federated Learning-Based Misbehavior Detection for the 5G-Enabled Internet of Vehicles," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 2, pp. 4656-4664, May 2024, [Online]. Available: <https://doi.org/10.1109/TCE.2023.3328020>.
- [8] L. Zhou, C. Ge, S. Hu, and C. Su, "Energy-Efficient and Privacy-Preserving Data Aggregation Algorithm for Wireless Sensor Networks," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 3948-3957, May 2020, [Online]. Available: <https://doi.org/10.1109/JIOT.2019.2959094>.
- [9] C.-I. Fan, S.-Y. Huang, and Y.-L. Lai, "Privacy-Enhanced Data Aggregation Scheme Against Internal Attackers in Smart Grid," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 666-675, Feb. 2014, [Online]. Available: <https://doi.org/10.1109/TII.2013.2277938>.
- [10] N. Hussain and P. Rani, "Comparative studied based on attack resilient and efficient protocol with intrusion detection system based on deep neural network for vehicular system security," in *Distributed Artificial Intelligence*, CRC Press, 2020, pp. 217-236.
- [11] Y.-N. Liu, Y.-P. Wang, X.-F. Wang, Z. Xia, and J.-F. Xu, "Privacy-preserving raw data collection without a trusted authority for IoT," *Computer Networks*, vol. 148, pp. 340-348, Jan. 2019, [Online]. Available: <https://doi.org/10.1016/j.comnet.2018.11.028>.
- [12] N. Kumar Agrawal et al., "TFL-IHOA: Three-Layer Federated Learning-Based Intelligent Hybrid Optimization Algorithm for Internet of Vehicle," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 3, pp. 5818-5828, Aug. 2024, [Online]. Available: <https://doi.org/10.1109/TCE.2023.3344129>.
- [13] J. Song, Y. Liu, J. Shao, and C. Tang, "A Dynamic Membership Data Aggregation (DMDA) Protocol for Smart Grid," *IEEE Systems Journal*, vol. 14, no. 1, pp. 900-908, Mar. 2020, [Online]. Available: <https://doi.org/10.1109/JSYST.2019.2912415>.
- [14] A. Da Veiga and J. Ophoff, "Concern for Information Privacy: A Cross-Nation Study of the United Kingdom and South Africa," in *Human Aspects of Information Security and Assurance*, N. Clarke and S. Furnell, Eds., *IFIP Advances in Information and Communication Technology*, vol. 593, Cham: Springer International Publishing, 2020, pp. 16-29, [Online]. Available: https://doi.org/10.1007/978-3-030-57404-8_2.
- [15] P. Rani, S. P. Yadav, P. N. Singh, and M. Almusawi, "Real-World Case Studies: Transforming Mental Healthcare With Natural Language Processing," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Bijju, and G. C. Deka, Eds., IGI Global, 2025, pp. 303-324, [Online]. Available: <https://doi.org/10.4018/979-8-3693-4203-9.ch016>.
- [16] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557-570, Oct. 2002, [Online]. Available: <https://doi.org/10.1142/S0218488502001648>.
- [17] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul: IEEE, Apr. 2007, pp. 106-115, [Online]. Available: <https://doi.org/10.1109/ICDE.2007.367856>.
- [18] A. Singh et al., "Resilient Wireless Sensor Networks in Industrial Contexts via Energy-Efficient Optimization and Trust-Based Secure Routing," *Peer-to-Peer Networking and Applications*, vol. 18, no. 3, p. 132, Jun. 2025, [Online]. Available: <https://doi.org/10.1007/s12083-025-01946-5>.
- [19] P. Rani and M. H. Falaah, "Real-Time Congestion Control and Load Optimization in Cloud-MANETs Using Predictive Algorithms," *NJF Intelligent Engineering Journal*, vol. 1, no. 1, pp. 66-76, 2024.

- [20] R. Lu, K. Heung, A. H. Lashkari, and A. A. Ghorbani, "A Lightweight Privacy-Preserving Data Aggregation Scheme for Fog Computing-Enhanced IoT," *IEEE Access*, vol. 5, pp. 3302-3312, 2017, [Online]. Available: <https://doi.org/10.1109/ACCESS.2017.2677520>.
- [21] H. Mahdikhani, S. MahdaviFar, R. Lu, H. Zhu, and A. A. Ghorbani, "Achieving Privacy-Preserving Subset Aggregation in Fog-Enhanced IoT," *IEEE Access*, vol. 7, pp. 184438-184447, 2019, [Online]. Available: <https://doi.org/10.1109/ACCESS.2019.2961270>.
- [22] J. K. Singh and A. K. Goel, "Study on fog computing: security & privacy challenges in terms of IoT," *Journal of Physics: Conference Series*, vol. 2007, no. 1, p. 012039, Aug. 2021, [Online]. Available: <https://doi.org/10.1088/1742-6596/2007/1/012039>.
- [23] A. Saleem et al., "FESDA: Fog-Enabled Secure Data Aggregation in Smart Grid IoT Network," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6132-6142, Jul. 2020, [Online]. Available: <https://doi.org/10.1109/JIOT.2019.2957314>.
- [24] M. Manicka Raja and S. Manoj Kumar, "Aggregated PSO for Secure Data Transmission in WSN Using Fog Server," *Intelligent Automation & Soft Computing*, vol. 34, no. 2, pp. 1017-1032, 2022, [Online]. Available: <https://doi.org/10.32604/iasc.2022.025665>.
- [25] I. Abunadi, A. Rehman, K. Haseeb, L. Parra, and J. Lloret, "Traffic-Aware Secured Cooperative Framework for IoT-Based Smart Monitoring in Precision Agriculture," *Sensors*, vol. 22, no. 17, p. 6676, Sep. 2022, [Online]. Available: <https://doi.org/10.3390/s22176676>.
- [26] R. Lu, K. Alharbi, X. Lin, and C. Huang, "A Novel Privacy-Preserving Set Aggregation Scheme for Smart Grid Communications," in *2015 IEEE Global Communications Conference (GLOBECOM)*, IEEE, 2015, pp. 1-6, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/7417029/>.
- [27] P. Paillier, "Public-Key Cryptosystems Based on Composite Degree Residuosity Classes," in *Advances in Cryptology — EUROCRYPT '99*, J. Stern, Ed., *Lecture Notes in Computer Science*, vol. 1592, Berlin, Heidelberg: Springer, 1999, pp. 223-238, [Online]. Available: https://doi.org/10.1007/3-540-48910-X_16.

A Novel Secure and Decentralized Model for Authentication and Data Integrity in IoT-Based Cyber-Physical Healthcare System

Hijriani Hijriani¹, Muhamad Yusuf¹, Reem Abduljaleel Khaleel² and Banan Yousif Ahmed³

¹Department of Laws, Universitas Sulawesi Tenggara, 93121 Kendari, Indonesia

²Department of Radiology Technologies, Dijlah University College, 10021 Baghdad, Iraq

³Al-Iraqia University, Scientific Affairs Department, 10001 Baghdad, Iraq

hijriani@un-sultra.ac.id, myusuf@un-sultra.ac.id, reem.abduljaleel@duc.edu.iq, banan.y.ahmed@aliraqia.edu.iq

Keywords: Internet of Things (IoT), Cyber-Physical Healthcare Systems (CPHS), Authentication and Data Integrity, Blockchain Technology, Security and Privacy.

Abstract: The integration of Internet of Things (IoT) technologies in healthcare has revolutionized patient monitoring and clinical decision-making through real-time data acquisition. However, the proliferation of connected medical devices introduces significant vulnerabilities regarding the security and integrity of sensitive health information. This paper presents a novel hybrid framework for Cyber-Physical Healthcare Systems (CPHS) that addresses these critical challenges through an innovative combination of IoT and blockchain technologies. Our solution implements advanced cryptographic protocols to establish robust data authentication mechanisms while preserving patient privacy. The proposed decentralized architecture fundamentally transforms healthcare data management by eliminating single points of failure inherent in traditional centralized systems. Key advantages include enhanced system resilience, complete transaction transparency, and immutable record-keeping that prevents data tampering. Comprehensive simulation results demonstrate the framework's effectiveness in protecting sensitive patient information while simultaneously improving operational efficiency in IoT-enabled healthcare environments. The research contributes both theoretically and practically by providing a secure, scalable solution that maintains data integrity without compromising system performance - a crucial advancement for modern digital healthcare infrastructure.

1 INTRODUCTION

Health systems have been transformed by integrating IoT (Internet of Things) into patient monitoring, diagnostics, and overall care delivery systems. Patients' health can be monitored in real-time, their outcomes can be improved, and healthcare management can be improved through Cyber-Physical Healthcare Systems (CPHS) connected to the Internet of Things, featuring Internet of Things-enabled sensors and devices [1]. These systems are becoming increasingly interconnected and exchanging data, introducing significant security and integrity challenges. Patients and healthcare providers must ensure sensitive medical information is protected and the system's operations are trustworthy. The vulnerability of IoT-based CPHS to cyberattacks, data tampering, and unauthorized access is a key concern since traditional centralized models often experience single points of failure, scalability issues, and potential breaches. It is,

therefore, crucial that the data exchanged within the system is authenticated using robust mechanisms [2].

As wearables and gadgets become more embedded in open environments with radio communication infrastructure, new security threats arise. During communication, patient wearable devices are at risk of security risks as well as the collected and transmitted data. Therefore, the entire ecosystem must be protected from internal and external threats. [3]. As edge devices and gateways, these networks use patient wearable devices to collect and transmit data. The threat of a network adversary, capturing or injecting malicious data into these devices, is therefore readily available [4].

A decentralized authentication and data integrity model that uses blockchains and advanced cryptography to enhance resilience, transparency, and tamper resistance. By eliminating centralized control, decentralized health systems can transmit and authenticate sensitive health information securely without relying on a single authority. The proposed

model further enhances IoT-based healthcare systems in terms of scalability, efficiency, and trustworthiness. It enables greater confidence in IoT-enabled healthcare solutions by protecting patient privacy, securing medical data, and fostering a safe and effective use of cyber-physical systems in medicine through this innovative approach. Tractia, an intelligent organization, predicts by 2025 that the blockchain technology sector will generate 19.9 billion in annual earnings [3]. Healthcare is increasingly using IoTs to provide patients and physicians with real-time services [5]. In order to achieve this, medical institutions and businesses should incorporate IoMT medical devices. In addition, there will be an increasing volume of inconsistent data as more Internet-connected medical devices (IoMT) are connected to the Internet [6]. CCS, which encapsulates all centre activities, currently faces significant challenges due to high latency (HL), network dependencies, and individual failure points.

A difficult aspect of instantaneous transactions is adjusting to them [7]. The concept of fog computing (EC) is based on the assumption that network edge services should be efficient in terms of time and resources. The Fog Layer (FL) manages edge-to-cloud operations. Figure 1 illustrates how fog IoMT works [8], [9]. In Figure 1, IOT is illustrated in a variety of contexts. In addition to transportation, industry, homes, healthcare, and education, there are many other domains where technology is used. As a result, the planner is able to establish and deliver a service more efficiently, thus reducing resource consumption and latency. With the help of the 5G network, mobile networks can connect to machines and operate [10].



Figure 1: The industrial IoT and its real-world applications.

2 LITERATURE REVIEW

The security and protection of healthcare data remain challenging despite advancements in emerging technologies such as the Internet of Things (IoT), Edge Computing, Artificial Intelligence, and Machine Learning [11], [12]. Modern attacks may pose a serious threat to the lives of patients whose sensitive and confidential information has been compromised or leaked.

The authors provide an analysis of EHR analytics using deep learning-based approaches [8]. The authors have discussed the limitations of existing frameworks, as well as models that are applied to heterogeneous data sets, patient records, and certificates. In basic image processing research, images are represented in more complicated, hierarchical ways and creatively processed using greater, more sophisticated structures [13]. Earlier this year, the author of an article [9] argued that Blockchain technologies are capable of facilitating cloud-based electronic health record sharing. Through blockchain technology, it is possible to share electronic health records (EHR) in the cloud. Blockchain and ABSC have been used to secure the sharing of EHRs [14]. The proposed model ensures the accuracy, security, and enforceability of EHRs. In addition to protecting users' privacy, cloud-based EHRs are also secured, enabling them to identify themselves. Authorized users can upload cloud-based EHRs, and signer identities remain anonymous, but they are not suitable for severe patients due to the high latency.

As outlined by the Author [15], a blockchain-enabled cloud computing approach can be used to exchange medical records. In addition to enhancing security and privacy, blockchain facilitates authorization and other functions. Consortium blockchains contain three key components: a network model, a mechanism for generating data, and a consensus mechanism. Currently, this solution has only been tested on Ethereum, so its efficiency cannot be guaranteed on other platforms due to the requirement of high bandwidth for data retrieval and storage.

These limitations can be overcome while securely exchanging information using blockchain technology [16]. The authors present modules for requesting and linking. The developed model enables patients to keep full control of their health data and ensure their privacy at all times. The current healthcare system could be made more secure, stable, and robust using blockchain technology. As a result, the developed strategy has the main weakness of

relying on the properties of a blockchain node to perform. A major disadvantage of developed strategies is that their performance is dependent on the properties of the blockchain nodes. Using MedShare, the Author [17] proposes an interoperability-focused cloud-based system for patient data exchange that can overcome existing barriers. Mediation is when independent healthcare providers do not want to share patient data with their consumers or with their competitors. Due to its reliability requirement, the proposed scheme relies on the public cloud, and the data transformer required to implement it is costly.

Attribute-based sign encryption is used in the Personal Health Record Exchange described by the author [18]. In CP-OABSC, server-aided signature verification is used in conjunction with ABE schemes and verified outsourced decryption. This paper uses attribute-based encryption techniques to demonstrate the robustness, reliability, and security of a hybrid encryption method.

3 METHODOLOGY

An edge-deployed hybrid deep-learning model will be developed using this research methodology by training it distributed. Using their local data, edge and fog devices update their pre-trained models and evaluate their models. As a consequence of the incorporation of blockchain technology, all users are considered trusted users [19] - [21]. When communicating data over the Internet of Things network, maintaining data integrity can be challenging due to the security and flexibility of the access control scheme. IoT security mechanisms also face the challenge of distinguishing between normal and attack scenarios. A medical sensor, actuator, and machine network interconnects several fog nodes [22] - [24]. For blockchain technology to be integrated into cloud-based industrial systems, a new framework for deep learning and blockchain deployment is needed. A blockchain stores and hashes meta-data, while a cloud stores backups. Due to differences in computing power across edge nodes, such frameworks are often unable to store the complete block due to their scalability issues [22], [25]. Based on our framework, the following scenarios are presented. Our aim in this chapter is to discuss how blockchain can be integrated with hybrid IOT and how it can be used in fog computing [26]. Based on our framework, the following scenarios may be considered.

Using the previous chain's hash function, a blockchain can be maintained. As a result, the system has become more verifiable. In addition, the blockchain utilizes consensus to verify the integrity of the hash chain when generating new blocks [27]. A smart healthcare application is an essential part of 5G networks. Communication through 5G networks requires smart antennas, and they are shown as essential elements of the overall 5G intelligent health network architecture. As a result of recent advancements, smart antennas are becoming more attractive for deployment in 5G [22]. Well-coordinated RF beams enable perfect transmission of signals. As attenuation increases, the location becomes less important, as the focus of interest diminishes. 5G networks (IoMT) are predicted to support intelligent healthcare through machine-to-machine (M2M) and internet-of-things (IoT) connections [28]. The strategies given have two fundamental shortcomings. First, many terminals cause dense networks. A high degree of density and scalability is needed for M2M and IoMT applications. As an additional anxiety, low-based [29] applications that use wireless sensors pose a security risk because they do not provide secure consumption methods. An overview of the proposed model's timeline and its functions is presented in the diagram. Figure 1 presents a timeline for the proposed model and its functions.

A solid approach has been developed to reduce the threat posed by attackers, addressing some of the concerns expressed about security in wireless sensors. By developing an adversarial threat model, we will be able to identify vulnerabilities that unauthorized users can exploit to gain access to authorized sessions. Our goal here is to examine three distinct protocols that employ salted hashes but have their unique architectures, methodologies, and explanatory texts. The development of a blockchain-based healthcare system capable of communicating with a cyber-physical system must be methodical. An overview of how such a system might be created is provided:

- Problem identification. To develop a cyber-physical system, one must first identify the problems that need to be addressed. It's our objective here to make healthcare data more secure, private, and interoperable.
- Requirement definition. A cyber-physical system begins with identifying the problem, followed by defining the exact requirements. System specifications include functional and performance requirements, security concerns, and scalability concerns.

- Stakeholder identification. Assign stakeholders to the system based on their use and interaction with it. The list includes healthcare providers, patients, insurers, regulators, and others who are involved in the healthcare industry.
- Use case meaning. List the specific use cases associated with the cyber-physical system according to the identified requirements and stakeholders. A stakeholder's specific actions must be identified.
- Blockchain platform selection. System requirements should be met by the blockchain platform. In addition to Ethereum and HyperLedger, there are other blockchain-based platforms [30], [31].
- Smart contract description. Defining the smart contracts for the use cases will allow for the actions to be executed. The terms of an SC are incorporated directly into the agreement, so it is self-executed.
- Data structure description. The storage of healthcare data should be based on blockchain technology. Besides defining the field and type of data to be used, we must also determine the encryption mechanism.
- Consensus mechanism description. Decide how blockchain transactions will be validated by consensus. If a proof of stake (PoS) consensus mechanism is more appropriate than a proof of work (PoW), which is computationally expensive, then a proof of stake might be more appropriate.
- System testing and deployment. A controlled environment should be used for testing and deploying the system after the design phase is complete. In addition to evaluating the scalability and security of the system, it should be tested for functionality.
- System monitoring and maintenance. Monitoring and maintaining the system is essential after deployment. Security assessments are carried out regularly, potential breaches are dealt with, and potential breaches are dealt with in order to ensure the system consistently meets stakeholder requirements.

The blockchain-based cyber-physical healthcare system we designed and deployed is secure, scalable, and interoperable.

3.1 Cyber-Physical Systems (CPSs)

Interoperability in healthcare can be enabled by combining blockchains and CPSs. A three-protocol analysis of blockchain integration with healthcare

CPSs is presented in this review. This merger has both advantages and disadvantages. The healthcare industry generates a great deal of personal information, which must be preserved, transported, and updated constantly. Blockchain technology enhances data privacy as well as making it transparent, immutable, and distributed. Healthcare CPSs can maintain secure communications and control data transfers with blockchains by using the basic protocol, session protocol, and cookie protocol. CPSs use a request-response protocol model as a standard method for transferring data between participating companies. The immutability and transparency of blockchain data make this architecture possible. Blockchains record transactions securely and immutably between users. Every data transfer is audited and verified as part of the built-in system. It will enable doctors and patients to interact confidentially and have access to better data [31].

CPS components communicate through channels established and maintained by the sessions protocol model. This method is significantly enhanced by incorporating blockchain technology. Blockchain technology allows individuals and organizations to communicate securely, ensuring the confidentiality of any data they exchange. Unlike centralized ledgers, blockchain does not require a central session manager, thus removing potential vulnerabilities. In online applications, cookie protocols are commonly used to manage sessions and customize user preferences. This system is more confidential and secure with the addition of blockchain. Using blockchain technology, cookies can be stored and verified decentralized. In addition to increasing users' trust in the system, this enhancement also prevents cookies from being tampered with. Patients can also select which third parties they wish to receive their cookie data from. Blockchain can have several advantages when integrated into healthcare CPSs' fundamentals, sessions, and cookie protocols. All communication channels should use encryption, and patients should be given control over their data for improved data security, interoperability, and privacy.

3.2 Proposed Protocol

A user-specified difficulty level is used to select a reliable miner who will mine the hash under the recommended protocol. The solution we offer incorporates a method for assessing the trust value of the checker and the author, unlike previous attempts. Miners are selected randomly for block generating and validating using a trusted procedure. The

proposed blockchain-based protocol for cyber-physical health care systems comprises the following core components:

- User registration. It is necessary to register patients and healthcare providers, providing personal information and authentication methods.
- Data collection. Healthcare data can be collected using a variety of devices, including medical equipment, electronic health records, and wearables.
- Data encryption. For healthcare data acquired on the blockchain to be securely stored, it is encrypted before it is stored.
- Data storage. Blockchains provide a secure, decentralized way to store encrypted healthcare information.
- Smart contract execution. Healthcare data is securely controlled by smart contracts executed on the blockchain and is accessible only to authorised users.
- Data access. Healthcare data is stored on the blockchain, which is accessible to patients and providers by verifying their identity.
- Data sharing. It is possible to share health data among authorized parties using blockchain technology securely and transparently.
- Consensus mechanism. Proof of Stake (PoS) is one of the consensus mechanisms used in blockchains for validating transactions and maintaining data integrity.
- Audit trail. Transactions and activities related to health are secured and transparently recorded on a blockchain.
- Compliance with regulations: In accordance with HIPAA and GDPR, blockchain-based healthcare systems secure and protect patient information.

Compared to current healthcare paradigms, blockchain-based systems are more efficient, effective, and secure and offer greater privacy, transparency, and accountability. For healthcare information to be confidential, additional processes or systems may be required based on stakeholders' needs and expectations.

3.3 Proposed Neural Network

Known for its capability to classify images, the convolutional neural network (CNN) is a deep learning knowledge-based neural network. An initial layer of linear convolutions (conv) and an additional layer of fully connected (FC) are included in the

model. The nonlinear function is applied over the linear function in this model. As a result, this nonlinear function impacts all layers of the input and pooling layers and minimizes their size. In many parts of images, multiple perceptrons are used to separate pixel values using learned weights and bias values. A learnable weight and bias value is used to train perceptrons. CNNs have several advantages, including using a local spatial domain for the input images, sharing parameters, and using fewer weights. There are a number of advantages to this technique, including its lower computational complexity and smaller memory requirements. Figure 2 illustrates the CNN architecture:

- Convolutional layer. It essentially resizes the input image to 224 x 224, which is CNN's standard size. Resized images undergo multiple convolutional layers with different receptive fields. Feature data are extracted from the input matrix using a convolution layer. Then, these data are mapped to successive layers through a sequence of mathematical operations over a sliding kernel matrix. A feature matrix is generated by combining the results of element-wise matrix multiplication at each coordinate. In physics, image processing, and statistics, recurrent convolutions are used as elegant linear models. Several axes are used to estimate convolution. Convolved images are calculated by multiplying input images by kernel filter coefficients and dividing by input images.

$$S(i, j) = \sum_m \sum_n I(-m, n) k(i - m, j - n). \quad (1)$$

- Pooling layer. Following the convolutional layer comes the pooling layer, which reduces the spatial domain representation, thereby reducing network computations. Typically, CNN uses a long and wide pooling kernel with a long and wide stride.
- Fully connected layer. As a result of convolution, CNN replicates this FC layer. There are normally two tensors in the tensor space, n1 and n2, where n1 is the input tensor size, and n2 is the output tensor size.
- Dropout. In deep learning, this layer is known as Drop. It is usually used to eliminate input overfits; the primary purpose is to improve conjecture and prediction. A weight is normally assigned to each node in a network.
- Softmax. Convolutional layers are followed by ReLU layers, which determine the nonlinearity of CNN and improve it accordingly.

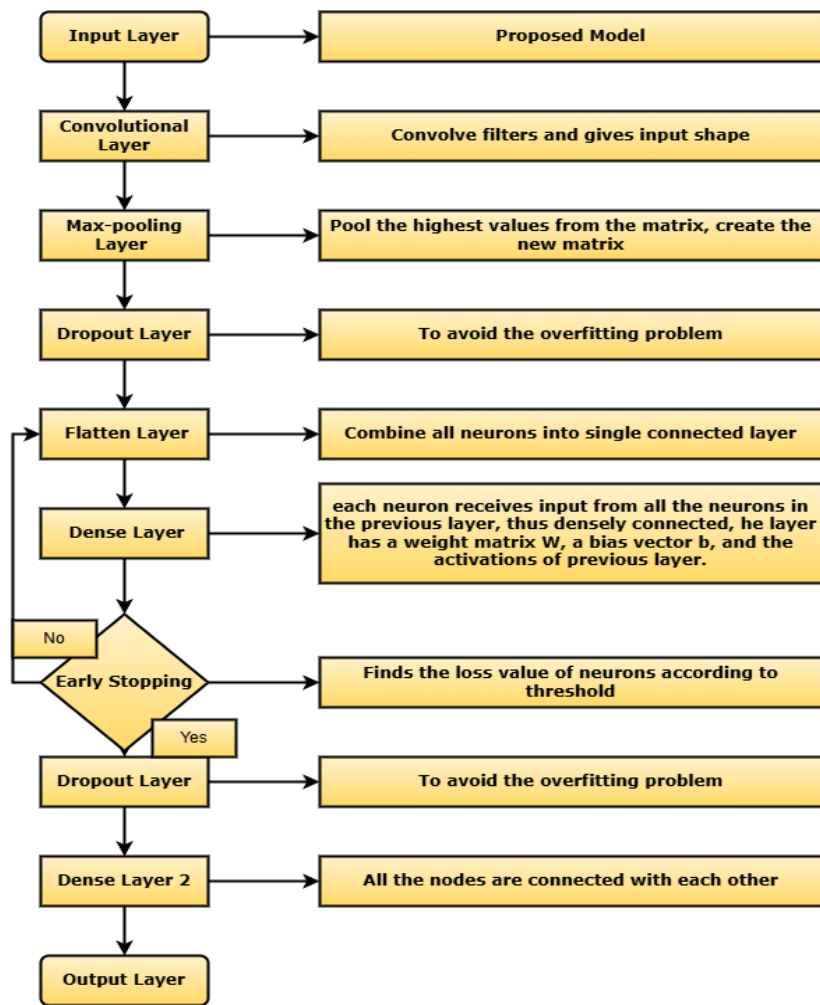


Figure 2: An overview of the convolutional neural network architecture.

4 RESULTS AND DISCUSSION

Models that are currently available were compared with the proposed framework. Compared to the benchmark models, the new framework performs better than the old framework (Fig. 3).

Using the benchmark model as a benchmark, the simulation included varying the percentage of malicious nodes to compare the proposed framework with its performance. A total of 150 malicious nodes were detected. When there are more than 50% of malicious nodes, accuracy declines significantly. In comparison with the other frameworks, the proposed one outperforms them due to its use of positive values to select honest miners. This performance advantage is further reflected in the cache efficiency, as shown in Figure 4, which illustrates the cache hit rate versus the number of blocks. Real blockchain scenarios typically do not have more than 50% malicious nodes.

Therefore, the proposed framework would be more reliable.

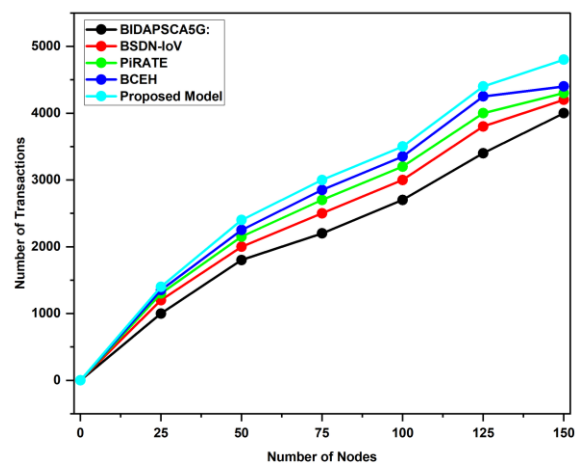


Figure 3: The number of transactions based on node count.

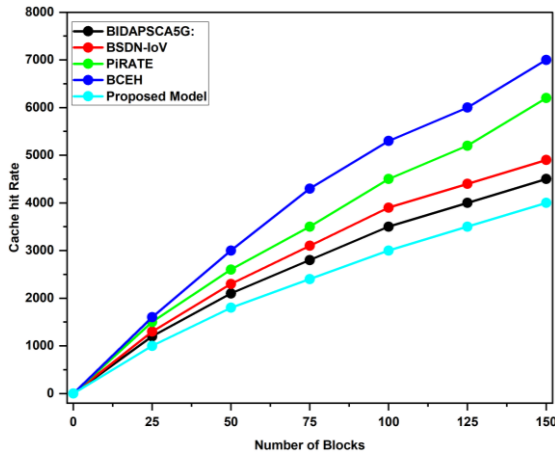


Figure 4: Cache hit rate versus number of blocks.

Figure 5 illustrates how the proposed framework enhanced security while reducing risks significantly. To assess how efficient single block creation can be, a second simulation was conducted using different numbers of nodes. A simulation based on the average time it takes to generate 80 (or 50) rounds of blocks was used to make sure the results were generalizable. A high level of safety and security can be maintained with less computational resources using this framework.

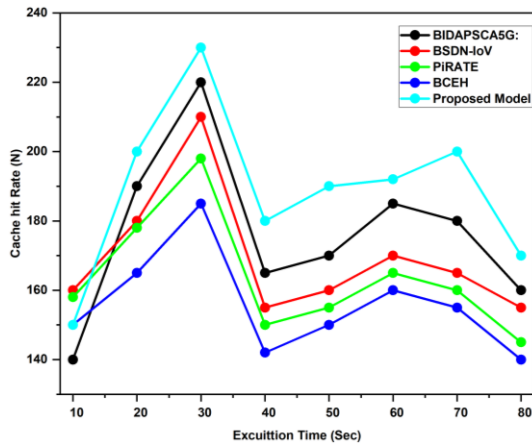


Figure 5: Cache hit rate versus execution time (sec).

Using Figure 6, we can see the time and number of records involved in the simulation. The number of records found ranged from 500 to 12,500. A comparison of the proposed framework with the benchmark model reveals that the proposed framework takes significantly less time to run for the same number of rounds as the benchmark model, as illustrated in Figure 6.

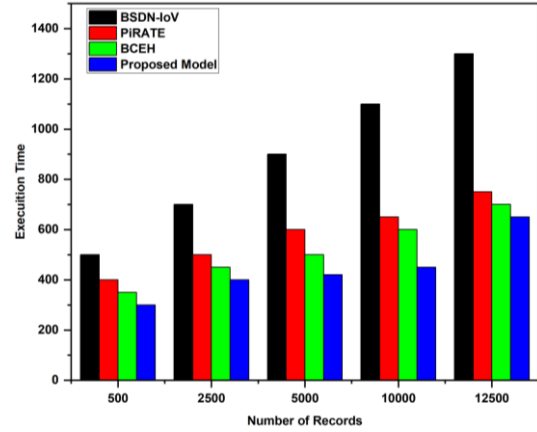


Figure 6: Execution time and record count simulation results.

Based on the simulation results, it is evident that the number of blocks and processing time are correlated significantly (Fig. 7). Our framework outperformed benchmark models based on a comparison we conducted with benchmark models.

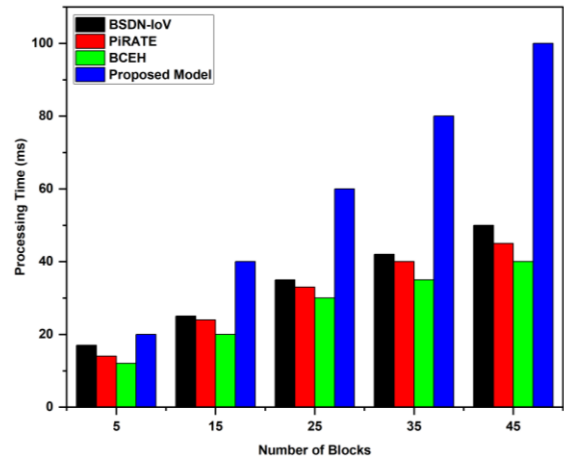


Figure 7: Process time and block count simulation results.

5 CONCLUSIONS

Cyber-Physical Healthcare Systems (CPHS) based on IoT are presented in this paper, leveraging blockchain technology and cryptography to ensure data integrity and security. In addition to effectively addressing critical security challenges, the proposed model offers robust protection for sensitive healthcare data against risks such as unauthorized access, data breaches, and information tampering. Simulation results demonstrate that the designed framework outperforms current methodologies by significantly

enhancing data management capabilities, system efficiency, and security safeguards, which collectively improve patient trust and operational reliability. Moving forward, researchers and practitioners should focus on expanding the framework's scalability, ensuring adaptability to diverse healthcare settings - including remote clinics and large hospitals - and thoroughly assessing practical deployment barriers. Further investigation is recommended into integrating emerging technologies, such as 5G connectivity, wearable healthcare devices, and intelligent analytics solutions, to optimize patient care delivery, enhance real-time monitoring capabilities, and maintain rigorous standards of patient privacy and data protection.

REFERENCES

- [1] P. Rani and M. H. Falaah, "Real-Time Congestion Control and Load Optimization in Cloud-MANETs Using Predictive Algorithms," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 66–76, 2024.
- [2] A. Singh et al., "Smart Traffic Monitoring Through Real-Time Moving Vehicle Detection Using Deep Learning via Aerial Images for Consumer Application," *IEEE Trans. Consum. Electron.*, vol. 70, no. 4, pp. 7302–7309, Nov. 2024, doi: 10.1109/TCE.2024.3445728.
- [3] A. Ali et al., "Performance analysis of AF, DF and DtF relaying techniques for enhanced cooperative communication," in *Proc. 6th Int. Conf. Innovative Computing Technology (INTECH)*, Dublin, Ireland: IEEE, Aug. 2016, pp. 594–599, doi: 10.1109/INTECH.2016.7845056.
- [4] P. Rani et al., "Federated Learning-Based Misbehavior Detection for the 5G-Enabled Internet of Vehicles," *IEEE Trans. Consum. Electron.*, vol. 70, no. 2, pp. 4656–4664, May 2024, doi: 10.1109/TCE.2023.3328020.
- [5] Z. Mushtaq et al., "Automatic Agricultural Land Irrigation System by Fuzzy Logic," in *Proc. 3rd Int. Conf. Information Science and Control Engineering (ICISCE)*, Beijing, China: IEEE, Jul. 2016, pp. 871–875, doi: 10.1109/ICISCE.2016.190.
- [6] A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "MedRec: Using Blockchain for Medical Data Access and Permission Management," in *Proc. 2nd Int. Conf. Open and Big Data (OBD)*, Vienna, Austria: IEEE, Aug. 2016, pp. 25–30, doi: 10.1109/OBD.2016.11.
- [7] M. Hasnain, M. F. Pasha, I. Ghani, B. Mehboob, M. Imran, and A. Ali, "Benchmark Dataset Selection of Web Services Technologies: A Factor Analysis," *IEEE Access*, vol. 8, pp. 53649–53665, 2020, doi: 10.1109/ACCESS.2020.2979253.
- [8] A. Ali et al., "Security, Privacy, and Reliability in Digital Healthcare Systems Using Blockchain," *Electronics*, vol. 10, no. 16, p. 2034, Aug. 2021, doi: 10.3390/electronics10162034.
- [9] P. Rani, P. N. Singh, S. Verma, N. Ali, P. K. Shukla, and M. Alhassan, "An implementation of modified blowfish technique with honey bee behavior optimization for load balancing in cloud system environment," *Wirel. Commun. Mob. Comput.*, vol. 2022, pp. 1–14, 2022.
- [10] A. Ali, M. Naveed, M. Mehboob, H. Irshad, and P. Anwar, "An interference aware multi-channel MAC protocol for WASN," in *Proc. Int. Conf. Innovations in Electrical Engineering and Computational Technologies (ICIEECT)*, Karachi, Pakistan: IEEE, Apr. 2017, pp. 1–9, doi: 10.1109/ICIEECT.2017.7916523.
- [11] S. S. Gill et al., "Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots," *Internet Things Cyber-Phys. Syst.*, vol. 4, pp. 19–23, 2024, doi: 10.1016/j.iotcps.2023.06.002.
- [12] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," *Comput. Electr. Eng.*, vol. 105, p. 108543, 2023.
- [13] N. Kumar, P. Rani, V. Kumar, S. V. Athawale, and D. Koundal, "THWSN: Enhanced energy-efficient clustering approach for three-tier heterogeneous wireless sensor networks," *IEEE Sens. J.*, vol. 22, no. 20, pp. 20053–20062, 2022.
- [14] A. Singh et al., "Blockchain-Based Lightweight Authentication Protocol for Next-Generation Trustworthy Internet of Vehicles Communication," *IEEE Trans. Consum. Electron.*, vol. 70, no. 2, pp. 4898–4907, May 2024, doi: 10.1109/TCE.2024.3351221.
- [15] Y. Wang, A. Zhang, P. Zhang, and H. Wang, "Cloud-Assisted EHR Sharing With Security and Privacy Preservation via Consortium Blockchain," *IEEE Access*, vol. 7, pp. 136704–136719, 2019, doi: 10.1109/ACCESS.2019.2943153.
- [16] Y. Zhuang, L. R. Sheets, Y.-W. Chen, Z.-Y. Shae, J. J. P. Tsai, and C.-R. Shyu, "A Patient-Centric Health Information Exchange Framework Using Blockchain Technology," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 8, pp. 2169–2176, Aug. 2020, doi: 10.1109/JBHI.2020.2993072.
- [17] Y. Yang et al., "Medshare: A Novel Hybrid Cloud for Medical Resource Sharing Among Autonomous Healthcare Providers," *IEEE Access*, vol. 6, pp. 46949–46961, 2018, doi: 10.1109/ACCESS.2018.2865535.
- [18] F. Deng, Y. Wang, L. Peng, H. Xiong, J. Geng, and Z. Qin, "Ciphertext-Policy Attribute-Based Signcryption With Verifiable Outsourced Designcryption for Sharing Personal Health Records," *IEEE Access*, vol. 6, pp. 39473–39486, 2018, doi: 10.1109/ACCESS.2018.2843778.
- [19] A. A. Shah, G. Piro, L. A. Grieco, and G. Boggia, "A qualitative cross-comparison of emerging technologies for software-defined systems," in *Proc. 6th Int. Conf. Software Defined Systems (SDS)*, IEEE, 2019, pp. 138–145. Accessed: Apr. 03, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8768566/>.

- [20] B. Jia, T. Zhou, W. Li, Z. Liu, and J. Zhang, "A Blockchain-Based Location Privacy Protection Incentive Mechanism in Crowd Sensing Networks," *Sensors*, vol. 18, no. 11, p. 3894, Nov. 2018, doi: 10.3390/s18113894.
- [21] A. Shah, G. Piro, L. Grieco, and G. Boggia, "A review of forwarding strategies in transport software-defined networks," in *Proc. 22nd Int. Conf. Transparent Optical Networks (ICTON)*, 2020, pp. 1–4, doi: 10.1109/ICTON51198.2020.9203103.
- [22] T. M. Fernández-Caramés, I. Froiz-Míguez, O. Blanco-Novoa, and P. Fraga-Lamas, "Enabling the Internet of Mobile Crowdsourcing Health Things: A Mobile Fog Computing, Blockchain and IoT Based Continuous Glucose Monitoring System for Diabetes Mellitus Research and Care," *Sensors*, vol. 19, no. 15, p. 3319, Jul. 2019, doi: 10.3390/s19153319.
- [23] A. Beebeejaun, "VAT on foreign digital services in Mauritius; a comparative study with South Africa," *Int. J. Law Manag.*, vol. 63, no. 2, pp. 239–250, Feb. 2021, doi: 10.1108/IJLMA-09-2020-0244.
- [24] A. Cirstea, F. M. Enescu, N. Bizon, C. Stirbu, and V. M. Ionescu, "Blockchain Technology Applied in Health: The Study of Blockchain Application in the Health System (II)," in *Proc. 10th Int. Conf. Electronics, Computers and Artificial Intelligence (ECAI)*, Iasi, Romania: IEEE, Jun. 2018, pp. 1–4, doi: 10.1109/ECAI.2018.8679029.
- [25] R. R. Bruce, J. P. Cunard, and M. D. Director, *From Telecommunications to Electronic Services: A Global Spectrum of Definitions, Boundary Lines, and Structures*. Oxford, UK: Butterworth-Heinemann, 2014. Accessed: Apr. 03, 2025. [Online]. Available: <https://books.google.com/books?id=AEeeBQAAQB-AJ>.
- [26] P. Rani, U. C. Garjola, and H. Abbas, "A Predictive IoT and Cloud Framework for Smart Healthcare Monitoring Using Integrated Deep Learning Model," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 53–65, 2024.
- [27] A. Yazdinejad, G. Srivastava, R. M. Parizi, A. Dehghantanha, K.-K. R. Choo, and M. Aledhari, "Decentralized Authentication of Distributed Patients in Hospital Networks Using Blockchain," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 8, pp. 2146–2156, Aug. 2020, doi: 10.1109/JBHI.2020.2969648.
- [28] A. Singh et al., "Resilient wireless sensor networks in industrial contexts via energy-efficient optimization and trust-based secure routing," *Peer-Peer Netw. Appl.*, vol. 18, no. 3, p. 132, Jun. 2025, doi: 10.1007/s12083-025-01946-5.
- [29] A. Ali, M. Naveed, M. Mehboob, H. Irshad, and P. Anwar, "An interference aware multi-channel MAC protocol for WASN," in *Proc. Int. Conf. Innovations in Electrical Engineering and Computational Technologies (ICIEECT)*, IEEE, 2017, pp. 1–9. Accessed: Apr. 03, 2025. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/7916523/>.
- [30] M. M. Nair, A. K. Tyagi, and R. Goyal, "Medical Cyber Physical Systems and Its Issues," *Procedia Comput. Sci.*, vol. 165, pp. 647–655, 2019, doi: 10.1016/j.procs.2020.01.059.
- [31] X. Liu, B. Xu, X. Wang, K. Zheng, K. Chi, and X. Tian, "Impacts of Sensing Energy and Data Availability on Throughput of Energy Harvesting Cognitive Radio Networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 747–759, Jan. 2023, doi: 10.1109/TVT.2022.3204310.

Privacy-Preserving Machine Learning Using Consortium Blockchain in Vehicular Social Networks

Amna Arak Radeini Asal and Shaimaa Jabbr Ali

*Department of Arabic, College of Education for Women, University of Anbar, 31001 Ramadi, Iraq
amn21w5005@uoanbar.edu.iq, shaimaaja@uoanbar.edu.iq*

Keywords: Vehicular Social Networks (VSNs), Privacy-Preserving Machine Learning, Consortium Blockchain, Threshold Paillier Cryptosystem, Support Vector Machine (SVM).

Abstract: It presents both opportunities for intelligent transportation systems and challenges for ensuring privacy during the data analysis process that VSNs (Vehicular Social Networks) generate. Our paper proposes a threshold Paillier cryptosystem and consortium blockchain for training Support Vector Machines (SVMs) on vertically partitioned datasets. Traditional approaches that rely on trusted third parties are insecure compared to blockchain-based collaboration. Most computations are performed locally and intermediate values are only shared if they are encrypted. This ensures high levels of privacy and efficiency. This model (PP-SVM) offers classification accuracy comparable to standard SVMs, resulting in privacy-preserving learning environments in virtual social networks. As a result of this approach, sensitive user data is effectively protected, and a robust sense of trust among network members is fostered. In addition to ensuring data integrity, consortium blockchain technology promotes collaborative learning by leveraging its inherently decentralized nature, facilitating secure interactions and shared decision-making processes. With the rapid evolution and increasing adoption of vehicular social networks, preserving user privacy has become increasingly crucial, demanding scalable and reliable security mechanisms.

1 INTRODUCTION

Vehicular Social Networks (VSNs) are a significant advancement in intelligent transportation systems, enabling vehicle-to-vehicle and infrastructure-to-network data exchange. As a result of these networks, traffic management, safety, and user experience can be improved [1]. A critical challenge remains, however: securing sensitive data and enabling collaborative, decentralized decision-making while protecting privacy. The increasing amount of real-time, location-based, and personal data being shared in VSNs makes privacy preservation essential. Blockchain-based machine learning (ML) is one promising solution for addressing privacy concerns while allowing intelligent data processing. In particular, consortium blockchains provide a secure and transparent platform for managing data and facilitating collaborative learning, as they are permissioned blockchains where trusted parties collaborate. A similar setup would allow vehicles to contribute data for training machine learning models without exposing their sensitive information to others.

Several scenarios have benefited from cloud and edge computing, including mobile social networks (VSNs) [2], [3]. VSNs can be optimized in terms of safety, convenience, amenity, and entertainment with the use of efficient methods [4], [5]. The application of machine learning and deep learning has captured the interest of researchers across a wide range of disciplines [6], [7]. Data analysis of VSNs is gradually advancing with the application of related technologies, including machine learning and deep learning. In many scenarios, support vector machines (SVMs) are frequently used due to their efficiency and robustness. Negative communication conditions can be detected using SVM, for example. VSNs collect data from multiple sources, including vehicle manufacturers, vehicle management agencies, and social networking application developers [8], [9].

Data sources differ between these entities, so their attributes are different. A VSN rarely has a comprehensive training dataset due to the limited number of sources available. Vehicle network service platforms control the majority of vehicle location and mobility trace attributes, and social network applications may control user preferences. As you

might expect, the performance of SVM classifiers is determined by datasets [10]. A VSN application trains SVM classifiers by combining datasets with sufficient attributes before reaching a more efficient model. An alternative approach is to segment the merged dataset vertically based on the datasets owned by the entities. SVM training presents several serious security challenges, however. VSNs contain extensive private information (e.g., vehicle location, user preferences), so companies are prohibited from using and sharing this information across borders due to constantly enacted regulations. VSN data, however, contain high-value information that makes data providers hesitant to share their original data. The value of shared data is lost if a privacy-preserving mechanism is not in place. Several security challenges are associated with intelligent and connected vehicles [9], [11], [12]. Data sharing between entities is also a problem that should be well-researched.

2 LITERATURE REVIEW

Several methods of machine learning have been investigated that preserve privacy, including linear regression, SVMs, naive Bayes classifiers, and logic regressions. In recent years, deep learning [13] has also gained attention. The author [14] trains a linear regression classifier using a hybrid approach based on vertically partitioned datasets. Two parties compute using garbled circuits in this protocol. Two parties require a crypto service provider, and multiple parties require an evaluator and crypto service provider simultaneously.

In [15], the author developed an algorithm using heteromorphic encryption, and Yao garbled circuits were separated. Crypto service providers and evaluators are essential for realizing the algorithm [16]. Several machine learning algorithms, such as linear regression, logistic regression, and neural networks, are supported [17], [18]. With the use of two-party computation, data is collected from data providers, and a model is trained securely. Neither server can work simultaneously with the other. An author [19] developed a framework using three servers to train linear regression models, logistic regression models, and neural networks.

2.1 Machine Learning in Vehicular Social Networks (VSNs)

Vehicular social networks (VSNs) are adopting machine learning (ML) techniques for a variety of

applications [20]. Transport and communication can be enhanced through these techniques by leveraging data collected from vehicles and their surroundings. Intelligent and adaptive services can be provided by VSNs with ML algorithms, addressing the increasing demands of modern transportation systems. A growing volume of vehicle data necessitates machine learning to extract valuable insights and optimize network performance. As a consequence of these applications, road safety is enhanced, vehicular services are improved, and intelligent transportation systems are promoted [21]. With ML algorithms, drivers can be warned of potential hazards in real time and reduce the risk of accidents by predicting potential hazards. Additionally, machine learning helps optimize traffic flow through the dynamic adjustment of traffic signals and efficient routing of vehicles, which in turn reduces congestion and increases travel time. Additionally, the use of machine learning in VSNs facilitates customized services, such as navigation and entertainment options, which enhances the driving experience.

2.2 Blockchain Technology in Vehicular Social Networks

VSNs can benefit from blockchain technology in terms of security and efficiency [22]. As a result of its inherent properties, such as immutability and transparency, blockchains are ideally suited to address the security and efficiency challenges facing VSNs. In addition to ensuring data integrity and authenticity, blockchain enables VSNs to streamline data sharing and access control processes. Secure transactions are enhanced by blockchain's decentralization, anonymity, and trust properties. By eliminating a central authority, the network will be less prone to failures and will become more resilient. In addition to protecting participants' identities, anonymity also prevents tracking and profiling by unauthorized parties. Participants in the blockchain network benefit from the tamper-proof and verifiable properties of blockchain, resulting in a high level of trust. Combined, these features improve VSN security and efficiency.

3 PROPOSED METHODOLOGY

3.1 SVM Overview

SVM is explained first by describing the notation. When a prime superscript is applied to a vector, it converts it to a row vector. A matrix containing two

vectors x and y is denoted by the scalar product. $x'y$, and a matrix containing $\|x\|$ 2-norm is called x . Taking an n -dimensional input space and using $m * n$ matrix a to represent m data points. Data points A are labelled +1 or -1 using a $m * n$ diagonal matrix D . x_i is represented by the class label D_{ii} or d_i in short. In mathematics, e denotes a column vector of ones of any dimension. A matrix with arbitrary dimensions is called an identity matrix.

Here is an example of a linear binary classification task. This problem can be solved with SVM by identifying the separating hyperplane ($w \cdot x = \gamma$) that maximizes the margin between the hyperplane and the closest data points. When a boundary is located at a different distance from each support vector, we use the "soft" margin. A "hard" margin is expressed as $\frac{1}{\|w\|}$, as illustrated below. Following is a primal program that combines the standard SVM solution with the objective of maximizing margin while minimizing error [23], [24]:

$$\min_{w,y} \frac{1}{2} w'w + ve'y, \quad (1)$$

$$s. t. D(Aw - ey) + y \geq e \text{ and } y \geq 0. \quad (2)$$

The method minimizes both the margin and error (w' and $e'y$). SVM allows soft margins or error by including the slack variable y in constraint (2). As the objective function (1) minimizes the slack or error, it will be larger than zero if the point lies within the margins. This parameter (a user parameter) balances the margin size with the error. With the help of this optimization problem, we will be able to calculate the vector of weights w and the bias b . As soon as w and y are calculated, we can identify the class of a new data object x using $f(x) = w'x - \gamma$, where $f(x) > 0$ indicates a positive class, otherwise a negative class.

3.2 Threshold Cryptosystems of Paillier

Typically, threshold cryptosystems have two components:

- 1) the public key is distributed, and the secret key is shared;
- 2) the secret key is decrypted and signed based on the shared representation.

When there should be no knowledge of the secret key of an individual, threshold schemes are vital. We have previously proposed threshold RSA encryption and decryption [20], DSS encryption and decryption [20] and Paillier in multiparty settings [19].

We have previously proposed threshold RSA encryption and decryption systems [25], SS encryption and decryption [26], DSS [26] and Paillier in multiparty settings. There are some cases (e.g., ElGamal) where multiparty techniques can easily be applied to two-party settings. However, there is still no solution to the problem of anti-malicious two-party threshold Paillier encryption.

3.3 System Model

According to Figure 1, a system such as ours is composed of three components: DD, DP, and BSP (Blockchain Service Platform).

Data Device. Sensors, mobile devices, and other data-generating devices are examples of this. Using these devices, valuable data can be collected and analyzed.

Data Provider: It is the responsibility of data providers to generate, collect, store, and analyze data from a variety of sources. The equipment and methods used by these participants resulted in varying data sets. As far as attributes are concerned, these diverse data sets complement one another. The participants in this collaborative machine-learning effort also serve as model trainers and provide data. This paper's scheme involves participants performing most training tasks locally.

Blockchain Service Platform. This platform uses blockchain technology. The BSP provides participants with a transparent platform for sharing data, allowing them access to all the data stored there. Moreover, the BSP protects data records from unauthorized alteration by maintaining their integrity. Additionally, it provides robust security measures that prevent data from leaving the participants' domain from being accessed. Participants and BSPs communicate using encrypted communications to ensure data confidentiality and prevent leakage.

Threat Model. A single role is assigned to the data provider in our scheme. In the security model, we view participants as honest but curious, implying that while they are curious about others' data, they will abide by the rules. In addition, as participants interact with the BSP extensively, potential threats during this interaction process are also considered.

3.4 The Construction of Secure SVM

3.4.1 Overview of Secure SVM Training with IoT Data

SVM models are trained using data collected from multiple IoT data providers in this method. A provider of IoT data preprocesses IoT data instances,

encrypts them locally with their private keys, and generates transactions to record them on a blockchain. In the global ledger, encrypted data can be accessed by data analysts who are training the SVM model. Interaction with each data provider is a necessary part of the training process.

3.4.2 Blockchain-Based Encryption of Data Sharing

Our blockchain transaction structure enables the storage of encrypted IoT data. Inputs and outputs are the two main fields of a transaction. There are three fields in the input form: the data provider's address, the type of IoT device, and the encrypted data. Output fields include the data analyst's address, encrypted data, and IoT device type. Address fields contain 32-byte hash values. Using the homomorphic encryption Paillier, the encrypted data was generated. The Bitcoin blockchain stores encrypted data instances that are 128 bytes long, assuming the private key is 128 bytes long. IoT device type segments are 4 bytes long.

3.5 Building Blocks

Gradient Descent. It is possible to optimize the parameters of an SVM using several methods. Sequential minimum optimization algorithms (SMO) and gradient descent algorithms (GD) are examples of these algorithms. It is a method for optimizing biquadratic SVM programs. Also, linear SVM and sparse data are well handled. There are many comparisons, dots, and divisions involved in SMO, which makes it complex. As a result of applying SMO cryptographically, there is a great deal of cost associated with computing and communication. Based on GD, SVM optimization is a simple and efficient algorithm that involves only a few comparisons and multiplications of vectors. For this reason, we selected GD as the optimization algorithm for optimizing SVM model parameters.

In the GD method, the primary SVM is converted into an empirical loss minimization problem by using a penalty factor.

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m L(w, b, (x_i, y_i)). \quad (3)$$

In this equation, the hinge-loss function appears on the right side

$$C \sum_{i=1}^m L(w, b, (x_i, y_i)) = C \sum_{i=1}^m \max\{0, 1 - y_i(wx_i - b)\}. \quad (4)$$

In most cases, $\frac{1}{m}$ is the misclassification penalty, and C is the penalty.

GD consists of the following forms:

$$x_{n+1} = x_n - \lambda \Delta Grad(x_n). \quad (5)$$

Secure Polynomial Multiplication. A secure addition can be described as a homomorphic addition based on Paillier's homomorphic property

$$[m_1 + m_2] = [m_1] * [m_2] \pmod{N^2}. \quad (6)$$

As an example of secure subtraction, consider the following:

$$[m_1 - m_2] = [m_1] * [m_2]^{-1} \pmod{N^2}. \quad (7)$$

The modular multiplicative inverse is $[m]^{-1}$, which performs the operation

$$[m] * [m]^{-1} \pmod{N^2} = 1. \quad (8)$$

Functions can be used to compute $[m]^{-1}$

$$\phi(N), [m]^{-1} = [m]^{\phi(N)-1}. \quad (9)$$

The manipulation of ciphertext can be used to obtain polynomial multiplication,

$$[am_1 + bm_2] = [m_1^b] \pmod{N^2}. \quad (10)$$

During the training model, the calculated median value must be shared between all three participants. A threshold homomorphic encryption scheme is used as the solution to protect the shared data, ensuring its security and ensuring gradient calculation accuracy. As a basis for judging the manner in which the gradients should be updated, we construct equations (11), (12), and (7) using additive homomorphic encryption.

$$[a] = \left[\sum_{i=1}^n a^i \right] = \prod_{i=1}^n [a^i], \quad (11)$$

$$[r_2] = \left[\sum_{i=1}^n r_2^i \right] = \prod_{i=1}^n [r_2^i], \quad (12)$$

$$[ar_1 + r_2] = [ar_1][r_2] = \left[\prod_{i=1}^{r_1} [a][r_2] \right] = [a]^{r_1} [r_2]. \quad (13)$$

This solution determined the gradient update method by comparing the encoded calculation result with the constant 1.

3.6 Data Sharing on BSP and Security Analysis

The BSP simplifies point-to-point communications by securely computing intermediate values. On-chain data is managed, and smart contracts execute queries. During the iteration process, every participant uploads data twice: once for calculating the intermediate values (IVs) and once for calculating the decrypted values (DVs).

3.7 The Format of IVs

Iteration Round. Data exchange round, managed by smart contracts, during collaborative model training.

DP ID: At the time of data upload, the data owner is automatically identified.

Training Intermediate Value. When training the model, the encrypted state's intermediate value is used. The encrypted state sums each participant's values and compares them to 1. This comparison uses three cryptographic parameters:

- r1: Comparative integer that is unencrypted.
- r2: For comparison, a positive integer is encrypted.
- r3: A positive integer that has not been encrypted is used for comparison.

3.8 Random Positive Integer

In the next iteration, the data instances will be determined by a random generator for each participant.

3.9 The Format of DVs

Iteration Round. Identifies the round of data exchange, similar to IVs.

DP ID. Provides information about who owns the data.

Decrypted Value. Decrypting the result based on each participant's private key allows participants to obtain the final decryption result collectively.

4 RESULT ANALYSIS AND DISCUSSION

Figures 1 and 2 illustrates the PP-SVM's classification accuracy, showing no observable performance degradation compared to the standard SVM. In spite of the threshold Paillier encryption, the encryption and decryption processes maintain

computational precision, ensuring that the classifier remains effective.

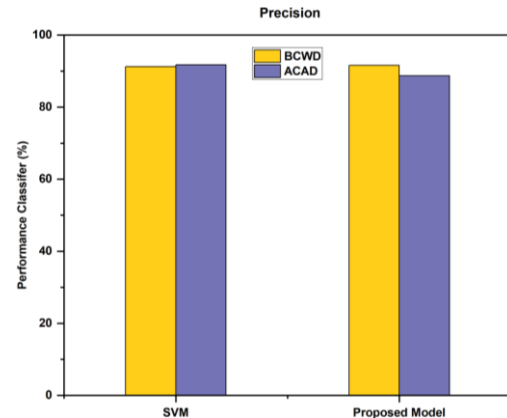


Figure 1: Performance of classifier accuracy.

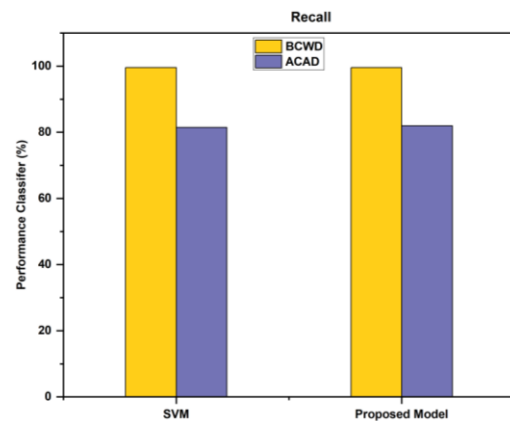


Figure 2: Recall comparison of SVM and PP-SVM on BCWD and ACAD.

In Figure 2, both the standard SVM and PP-SVM models are compared using BCWD and ACAD datasets. There is little difference in recall scores between the two models, which indicates that PP-SVM is not affected by the privacy-preserving mechanism. Data privacy is protected while maintaining effectiveness in the proposed model.

In this study, the standard SVM model is compared with the proposed PP-SVM model on two datasets: BCWD and ACAD. BCWD dataset has two yellow bars, one representing performance on the ACAD dataset and one representing performance on the BCWD dataset. Only slight differences occur in accuracy between the two models across the datasets. The proposed privacy-preserving PP-SVM model performs as well as a standard SVM without sacrificing performance, proving its efficiency in secure data processing.

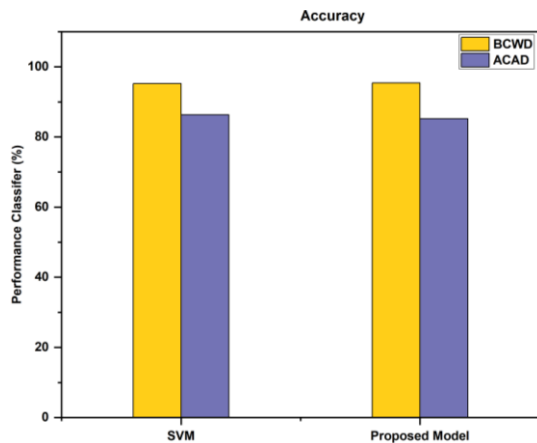


Figure 3: Accuracy with different numbers of VDPs.

Both datasets are represented in Figure 4. In general, the time spent in computation remains relatively low, with communications overhead taking up the majority of the time. Communication time remains elevated even as the training phase is completed quickly. Data sharing on consortium blockchains is often delayed due to consensus mechanisms among participating nodes, which require unavoidable delays.

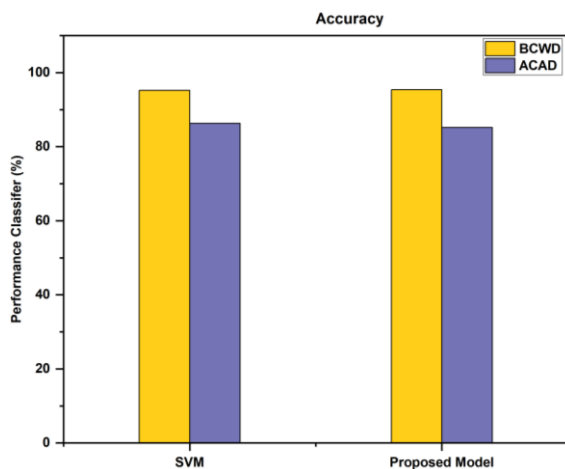


Figure 4: Performance of classifier efficiency.

5 CONCLUSIONS

SVM classifiers can be trained in VSNs using a privacy-preserving technique based on consortium blockchains and threshold Paillier encryption. As a result of the proposed model, data providers will no longer have to depend on trusted third parties to train

accurate classifiers, thereby significantly enhancing user autonomy, reducing privacy risks, and effectively safeguarding data confidentiality. This novel system strategically minimizes communication overheads and costs while ensuring high computing precision by encrypting only intermediate computation values, subsequently optimizing them through gradient descent algorithms. A thorough evaluation of the PP-SVM model demonstrates that its classification accuracy is consistently comparable to traditional SVM methods, and confirms its robustness and adaptability under diverse operational configurations and settings. Furthermore, consortium blockchain technology enhances data sharing processes by improving transparency, enforcing stringent security protocols, and maintaining strong access control mechanisms. Consequently, this framework represents a practical, scalable, and robust solution for securing machine learning processes in decentralized networks where privacy preservation and secure collaboration are critical.

REFERENCES

- [1] N. Hussain, P. Rani, N. Kumar, and M. G. Chaudhary, "A deep comprehensive research architecture, characteristics, challenges, issues, and benefits of routing protocol for vehicular ad-hoc networks," *Int. J. Distrib. Syst. Technol. (IJ DST)*, vol. 13, no. 8, pp. 1–23, 2022.
- [2] N. Cheng et al., "Big data driven vehicular networks," *IEEE Netw.*, vol. 32, no. 6, pp. 160–167, 2018.
- [3] Y. Zhang et al., "BDS: a centralized near-optimal overlay network for inter-datacenter data replication," in *Proc. 13th EuroSys Conf.*, Porto, Portugal: ACM, Apr. 2018, pp. 1–14, doi: 10.1145/3190508.3190519.
- [4] A. M. Vegni and V. Loscri, "A Survey on Vehicular Social Networks," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 4, pp. 2397–2419, 2015, doi: 10.1109/COMST.2015.2453481.
- [5] N. Hussain and P. Rani, "Comparative studied based on attack resilient and efficient protocol with intrusion detection system based on deep neural network for vehicular system security," in *Distributed Artificial Intelligence*, CRC Press, 2020, pp. 217–236.
- [6] N. Kato et al., "The Deep Learning Vision for Heterogeneous Network Traffic Control: Proposal, Challenges, and Future Perspective," *IEEE Wirel. Commun.*, vol. 24, no. 3, pp. 146–153, Jun. 2017, doi: 10.1109/MWC.2016.1600317WC.
- [7] G. Xu, H. Li, S. Liu, K. Yang, and X. Lin, "VerifyNet: Secure and verifiable federated learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 911–926, 2019.
- [8] D. X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in *Proc. IEEE Symp. Security and Privacy (S&P)*, 2000, pp. 44–55. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/848445/>

- [9] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, "Machine Learning Classification over Encrypted Data," in Proc. 2015 Network and Distributed System Security Symp., San Diego, CA: Internet Society, 2015, doi: 10.14722/ndss.2015.23241.
- [10] G. Ansari, P. Rani, and V. Kumar, "A novel technique of mixed gas identification based on the group method of data handling (GMDH) on time-dependent MOX gas sensor data," in Proc. Int. Conf. Recent Trends in Computing (ICRTC 2022), Springer, 2023, pp. 641–654.
- [11] Y. Li, Q. Luo, J. Liu, H. Guo, and N. Kato, "TSP Security in Intelligent and Connected Vehicles: Challenges and Solutions," IEEE Wirel. Commun., vol. 26, no. 3, pp. 125–131, Jun. 2019, doi: 10.1109/MWC.2019.1800289.
- [12] A. Singh et al., "Blockchain-Based Lightweight Authentication Protocol for Next-Generation Trustworthy Internet of Vehicles Communication," IEEE Trans. Consum. Electron., vol. 70, no. 2, pp. 4898–4907, May 2024, doi: 10.1109/TCE.2024.3351221.
- [13] R. Shokri and V. Shmatikov, "Privacy-Preserving Deep Learning," in Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur., Denver, CO, USA: ACM, Oct. 2015, pp. 1310–1321, doi: 10.1145/2810103.2813687.
- [14] A. Gascón et al., "Secure linear regression on vertically partitioned datasets," IACR Cryptol. EPrint Arch., vol. 2016, p. 892, 2016.
- [15] V. Nikolaenko, U. Weinsberg, S. Ioannidis, M. Joye, D. Boneh, and N. Taft, "Privacy-Preserving Ridge Regression on Hundreds of Millions of Records," in Proc. IEEE Symp. Security and Privacy (S&P), Berkeley, CA: IEEE, May 2013, pp. 334–348, doi: 10.1109/SP.2013.30.
- [16] P. Rani and R. Sharma, "An experimental study of IEEE 802.11n devices for vehicular networks with various propagation loss models," in Proc. Int. Conf. Signal Processing and Integrated Networks, Springer, 2022, pp. 125–135.
- [17] P. Mohassel and Y. Zhang, "SecureML: A System for Scalable Privacy-Preserving Machine Learning," in Proc. IEEE Symp. Security and Privacy (SP), San Jose, CA, USA: IEEE, May 2017, pp. 19–38, doi: 10.1109/SP.2017.12.
- [18] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," Comput. Electr. Eng., vol. 105, p. 108543, 2023.
- [19] P. Mohassel and P. Rindal, "ABY3: A Mixed Protocol Framework for Machine Learning," in Proc. 2018 ACM SIGSAC Conf. Comput. Commun. Secur., Toronto, Canada: ACM, Oct. 2018, pp. 35–52, doi: 10.1145/3243734.3243760.
- [20] M. Shen, J. Zhang, L. Zhu, K. Xu, and X. Tang, "Secure SVM Training Over Vertically-Partitioned Datasets Using Consortium Blockchain for Vehicular Social Networks," IEEE Trans. Veh. Technol., vol. 69, no. 6, pp. 5773–5783, Jun. 2020, doi: 10.1109/TVT.2019.2957425.
- [21] J. Cui, F. Ouyang, Z. Ying, L. Wei, and H. Zhong, "Secure and Efficient Data Sharing Among Vehicles Based on Consortium Blockchain," IEEE Trans. Intell. Transp. Syst., vol. 23, no. 7, pp. 8857–8867, Jul. 2022, doi: 10.1109/TITS.2021.3086976.
- [22] H. Shen, J. Zhou, Z. Cao, X. Dong, and K.-K. R. Choo, "Blockchain-Based Lightweight Certificate Authority for Efficient Privacy-Preserving Location-Based Service in Vehicular Social Networks," IEEE Internet Things J., vol. 7, no. 7, pp. 6610–6622, Jul. 2020, doi: 10.1109/JIOT.2020.2974874.
- [23] H. H. S. Office for Civil Rights, "Standards for privacy of individually identifiable health information. Final rule," Fed. Regist., vol. 67, no. 157, pp. 53181–53273, 2002.
- [24] V. Vapnik and V. Vapnik, Statistical Learning Theory. New York: Wiley, 1998.
- [25] R. Gennaro, S. Jarecki, H. Krawczyk, and T. Rabin, "Secure Distributed Key Generation for Discrete-Log Based Cryptosystems," in Advances in Cryptology – EUROCRYPT '99, J. Stern, Ed., Lecture Notes in Computer Science, vol. 1592. Berlin, Heidelberg: Springer, 1999, pp. 295–310, doi: 10.1007/3-540-48910-X_21.
- [26] C. Hazay, G. L. Mikkelsen, T. Rabin, and T. Toft, "Efficient RSA Key Generation and Threshold Paillier in the Two-Party Setting," in Topics in Cryptology – CT-RSA 2012, O. Dunkelman, Ed., Lecture Notes in Computer Science, vol. 7178. Berlin, Heidelberg: Springer, 2012, pp. 313–331, doi: 10.1007/978-3-642-27954-6_20.

Blockchain-Driven Security, Privacy and Reliability for Digital Healthcare Systems

Radhiya Sulaiman Nasser Alhabsi and Alla Salim Mohammed Almukhaini

University of Nizwa, 616 Nizwa, Sultanate of Oman
RadhiyaAlhabsi@unizwa.edu.om, 12468128@uofn.edu.om

Keywords: IoT-Blockchain Platform, Data Integrity, Smart Contracts, Sensor Data Management, Decentralized Security.

Abstract: With the integration of digital technologies in healthcare, several transformative advances have been made, including the management and sharing of patient data. Security, privacy, and system reliability are among the challenges presented by digitizing health data. Health data integrity, confidentiality, and reliability are all ensured by blockchain technology because of its decentralized, immutable nature. The purpose of this paper is to explore the potential of blockchain technology by using it to address medical imaging, diagnosis, and secure data sharing. Combining distributed ledger technology and cryptography can improve the security, privacy, and interoperability of healthcare systems. Using blockchain-driven AI models, the study illustrates how medical imaging applications can be significantly enhanced through secure, auditable, and transparent data-sharing practices, thus increasing stakeholder trust and substantially enhancing the scalability and reliability of AI-driven systems. By effectively combining blockchain's decentralized security features with advanced artificial intelligence techniques, diagnostic accuracy can be improved, patient care and clinical decision-making processes can be optimized, and strict regulatory compliance in medical imaging environments can be consistently maintained. This integrated approach also facilitates smoother collaboration among medical professionals.

1 INTRODUCTION

With the advent of digital technologies, the storage, sharing, and analysis of healthcare data have been transformed. With EHRs, telemedicine platforms, and wearable health devices, patients now have real-time access to data and can make better decisions regarding their health. Patient and healthcare provider data are increasingly digitizing and being shared across platforms, resulting in greater risks of data breaches, unauthorized access, and system failures. Digital data can be secured and managed in a decentralized, immutable, and transparent manner with blockchain technology. Healthcare has found extensive use for Blockchain's underlying features, including distributed ledger technology (DLT), cryptographic security, and smart contracts [1]. By means of blockchain technology, sensitive health data can be stored and shared in a secure, tamper-proof manner that ensures data integrity and confidentiality while allowing patients greater control over their health information. The decentralized nature of Blockchain can also mitigate the risks associated with

centralized data management, where a single point of failure may compromise system reliability [2].

A crucial component of modern healthcare is imaging and diagnostics, which improve outcomes and help make better decisions for patients. An X-ray or MRI generates a huge amount of complicated data that must be interpreted immediately. Healthcare systems worldwide face substantial challenges due to a global shortage of radiologists and diagnosticians, as well as an increase in the complexity of medical imaging data. As a result of artificial intelligence, image analysis, anomaly detection, and predictive diagnosis have been transformed [3], [4]. The use of artificial intelligence in medical imaging is not without challenges, including data security, interoperability, and trustworthiness. In order for AI-driven healthcare systems to be open, privacy-conscious, and efficient, comprehensive frameworks are needed. There are a number of difficulties that blockchain technology can solve [5].

Security, transparency, and accountability have all been improved through the decentralized and immutable ledger system blockchain. Blockchain can solve a variety of healthcare problems, including

ensuring the integrity of medical records, tracing AI training, and improving interoperability. As a result of Blockchain and AI, medical imaging applications can become more scalable and reliable, as sensitive patient data can be secured, regulatory compliance can be ensured, and collaboration between institutions can be facilitated in the development of AI models [6],[7],[8],[9]. Blockchain and artificial intelligence might transform medical imaging. A blockchain-based system ensures the sharing of imaging datasets among institutions while protecting the privacy and ownership of patient data. As a result, there may also be fewer data silos and biases in AI training since it creates decentralized training settings [10]. By utilizing AI-driven diagnostic models, professionals, patients, and regulators can benefit by enhancing trust and understanding.

Medical imaging and diagnosis could be improved by Blockchain and AI, as discussed in this article. As a first step, it addresses the challenges associated with standard medical imaging workflows as well as the limitations of independent AI solutions. This section examines blockchain technology's potential to resolve these issues with a focus on the safe sharing of data, the validation of AI models, and compliance with regulatory requirements. Author [1], cutting-edge blockchain-driven artificial intelligence implementations and research are discussed in relation to medical imaging and diagnostic imaging. In addition to its technological effects, this integration has a social impact. As blockchain and artificial intelligence become more widespread, sophisticated diagnostic technologies can become more affordable, especially in impoverished areas. Additionally, they improve the accuracy, efficiency, equity, and patient-centeredness of medical imaging.

Medical imaging and diagnostics may undergo a revolution thanks to Blockchain and artificial intelligence. Through the implementation of these technologies, long-standing problems may be solved, and new opportunities may be created to improve the health ecosystem's resilience, security, and efficiency. Further research into its implications and applications will be encouraged by this study, as well as illuminating this intriguing junction. Diagnoses and treatment plans are aided by medical imaging in healthcare. Increasingly sophisticated imaging modalities and the desire to improve diagnostic accuracy have led to the implementation of AI in this field as an important facilitator of innovation. The use of artificial intelligence to analyse images, detect anomalies, and predict outcomes leads to improved diagnostic efficiency and accuracy [11]. The security,

trustworthiness, and scalability of medical imaging limit its potential.

AI for medical imaging is challenging because healthcare data is delicate. Medical imaging datasets usually contain personally identifiable information, which is subject to strict privacy laws and compliance with HIPAA and GDPR [12]. Data breaches in centralized data management systems raise privacy and security concerns. Additionally, the quality and reliability of AI algorithms are strongly influenced by diversified and high-quality datasets. In the absence of transparent data infrastructures and institutional resistance to sharing private information, AI systems cannot function optimally. Having a model that is easily explainable and auditable is another critical issue. AI decision-making should be transparent so that clinicians and regulators can foster trust and accountability. In traditional AI systems, decision processes tend to be hidden within black boxes, making tracking them difficult [13], [14]. Regulatory compliance can be complicated when there is a lack of transparency in high-stakes healthcare applications. Medical imaging could benefit from blockchain technology, but its use with AI is still being explored. With Blockchain's decentralized, immutable ledger structure, patients' data can be exchanged safely, and AI models can be trained collaboratively without compromising their privacy. An AI algorithm's traceability and dependability can be improved using Blockchain by tamper-proofing training data and model validation. There are few academic publications and real-world applications on blockchain-driven artificial intelligence in medical imaging that address implementation, scalability, and interoperability.

2 LITERATURE REVIEW

There are a number of traditional components to the healthcare system, including doctors, patients, laboratories, hospitals, electronic health records, and patient information systems. Additionally, they can be used for monitoring, diagnosing, treating, conducting medical research, making decisions, and managing hospitals. The above healthcare services require key-enabling technologies to be controlled and managed.

2.1 Internet of Things-Based Smart Healthcare Systems

Additionally, it is necessary to manage the physical assets of the health care system so that their location

and number can be tracked in the event of an emergency, including oxygen cylinders, nebulizers, wheelchairs, oxygen meters, heart monitors, personal protective equipment, and surgical tools. [15], [16]. Using an IoT-based healthcare asset management application, these assets can be tracked quickly. Their discussion includes, for instance, how advanced medical facilities are required in a short amount of time to treat COVID-19 [17], [18]. A cloud-based data centre with hybrid communication, colossal medical screening, and hybrid communication capabilities is proposed to assist hospitals, rescue teams, and first aid units in healthcare monitoring. An edge-enabled IoT healthcare management system for monitoring, diagnosing, and managing patient data was presented by Author [19]. This proposal proposes a database-based system to manage and monitor medical tasks. A breach of the system is possible because of its centralized storage.

2.2 Smart Healthcare Systems Using Artificial Intelligence

Healthcare companies use intelligent and predictive services to provide precise, effective care and to streamline administrative tasks [20]. The healthcare industry generates and transmits a massive amount of data through IoT devices. Considering how much data is generated in the healthcare sector, AI technology could improve data management and data improvisation. Healthcare systems using artificial intelligence offer numerous benefits over those relying on cumbersome analytics and decision-making. It also helps clinicians make decisions based on medical data (training data) to provide valuable insights into medical diagnosis, treatment, and clinical decision support. According to [21], osteoporosis (dead bones) is usually detected by X-rays and magnetic resonance imaging (MRI). By incorporating AI-based features, the authors optimized the classification of osteoporosis patients based on ultrasound data.

Using artificial intelligence, a healthcare author [22] predicts heart attacks and cancer, as well as finds intelligent patterns in collected data. They developed a lightweight, secure and AI-based healthcare ecosystem communication scheme. Simulations were conducted to evaluate the scheme's performance, including end-to-end delays and throughput. As a result of their computation time comparison, decision trees are faster than other algorithms, and SVM is more accurate than existing algorithms [23]. An AI algorithm has been applied to sustainable development by the author [24].

Researchers have identified a number of scenarios in which AI-based question recommender systems may be useful. The study authors focused exclusively on enhancing AI-based question recommendations for both security screenings and financial services, as well as their application in healthcare. There has been much attention given to federated learning in healthcare recently for data offloading and privacy protection. The Author [25] proposed a framework based on federated learning to monitor healthcare data locally. As a result of their framework, medical professionals are also able to detect skin diseases more effectively. In addition to protecting patients' privacy, the researchers' proposed framework reduces operators' costs [26].

The author [27] also reviewed the problem of training machine learning models on the diverse data of healthcare. In this study, message queueing telemetry transport (MQTT) and decentralized algorithms were employed for brain tumour segmentation, which is referred to as federated learning. According to their results, the proposed framework is more accurate and latency-efficient in the regular operation of healthcare systems.

2.3 A Blockchain-Based Healthcare System

The author developed a system that prevents security breaches in electronic healthcare records by utilizing blockchain technology [28]. Security was provided by swarm intelligence and private blockchain technology for the IoT network. The recent COVID-19 pandemic has provided an opportunity to use blockchain technology to solve diabetes problems, according to Author [29]. As a result of the pandemic, diabetes patients were disproportionately affected by a constant monitoring system, oxygen, insulin pumps, and medical beds powered by blockchain technology. An interplanetary file system, smart contracts, and the new economy movement were used to develop a proof-of-concept model encrypting and authenticating patient data. It is impressive how fast, cheap, and efficient their system is in terms of transaction speed and fees, as well as power consumption [30].

2.4 Healthcare Technology Based on Cloud Computing

Healthcare industries are increasingly collaborating and connecting through cloud computing, edge computing, and fog computing in the Cloud [31]. The implementation of smart healthcare ecosystems is

facing a variety of challenges, such as high operational costs, privacy concerns, and centralized data storage issues. In healthcare workflows, cloud technology promotes accessibility, faster response times, better personalised care, and load balancing. Using fog and edge computing techniques, wearable devices and doctors can communicate more quickly. A wearable device that cannot communicate rapidly with medical practitioners can endanger the lives of ICU patients. Fog and edge computing enable medical devices to process their data locally, as opposed to transmitting it from a global model, making remote monitoring and patient engagement easier.

3 METHODOLOGY

The blockchain system ensures the safety and security of hospitals' EHR data by increasing their data security levels. At the Central Server, the data is stored on a blockchain, increasing its security. Using blockchain technology, data is stored in a highly secure manner thanks to a hash algorithm linking each block of data to another and a consensus mechanism connecting distributed peer-to-peer networks.

Using hashes of the previous blocks in a blockchain, it is possible to link all data blocks together in order of arrival to form a chain of blocks. Whenever data within a block is changed, the hash value will change as well. There will be an immediate indication of changes, and the original block will be retrieved as well. By hashing, changes can be flagged. A blockchain consists of several nodes whose hash values propagate from one block to another.

Blockchains are built on Ethereum platforms, as mentioned previously [14]. A Proof-of-Stake consensus mechanism has replaced Proof-of-Work, where mining nodes had to solve cryptographic puzzles to create a new block of currency. Consequently, processing costs increased due to the computational requirements for solving this problem. While Proof-of-Stake does not require solving puzzles, it is quicker and less expensive.

Using the web3.js library, we developed the blockchain interface API to connect with the Ethereum blockchain. In this library, you will find JavaScript functions related to multi-blockchains.

3.1 The Data Layer

Consensus blockchains record and share data via their block producer nodes. Blockchain technology

prevents data from being altered or changed once it is recorded. Authentication and interaction with the data layer are handled by Data Sharing Modules, which enable unified data, improved authorization mechanisms, and more efficient access controls. A data layer can be accessed and interacted with by readers and writers using the application layer. Messages will be posted and received by the application layer, and the Data Sharing Module will create blockchain-compatible data structures. The Key Management Service (KMS) of Amazon Web Services manages and creates encryption keys for data storage. There is no way to expose any part of the KMS environment to the Master Key that encrypts Data Keys, which are encrypted and stored in the Blockchain.

3.2 Smart Contact

In the Smart Contract, 10 patient records are sent to mobile devices. In addition to the patient's personal information, there is information about the visit, information about laboratory tests, and information about drug orders. Solidity is the language we use to develop Ethereum's smart contracts. Patient information for a specific PID can be accessed using the mapping object, which stores patient information as a dictionary. When creating patient information, the system compares the Hash for a new visit to the Hash for the patient's most recent existing visit. The system will not be able to read new visitor information if that hash value repeats.

3.3 Application Programming Interface

ABlockchains are connected and interacted with through an Application Programming Interface (API). In the mobile application, API functions can be used to read and write patient data. With Node.js APIs, Amazon's Key Management System (AWS KMS) and Simple Storage Service (S3) are integrated, and two keys are encrypted: search hashes and encryption/decryption keys. Data that is encrypted will remain encrypted for as long as it exists without a means to decrypt it unless explicitly authorized.

A master key can be generated through the Amazon Key Management System (Amazon KMS). They are stored on an Amazon server with a very high level of security. Our application will use these keys to encrypt and decrypt data keys. In the blockchain system, patients' data is searched using a hashing key,

and API Servers and mobile applications encrypt and decrypt data using an encryption/decryption key.

3.3.1 Initialization

The initialization process involves the registration of three entities (Cloud, DO, and DU) and the publication of resources.

3.3.1.1 Registration of Cloud

Cloud, DO, and DU must register in Blockchain before becoming legitimate users or nodes. KGen, AGen, SynData, and ISave, each of which is implemented by Blockchain, are introduced before designing the registration workflow. Input null is passed to KGen(), and output KW is output; input. K_{pub} is passed to AGen(), and output Addr is output. SynData() synchronizes data from time. t_i and t_j . Inputs t_i and t_j , outputs blockdata_{ij}. It is an interface for blockchain storage that accepts addresses, signatures, private keys, storage contents, and outputs trans.

3.3.1.2 Registration of User

A two-step registration workflow is designed for DO and DU. Since DO and DU share the same registration process, we collectively refer to it as user registration.

- User->Blockchain: $URequest || t_l || t_m$. A user registers with Blockchain.
- Blockchain->User:
 $E(ks_2, Addr_{user} || K_{user_w}) || E(K_{pubU}, ks_2) || blockdata_{lm}$
 K_{user_w} is generated by Blockchain, and $AGen(L_{pub_{user_w}})$ creates $Addr_{user}$. Then, $E(ks_2, Addr_{user} || K_{user_w}) || E(K_{pubU}, ks_2) || blockdata_{lm}$ is sent to the user. $Addr_{user} || K_{user_w}$ represents the data generated by Blockchain from t_l to t_m , which is decrypted by the $D(K_{priU}, ks_2)$ and $D(ks_2, Addr_{user} || K_{user_w})$. To synchronize $blockdata_{lm}$ with the local database, the user normally calls $SynData(t_l, t_m)$.

3.3.1.3 Resource Publishing

A resource publisher uploads resources to a cloud and publishes their metadata on a blockchain. The first step in designing a registration workflow is to introduce the functions $ResUp()$ and $ISend()$ in the Blockchain. Uploading resources to the Cloud is

performed by DO, which inputs and outputs $resUpURL$ and $resInfo$; the $ISend()$ function initiates blockchain transactions by inputting. $K_{pri_{from}}$, $Addr_{from}$, $Addr_{to}$, index, content, and timestamp, and outputting $tran_{ID}$. In line 2, the smart contract $Tx_{regresource}$ configures the Blockchain, lines 3-8 encrypt access control permissions, lines 9 and 10 assign $resInfo$, line 11 calls $ISend$, and line 12 returns $tran_{ID}$.

3.3.2 Access Control

In cloud access control, users request cloud resources, and the Cloud determines whether users can access the resources based on permissions stored in Blockchain. Users with rights can access resources in the Cloud, and Blockchain records the access. We introduce a function called $VerifyCap$ and an interface called $IQuery$ before designing access control workflows. With $VerifyCap()$, realized by Cloud, two permission sets are checked for inclusion. If $resCAP_1$ contains $resCAP_{s_2}$, returns true; otherwise, returns false. It takes input, K_{priC} , and $resCAP_{s_2}$, and outputs true or false.

A transaction query interface is $IQuery()$, which is realized by Blockchain, which inputs index, $Addr_1$ and $Addr_2$, where index represents the query index value, $Addr_1$ and $Addr_2$, which represents trading parties and outputs the train.

- Cloud->User: access information for responses. A resource can only be accessed by a User if the Response access Info is true; otherwise, the Resource cannot be accessed.
- Cloud->Blockchain:
 $Txlog_{record}(K_{pri_{cloud_w}}, Addr_{cloud}, Addr_{user}, Addr_{DO}, K_{pubC}, resInfo, timestamp)$
 . An access log is published in the Blockchain using $Tx_{log_{record}}$, a smart contract created by the Cloud.

3.3.3 Authorization

Direct and indirect authorizations are divided into two categories. DUs are authorized directly by the DO, whereas indirect authorizations are authorized indirectly by the granted DU:

3.3.3.1 Direct Authorization

Owners of resources directly authorize others to access their resources. Publishing resources requires direct authorization workflows.

- DU1->Cloud:
 $E(ks_5, auth_{flag} || Addr_{U1} || Addr_{DO} || resInfo)$
 $|| E(K_{pubC}, ks_5)$. As part of the authorization request, DU1 sends an authorization flag to Cloud, $auth_{flag}$, which is the authorization key. $Addr_{U1}, Addr_{DO}$ resInfo are obtained from cloud decryption.
- Cloud->Blockchain: $hash_{resID} || Addr_{DO}$. Cloud computes Hash(resID) to get $hash_{resID}$, and calls QFCapChain ($Addr_{DO}, Addr_{cloud}, hash_{resID}$).
- Blockchain->Cloud: $resCAP_s$. Cloud receives $resCAP_s$ from Blockchain
- Cloud:
 $erifyCap(resInfo.resCAP, K_{priC}, resCAP_s)$
 . To ensure that DU1's permission requests are within the scope of the host's authorization, Cloud checks that $Addr_{DO}$ is the host of the Resource.
- Requests that fall outside the authorization scope are immediately terminated.
- Cloud->DO:
 $E(ks_6, auth_{flag} || Addr_{U1} || resInfo) || E(K_{pubDO},$
 . A request to authorize DU1 is sent from the Cloud to the DO.
- DO->Blockchain:
 $Tx_{publish}(K_{pri_{DOw}}, Addr_{DO}, Addr_{U1}, K_{pubC}, resInfo, \dots)$
 Using $Tx_{publish}$, DO publishes authorization to the Blockchain. $Tx_{publish}$ is a smart contract,
- DO->DU1: h_{flag} . DU1 receives $auth_{flag}$ from DO.

3.3.3.2 Indirect Authorization

DUs with indirect authorization provide authorization to at least five entities: DOs (resource owners), DU1s (authorizers), DU2s (requesters), Blockchains, and cloud-based technologies. Processes specific to interaction.

4 RESULTS AND DISCUSSION

A comparison of packets sent versus rounds for four models is shown in Figure 1: Medrec, Medchain, Medblock, and the Proposed Model. Each model sends more packets as rounds increase, but the Proposed Model consistently sends fewer packets, showing superior efficiency. The most packets are sent by Medrec, followed by Medchain and Medblock. As a result of the Proposed Model, the

communication load is reduced, resulting in enhanced data transmission efficiency.

A comparison between the proposed framework and benchmark models is illustrated in Figure 2. Round and transaction numbers are analyzed during the analysis. In comparison with the benchmark models, the proposed method achieved a greater number of transactions. In the proposed approach, smart contracts are used to minimize delays and allow for more transactions at the sink, which leads to an improvement in transaction volume.

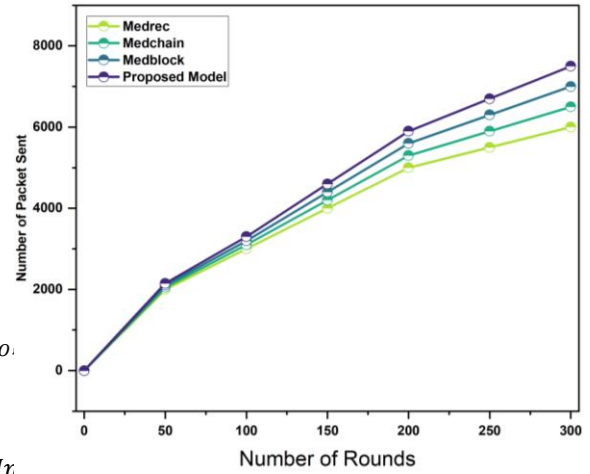


Figure 1: Communication efficiency: packets sent versus rounds.

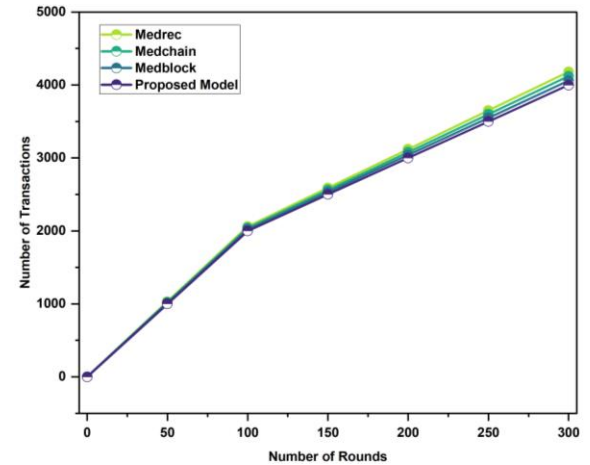


Figure 2: Number of transactions per round.

According to Figure 3, dead nodes are compared to rounds in terms of the number of dead nodes. Based on simulation results, the proposed method generates fewer dead nodes than Medrec, Medchain, and Medblock.

A summary of the simulation results, as well as the number of live nodes, can be found in Figure 4. The maximum number of live nodes observed after 300 rounds was 4000. A comparison of the proposed method with the benchmark models clearly shows that it maintains a higher number of live nodes for a given number of rounds. In this way, our framework demonstrates its superior efficiency.

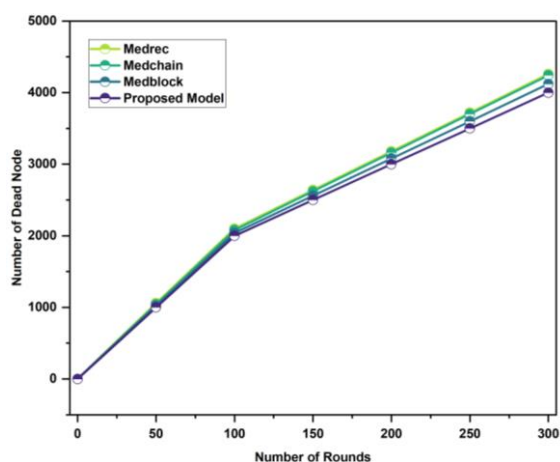


Figure 3: Number of dead nodes versus rounds.

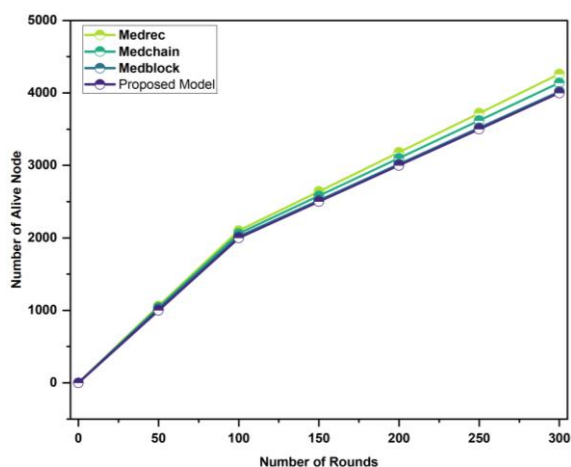


Figure 4: Number of live nodes versus rounds.

5 CONCLUSIONS

Using Blockchain technology, we developed a platform for securing, integrating, and presenting data from IoT devices. It was found that the proposed system registered devices and transmitted sensor data faster than conventional systems, based on performance evaluations. The study illustrates how medical imaging applications can be significantly

enhanced through secure, auditable, and transparent data-sharing practices, thus increasing stakeholder trust and substantially enhancing the scalability and reliability of AI-driven systems. By effectively combining blockchain's decentralized security features with advanced artificial intelligence techniques, diagnostic accuracy can be improved, patient care and clinical decision-making processes can be optimized, and strict regulatory compliance in medical imaging environments can be consistently maintained. This integrated approach also facilitates smoother collaboration among medical professionals, encourages patient engagement by ensuring data ownership and consent management, and promotes innovation in personalized healthcare solutions through trustworthy and interoperable data exchanges.

REFERENCES

- [1] P. Rani, S. Verma, S. P. Yadav, B. K. Rai, M. S. Naruka, and D. Kumar, "Simulation of the lightweight blockchain technique based on privacy and security for healthcare data for the cloud system," *Int. J. E-Health Med. Commun. (IJEHMC)*, vol. 13, no. 4, pp. 1–15, 2022.
- [2] A. Singh et al., "Blockchain-Based Lightweight Authentication Protocol for Next-Generation Trustworthy Internet of Vehicles Communication," *IEEE Trans. Consum. Electron.*, vol. 70, no. 2, pp. 4898–4907, May 2024. [Online]. Available: <https://doi.org/10.1109/TCE.2024.3351221>.
- [3] K. Devarapu, K. Rahman, A. Kamisetty, and D. Narsina, "MLOps-Driven Solutions for Real-Time Monitoring of Obesity and Its Impact on Heart Disease Risk: Enhancing Predictive Accuracy in Healthcare," *Int. J. Reciprocal Symmetry Theor. Phys.*, vol. 6, pp. 43–55, 2019.
- [4] C. R. Thompson, R. R. Talla, J. C. S. Gummadi, and A. Kamisetty, "Reinforcement Learning Techniques for Autonomous Robotics," *Asian J. Appl. Sci. Eng.*, vol. 8, no. 1, pp. 85–96, 2019.
- [5] P. K. Gade, "MLOps Pipelines for GenAI in Renewable Energy: Enhancing Environmental Efficiency and Innovation," *Asia Pac. J. Energy Environ.*, vol. 6, no. 2, pp. 113–122, Dec. 2019. [Online]. Available: <https://doi.org/10.18034/apjee.v6i2.776>.
- [6] R. K. Karanam et al., "Neural Networks in Algorithmic Trading for Financial Markets," *Asian Account. Audit. Adv.*, vol. 9, no. 1, pp. 115–126, 2018.
- [7] R. Mohammed et al., "Optimizing Web Performance: Front End Development Strategies for the Aviation Sector," *Int. J. Reciprocal Symmetry Theor. Phys.*, vol. 4, pp. 38–45, 2017.
- [8] D. Narsina et al., "AI-Driven Database Systems in FinTech: Enhancing Fraud Detection and Transaction Efficiency," *Asian Account. Audit. Adv.*, vol. 10, no. 1, pp. 81–92, 2019.

- [9] M. Rodriguez et al., "Oracle EBS and Digital Transformation: Aligning Technology with Business Goals," *Technol. Manag. Rev.*, vol. 4, pp. 49–63, 2019.
- [10] P. Rani, K. Ur Rehman, S. P. Yadav, and L. Hussein, "Deep Learning and AI in Behavioral Analysis for Revolutionizing Mental Healthcare," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra et al., Eds., IGI Global, 2025, pp. 263–282. [Online]. Available: <https://doi.org/10.4018/979-8-3693-4203-9.ch014>.
- [11] H. P. Kommineni, "Cognitive Edge Computing: Machine Learning Strategies for IoT Data Management," *Asian J. Appl. Sci. Eng.*, vol. 8, no. 1, pp. 97–108, Oct. 2019. [Online]. Available: <https://doi.org/10.18034/ajase.v8i1.123>.
- [12] S. Kothapalli et al., "Code Refactoring Strategies for DevOps: Improving Software Maintainability and Scalability," *ABC Res. Alert*, vol. 7, no. 3, pp. 193–204, Dec. 2019. [Online]. Available: <https://doi.org/10.18034/ra.v7i3.663>.
- [13] R. R. Kundavaram et al., "Predictive Analytics and Generative AI for Optimizing Cervical and Breast Cancer Outcomes: A Data-Centric Approach," *ABC Res. Alert*, vol. 6, no. 3, pp. 214–223, Dec. 2018. [Online]. Available: <https://doi.org/10.18034/ra.v6i3.672>.
- [14] P. Rani, D. S. Mohan, S. P. Yadav, G. K. Rajput, and M. A. Farouni, "Sentiment Analysis and Emotional Recognition: Enhancing Therapeutic Interventions," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra et al., Eds., IGI Global, 2025, pp. 283–302. [Online]. Available: <https://doi.org/10.4018/979-8-3693-4203-9.ch015>.
- [15] E. Mezghani, E. Exposito, and K. Drira, "A Model-Driven Methodology for the Design of Autonomic and Cognitive IoT-Based Systems: Application to Healthcare," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 1, no. 3, pp. 224–234, Jun. 2017. [Online]. Available: <https://doi.org/10.1109/TETCI.2017.2699218>.
- [16] M. N. Kadhim, A. H. Mutlag, D. A. Hammood, and N. B. H. Ismail, "Identification of Vehicle Logos in Deep Learning: A Comprehensive Survey," *JT*, vol. 7, no. 1, pp. 37–47, Mar. 2025.
- [17] G. Tomasicchio, A. Ceccarelli, A. D. Matteis, and L. Spazzacampagna, "A space-based healthcare emergency management system for epidemics monitoring and response," in *IET Conf. Proc.*, vol. 2021, no. 14, pp. 195–199, Apr. 2022. [Online]. Available: <https://doi.org/10.1049/icp.2022.0571>.
- [18] S. Verma et al., "An automated face mask detection system using transfer learning based neural network to preventing viral infection," *Expert Syst.*, p. e13507, 2024.
- [19] A. F. Subahi, "Edge-Based IoT Medical Record System: Requirements, Recommendations and Conceptual Design," *IEEE Access*, vol. 7, pp. 94150–94159, 2019. [Online]. Available: <https://doi.org/10.1109/ACCESS.2019.2927958>.
- [20] M. U. Rehman et al., "A Novel Chaos-Based Privacy-Preserving Deep Learning Model for Cancer Diagnosis," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 6, pp. 4322–4337, Nov. 2022. [Online]. Available: <https://doi.org/10.1109/TNSE.2022.3199235>.
- [21] D. Miranda, R. Olivares, R. Munoz, and J.-G. Minonzio, "Improvement of Patient Classification Using Feature Selection Applied to Bidirectional Axial Transmission," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 69, no. 9, pp. 2663–2671, Sep. 2022. [Online]. Available: <https://doi.org/10.1109/TUFFC.2022.3195477>.
- [22] M. Wazid, J. Singh, A. K. Das, S. Shetty, M. K. Khan, and J. J. P. C. Rodrigues, "ASCP-IoMT: AI-Enabled Lightweight Secure Communication Protocol for Internet of Medical Things," *IEEE Access*, vol. 10, pp. 57990–58004, 2022. [Online]. Available: <https://doi.org/10.1109/ACCESS.2022.3179418>.
- [23] G. Ansari, P. Rani, and V. Kumar, "A novel technique of mixed gas identification based on the group method of data handling (GMDH) on time-dependent MOX gas sensor data," in *Proc. Int. Conf. Recent Trends Comput. (ICRTC 2022)*, Springer, 2023, pp. 641–654.
- [24] C. M. Parra, M. Gupta, and D. Dennehy, "Likelihood of Questioning AI-Based Recommendations Due to Perceived Racial/Gender Bias," *IEEE Trans. Technol. Soc.*, vol. 3, no. 1, pp. 41–45, Mar. 2022. [Online]. Available: <https://doi.org/10.1109/TTS.2021.3120303>.
- [25] H. Elayan, M. Aloqaily, and M. Guizani, "Sustainability of Healthcare Data Analysis IoT-Based Systems Using Deep Federated Learning," *IEEE Internet Things J.*, vol. 9, no. 10, pp. 7338–7346, May 2022. [Online]. Available: <https://doi.org/10.1109/JIOT.2021.3103635>.
- [26] P. Rani, S. P. Yadav, P. N. Singh, and M. Almusawi, "Real-World Case Studies: Transforming Mental Healthcare With Natural Language Processing," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra et al., Eds., IGI Global, 2025, pp. 303–324. [Online]. Available: <https://doi.org/10.4018/979-8-3693-4203-9.ch016>.
- [27] B. C. Tedeschini et al., "Decentralized Federated Learning for Healthcare Networks: A Case Study on Tumor Segmentation," *IEEE Access*, vol. 10, pp. 8693–8708, 2022. [Online]. Available: <https://doi.org/10.1109/ACCESS.2022.3141913>.
- [28] N. A. Kadhim, A. A. Obed, A. J. Abid, A. L. Saleh, and R. J. Hassoon, "A Systematic Review for Reconfiguring Photovoltaic Arrays under Conditions of Partial Shading," *EETJ*, vol. 1, no. 1, pp. 20–34, Jun. 2024.
- [29] G. Subramanian and A. Sreekantan Thampy, "Implementation of Blockchain Consortium to Prioritize Diabetes Patients' Healthcare in Pandemic Situations," *IEEE Access*, vol. 9, pp. 162459–162475, 2021. [Online]. Available: <https://doi.org/10.1109/ACCESS.2021.3132302>.
- [30] P. Rani, U. C. Garjola, and H. Abbas, "A Predictive IoT and Cloud Framework for Smart Healthcare Monitoring Using Integrated Deep Learning Model," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 53–65, 2024.
- [31] V. K. Prasad, M. D. Bhavsar, and S. Tanwar, "Influence of Monitoring: Fog and Edge Computing," *Scalable Comput. Pract. Exp.*, vol. 20, no. 2, pp. 365–376, May 2019. [Online]. Available: <https://doi.org/10.12694/scpe.v20i2.1533>.

SECTION 3

CONTROL AND AUTOMATION

Proceedings of the 13th International Conference on Applied Innovations in IT (ICAIIIT)

JULY 2025

An Efficient Intrusion Detection System for IoT Using Feature Engineering and Machine Learning

Basima Kzar Hassan¹, Khawla Rashige Hassen² and Sakna Jahiya Faraj³

¹Department of Economic Studies, Center for Basra and Arab Gulf Studies, University of Basrah, 61004 Basra, Iraq

²Department of Field Crops, College of Agriculture, University of Basrah, 61004 Basra, Iraq

³Department of Economics, College of Economics and Administration, University of Basrah, 61004 Basra, Iraq
basima.hasan@uobasrah.edu.iq, khawla.hassan@uobasrah.edu.iq, sakna.al-sary@uobasrah.edu.iq

Keywords: Intrusion Detection System (IDS), Internet of Things (IoT), Machine Learning, Feature Engineering, SMOTE.

Abstract: IoT devices have limited computing and security capabilities, making them vulnerable to cyberattacks. This rapid expansion of IoT has introduced unprecedented connectivity but also heightened security vulnerabilities. The security of IoT environments, therefore, depends on efficient and lightweight intrusion detection systems (IDS). Using advanced feature engineering and machine learning algorithms, this study develops high-performance IDS designed for IoT networks. Data imbalance is addressed with preprocessing techniques, feature extraction, and synthetic minority oversampling techniques (SMOTE). Multi-dataset training and testing included K-nearest neighbour models, sequential minimalism optimization models, random forest models, and stacking ensembles. A transfer learning model such as VGG-16 and DenseNet was also incorporated to improve classification accuracy. It has been demonstrated that the proposed models, particularly the ensemble and RF-based approaches, are highly accurate, precise, and recallable. IoT environments with limited resources can benefit from the proposed IDS framework because it effectively identifies malicious traffic while maintaining computational efficiency.

1 INTRODUCTION

Internet-connected devices are called IoT devices. Things such as wireless sensors, smart cameras, and smart televisions are being connected to the IoT, which has undergone rapid changes [1], [2]. IoT devices are rapidly spreading across the Internet, with more than 2 billion connected in 2017. Data generated by Internet-of-things devices is projected to reach 73.1 zettabytes by 2025, according to experts [3]. As the IoT grows in popularity and offers many benefits, their security can often be limited even though IoT generates vast amounts of data and is growing at an unprecedented rate. IoT has been rapidly adopted due to the huge amount of data generated by billions of connected devices. Even as IoT devices become increasingly popular, they lack strong security protections.

IoT devices lack sufficient security capabilities, so it's essential to build NIDS that can detect and prevent attacks on IoT networks quickly and reliably [4]. As well as machine learning techniques, publicly available data on network traffic can also be used for intrusion detection in IoT [5], [6]. Machine

learning models tend to be more complex and accurate when these datasets contain fewer redundant or irrelevant features [7]. Machine learning models are often developed through feature reduction in order to develop efficient NIDS. In addition to reducing computational costs and latency, model generalization is also increased.

The problem of excessive features can be solved using several techniques, including feature selection and feature extraction. Feature selection selects the most informative features from an initial set of features [8]. The selected features remain semantically interpretable while the dimensionality is reduced. By mathematically projecting the original features into a low-dimensional space, feature extraction is performed [9], [10]. There is no intuitive meaning to the extracted features, which may reduce dimensionality, but they may reduce complexity. In the IoT, lightweight and efficient IDS can be easily created by carefully selecting original features that are relevant to the device in question. As an alternative, a feature extraction technique reduces the overall dimensions of the data while preserving critical information. The optimization of the

efficiency and interpretability of intrusion detection models, which are tailored to the limitations and complexity of IoT devices and networks, can help IoT ecosystems improve their cybersecurity posture.

Many sectors are experiencing an increase in cyber threats, such as IoT, online banking, industrial systems, and healthcare. A smart home, a smart city, and even wearables are all made possible by the IoT. There are, however, some limitations to IoT devices operating over public networks, including limited computational power, storage, and bandwidth [11]. Since these limitations limit their attack potential, they are more vulnerable to attacks than traditional endpoint devices. There is room for improvement in intrusion detection techniques, especially when it comes to improving their accuracy and adaptability, regardless of the number of intrusion detection techniques proposed in the literature.

Here are some contributions to the ongoing study that address these challenges:

- A literature review is presented on the use of machine learning and deep learning for intrusion detection based on numerical and image-based datasets.
- To correct for class imbalances, datasets are preprocessed and balanced using the SMOTE.
- The combination of multiple feature extraction techniques with stacked machine learning methods, including KNN, SMO, and RF, is used to identify malicious or benign network traffic.
- Performing experiments to validate the proposed models' performance and reliability.

2 LITERATURE REVIEW

IoT architectures refer to mechanisms for interconnecting addressable electronic gadgets using radio and telecommunication infrastructures in order to establish interconnection [12]. The communication infrastructure additionally includes living life, such as people, animals, and plants, where it is used for monitoring, improving the quality of life, and reducing resource utilization [13]. WSNs anchor IoT. In addition to collecting information, WSNs do a number of other things [14]. Currently, IoT security frameworks and sensor centres are lacking, and gadgets are not protected against attacks such as Denial of Service (DoS), Man-in-the-middle attacks, etc [15], [16]. IoT verification and access control conventions are currently undergoing a lot of activity, but additional mechanisms are needed according to requirements. Experts have expressed concerns about IoT, its network scheme, and safety concerns [17].

Within the network design architecture, IoT requires consistent data handling and communication.

2.1 Machine Learning-Based Intrusion Detection Systems

In this study, we attempted to determine what set of hyperparameters would be most appropriate for use in NIDS. Data from UNSW-NB15 was analyzed using DFF and LSTM architectures. As compared to DFF, LSTM performed slightly better, but the relu activation function outperformed them all. SGD performed less accurately than the majority of optimisers, except for the majority of them. In their opinion, the best settings for the hyperparameters were relu, adam, and nodes configuring input and output rules at 0.75. DFF had the best accuracy at 98.8%, while LSTM had the best accuracy at 98%. This paper does not drop the flow identifier features, nor does it evaluate their best-claimed set of hyperparameters on a different dataset. As an FE tool, [18] proposed an LSTM-based AE neural architecture composed of dense layers and LSTMs. RF classifiers are applied after extraction to detect attacks. We evaluated the proposed methodology using several datasets, including UNSW-NB15, ToN-IoT, and NSL-KDD [19]. This study indicates that the selected classifier is able to detect more objects without relying on compression to do so. The reduction of dimensions has significantly reduced training time.

In the literature, researchers continue to develop new FR methods and ML models based on the negative habits discussed in [20]. Research in this domain is always able to find combinations or variations to achieve slightly better results numerically. If you are applying the algorithm to a specific dataset, you should modify all hyperparameters used. Almost all papers have used a single dataset, making it difficult to conclude that their proposed techniques are generalizable. Depending on which dataset you use, the information presented will vary. This means the results of these proposed techniques may vary depending on the dataset to which they are applied. Due to the above experimental issues, ML-based NIDS cannot be deployed in the operational arena despite extensive academic research. Despite this, machine learning tools have been applied successfully to commercial scenarios, compared with other applications. A suitable ML model is necessary before deployment because of NIDS' high error costs. Our best combination will be generalizable if we compare all datasets. Analyzing the extracted dimensions requires

calculating and comparing the variance and correlation between PCA and LDA.

Several evaluation metrics were used in the evaluation of these algorithms, including KDD CUP 99, NSL-KDD, CIDDs, and CICIDS2017. DBN, which enhanced detection accuracy from 5% to 10%, was found to be more effective in DL models than ML models. According to [21], heuristic intrusion detection can achieve an accuracy of 85.5% to 95.2%. Gradient decomposition algorithms are used to train IDSs beforehand, followed by retraining and testing using KDD20+ and KDDTest+ datasets.

By combining NSL-KDD and UNSW-15 datasets, [22] experiments with hybrid sampling-based intrusion detection. With SMOTE and OSS combined, RF, CNN, BiLSTM, CNN-BiLSTM, AlexNet-5, and CNN-BiLSTM models are cross-trained. For the datasets mentioned, CNN-BiLSTM had an accuracy of 83.58% and 77.16%, respectively, outperforming other algorithms. Using RNN and GRNN algorithms, the author investigated the effectiveness of bi-RNN intrusion detection [23]. In the evaluation, Bi-RNN was found to achieve the best accuracy with 99.04% when compared to other methods utilizing 10% KDD datasets [24].

2.2 IoT Threats

As IoT networks are distributed, their architecture is layered, with each layer performing its tasks sequentially to maintain the platform's efficiency. Researchers have found that a variety of attacks can be carried out at every layer of the network for the purpose of breaching its security, including protocols and gateways [25]. There has been a lot of recent attention paid to encapsulated attacks that occur in the whole network, such as DoS, Probe, U2R, and R2L [26].

3 METHODOLOGY

These sections describe the dataset, how it was preprocessed, how features were extracted, and how the dataset was evaluated. Modelling is primarily intended to identify regular traffic versus malicious traffic. This is accomplished by developing several models, representing them in image format, and comparing them. A set of datasets as described in the research paper "Detecting intrusions use network traffic profiling and machine learning for IoT applications using network traffic profiling". The first step is to provide the machine learning module with binary visualization data that is compressed. This

repository also provides the packet capture images from each of the five attack scenarios, along with their associated PCAP files, in attackScenario.zip and attackSenarioImages.zip.

3.1 Data Processing

To improve the training process for machine learning models, it is essential to process data first. You can download all datasets for research purposes for free. To reduce storage size and eliminate redundant samples (flows), duplicate samples (flows) are removed. Source and destination IP addresses, ports, and timestamps are stripped from flow identifiers to prevent bias against attackers or victims. A numerical value is encoded for strings, non-numeric features, and other features. In the datasets, protocol and service characteristics are stored in string form, while ML models prefer numerical information. To encode features, one can use hot encoding, while the other can use label encoding. By adding X features to a feature, it becomes X categories. A category represents 1, and absence represents 0. Due to the increased number of dimensions in the dataset, ML models may struggle with performance and efficiency. Thus, each category is assigned a numerical value according to the label encoding technique.

Our dataset consists only of numerical values, eliminating nan, dash, and infinity. A 1 is displayed when a boolean feature is true, while a 0 is displayed when a boolean feature is false. In an effort to reduce complexity, all feature values are scaled between 0 and 1 based on the min-max feature scaling method. The ML model gives equal weight to all features, but since network traffic features tend to have large values, it may give them a heavier weight since they tend to have high values. The values of every feature are calculated by plugging in a new feature value from 0 to 1, the previous feature value, X^* , and the maximum and minimum of the feature. Datasets for training and testing are separated according to the label features, which is necessary because there is a class imbalance between the two datasets:

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}}. \quad (1)$$

3.2 Synthetic Minority Oversampling Technique (SMOTE)

Using SMOTE [23], balanced datasets are generated by oversampling minority classes. Synthetic examples are included along with the neighbours

nearest to the minority class so that the minority class can be oversampled. If oversampling is necessary, the k-nearest neighbours may be selected at random. This method is illustrated in Figure 1 by Y_i representing the point in question, Y_{i1} through Y_{i4} representing nearest neighbours, and W_1 through W_4 representing the synthetic data generated by randomized interjection.

Using a feature vector (sample) based on the nearest neighbour, synthetic samples can be generated. The feature vector is calculated by multiplying the difference by a random number between 1 and 0. Using this method, two distinct characteristics are considered when selecting a random point on the line segment. The following is a detailed description of SMOTE:

- Rank samples according to their k-nearest neighbours.
- Samples are selected randomly using the KNN algorithm.
- To calculate the new samples, add the original samples to the difference and then add the gap (0,1).
- Make the minority sample larger by adding new samples. The final step is to create a new dataset.

The oversampling of minority samples by SMOTE has some weaknesses since majority samples are not taken into account when creating minority samples. As a result, minority samples may be generated around the positive examples, escalating the problem of borderline and noisy examples in learning.

3.3 Feature Extraction

3.3.1 Overview of Feature Extraction Process

When a character recognition system is preprocessed, features are extracted. An input pattern must be correctly assigned to one of the possible output classes in order for pattern recognition to succeed. In general, two general stages are involved in this process: Feature selection and classification. Poorly selected features cannot be recognized by the classifier, so feature selection is very important. According to Lippman, the following features should be selected:

Feature information should be sufficient to generate discriminant functions efficiently and be insensitive to irrelevant variations in input, as well as limited in number in order to minimize training data requirements. To build a pattern classification, feature extraction is important for determining the

relevant characteristics of each class. As part of this process, relevant features are extracted to form feature vectors using an object/alphabet. Input units and target output units can be recognized by classifiers when these feature vectors are combined. By considering these features, the classifier can differentiate between classes more easily, making it easier to classify. The process of extracting features from raw data involves retrieving the most relevant information. Identifying the parameters that define characters allows them to be accurately and uniquely defined. A feature vector represents a character's identity in a feature extraction process. When generating features for a variety of instances of a symbol, the goal is to maximize the recognition rate using the simplest elements possible.

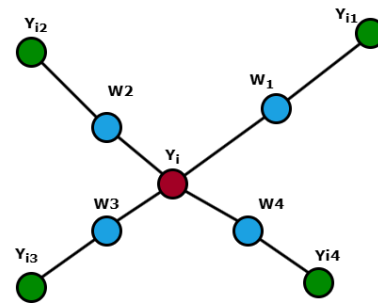


Figure 1: Generate data in the SMOTE algorithm.

3.3.2 Deep Learning-Based Feature Extraction with Pre-Trained Models

Pre-trained models are neural networks that have been trained using a lot of data. Pretrained models include the following. Models that were pre-trained and supervised. It is one of the most popular pre-training datasets for supervised learning. As part of the ImageNet classification challenge, [27] used deep neural networks to outperform humans. InceptionNet [28] is another deep neural network that uses parallel convolutional filters. ResNet [29] introduces skip connections to ease training and becomes much deeper as performance improves. DenseNet [30] densely connected blocks [31] are carefully designed to be mobile-friendly, and the structure can be further optimized through network architecture searches [32]. DenseNet, ResNet_50, and VGG-16 were used. A visibility evaluation model was constructed using the extracted features, and the variables were input into a Support Vector Regression model, followed by a visibility evaluation of the effective area:

- ImageNet was used to train the VGG-16 network [33], [34]. A large training set enables

the VGG-16 to perform well even with small image data sets. There are 16 convolution layers in the VGG-16 network, along with a 3x3 receptive field. The pooling layer, which has a size of 2x2, is one of five. Immediately following the last layer of Max pooling are three fully connected layers. There are three fully interconnected layers in the following layers. The Softmax classifier is the final layer of the algorithm. There are no hidden layers that are not activated by ReLu [35].

- The DenseNet is an artificial neural network based on dense convolutions [36]. A DenseNet architecture consists of interconnected layers. The number of layers in a network is $N(N+1)/2$. DenseNet comprises an initial layer of convolutions and a layer of pooling. Pre-trained models are neural networks that have been trained using a lot of data.

3.4 Image Filter

With an image filter (IF), high-quality images are removed from highly corrupted images by removing impulse noise from additive identical independent distributions (i.i.d.). To implement the proposed filter, fuzzy numbers need to be constructed, fuzzy filtering needs to be performed, and a genetic learning process needs to be applied. Fuzzy numbers are constructed by constructing the image knowledge base from sample images or noise-free images. Additionally, fuzzy decision processes, fuzzy means, and fuzzy inference mechanisms are all part of fuzzy filtering. The optimization of image knowledge bases through genetic algorithms is finalized by genetic learning. As a method of removing additive impulse noise from highly corrupted images, we will use the genetic learning method to tune the parameters of membership functions.

3.4.1 Auto-Color Correlogram

For optimal accuracy, a number of models were evaluated using a variety of filters, individual learning algorithms, and stacked models. Autocorrelograms were used to develop the first four models. A KNN model, an SMO model, an RF model, and other algorithms were used. Stacking KNN and SMO models was also created.

3.4.2 Fuzzy Color and Texture Histogram

Fuzzy Color and Texture Histogram (FCTH) is a quantitative histogram that includes information about colour and texture [8]. In this feature result,

three fuzzy units have been combined. Segmenting the image into blocks is done initially. Every block passes through the fuzzy units. The first unit involves extracting the fuzzy linking histogram using fuzzy rules. Based on the HSV colour space, this histogram can be viewed. A fuzzy system with three inputs generates a 10-bin histogram using twenty rules; each bin corresponds to a colour.

3.5 Performance Measures and Validation

Evaluation of classifiers is typically based on calculating fundamental performance metrics. In this study, the classifier's effectiveness was measured using key statistical values including true positives, true negatives, false positives, and false negatives. From these, important performance indicators such as accuracy, precision, recall, and the F1-score were derived to comprehensively assess the model's predictive capabilities. These metrics provide insight into the classifier's correctness, reliability, sensitivity, and balanced performance. The detailed methodology for calculating these metrics follows the framework established by [37].

4 RESULTS AND DISCUSSION

As shown in Figure 2, four models are compared for accuracy, precision, recall, and F1-score. According to this example, the K-Nearest Neighbor (KNN) model with $k = 1$ provides the best accuracy (93%), precision (91.4%) and test coverage (91%).

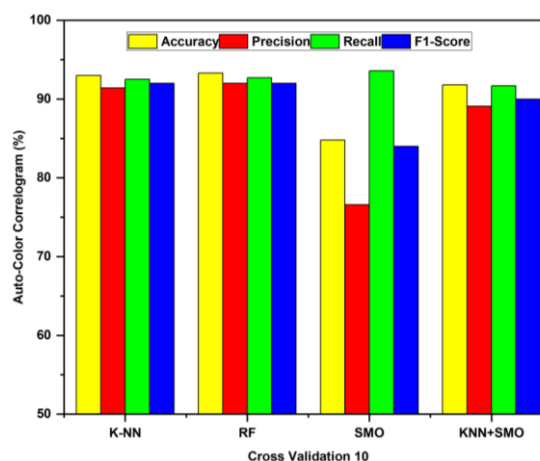


Figure 2: Performance comparison of classification models based on four methods using 10-fold cross-validation.

We evaluated five models, and Figure 3 shows their accuracy results. On the 90-10 dataset, the Random Forest (RF) model was 94.6% accurate and 94% precise, whereas the stacked model was 94.6% accurate and 94% precise. However, Random Forests outperform stacked models when it comes to recall, achieving 100% as opposed to 94.7% for stacked models. The Random Forest model remains accurate with 95% accuracy even when data is split 70/30. As shown in Figure 4, the accuracy results of each model are summarized. As a result of 90% training and 10% testing, the RF model yielded the highest accuracy and precision, respectively, of 94.22% and 91.02%.

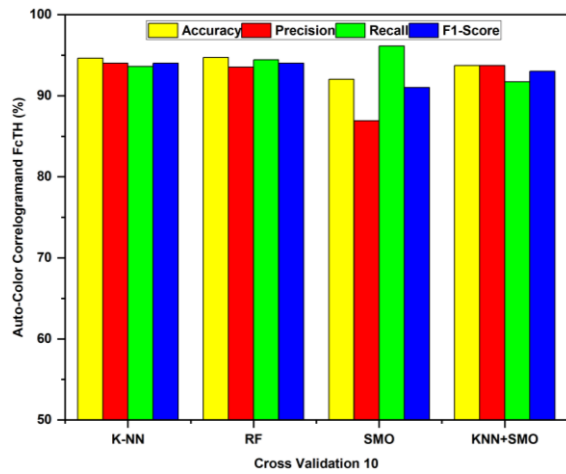


Figure 3: Performance comparison of classification models on FCTH features using 10-fold cross-validation.

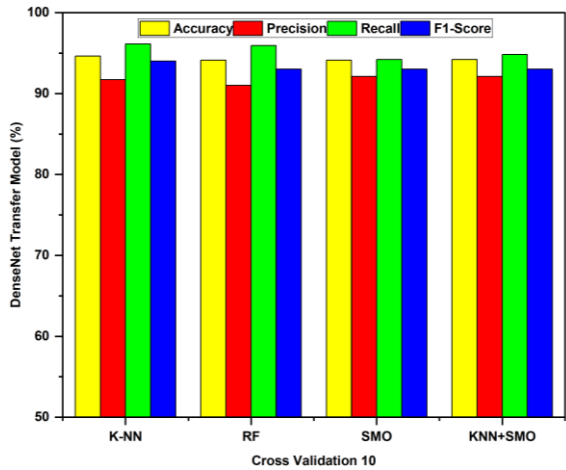


Figure 4: Comparison of four models with DenseNet transfer model.

A comparison of the four models is presented in Figure 5. When 90% of the training data are combined with 10% of the testing data, a model's accuracy and precision are 94.22% and 92.42%, respectively. With 70/30 training data, the KNN model with $k = 5$ shows 95.8% accuracy and 95.1% precision.

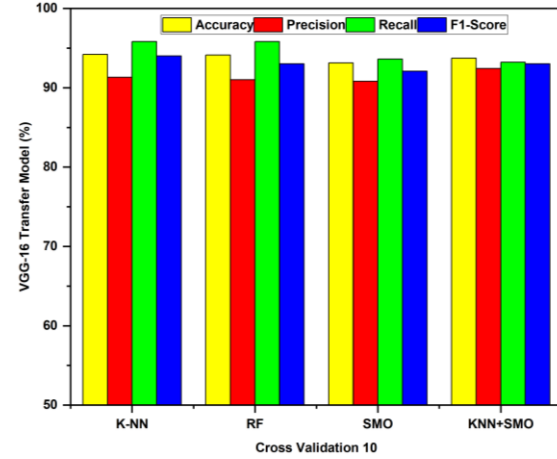


Figure 5: Comparison of four models with VGG-16 transfer model.

5 CONCLUSIONS

An IDS for the IoT is developed by integrating feature engineering techniques with traditional and advanced machine learning models. By effectively preprocessing data, performing thorough feature analysis, and utilizing SMOTE for handling class imbalance, the system significantly enhances the accuracy of intrusion detection across diverse attack scenarios in IoT networks. Multiple classifiers, such as Random Forest, K-Nearest Neighbors (KNN), Sequential Minimal Optimization (SMO), and various ensemble-based methods, exhibited commendable performance, with Random Forest and ensemble approaches consistently outperforming others in terms of accuracy, sensitivity, and precision metrics. The robustness and adaptability of the proposed system were further improved by incorporating image-based feature representations and employing advanced transfer learning techniques using pre-trained models like VGG-16 and DenseNet. These deep learning strategies effectively captured complex patterns within the data, providing superior generalization capabilities. Consequently, the proposed framework delivers a practical, reliable, and scalable solution suitable for real-time threat

detection in resource-constrained IoT environments. Future advancements in IDS applications can greatly benefit from extending this framework towards real-time deployment scenarios and integrating federated learning methods to enhance data privacy, decentralization, and collaborative learning across distributed IoT devices.

REFERENCES

- [1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015, [Online]. Available: <https://doi.org/10.1109/COMST.2015.2444095>.
- [2] B. Bhola et al., "Quality-enabled decentralized dynamic IoT platform with scalable resources integration," *IET Communications*, 2022.
- [3] W. Zhou, Y. Jia, A. Peng, Y. Zhang, and P. Liu, "The Effect of IoT New Features on Security and Privacy: New Threats, Existing Solutions, and Challenges Yet to Be Solved," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1606–1616, Apr. 2019, [Online]. Available: <https://doi.org/10.1109/JIOT.2018.2847733>.
- [4] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, "Network Intrusion Detection for IoT Security Based on Learning Techniques," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2671–2701, 2019, [Online]. Available: <https://doi.org/10.1109/COMST.2019.2896380>.
- [5] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "A Detailed Investigation and Analysis of Using Machine Learning Techniques for Intrusion Detection," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 686–728, 2019, [Online]. Available: <https://doi.org/10.1109/COMST.2018.2847722>.
- [6] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," *Computers & Electrical Engineering*, vol. 105, p. 108543, 2023.
- [7] B. A. Tama, M. Comuzzi, and K.-H. Rhee, "TSE-IDS: A Two-Stage Classifier Ensemble for Intelligent Anomaly-Based Intrusion Detection System," *IEEE Access*, vol. 7, pp. 94497–94507, 2019, [Online]. Available: <https://doi.org/10.1109/ACCESS.2019.2928048>.
- [8] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999, [Online]. Available: <https://researchcommons.waikato.ac.nz/bitstream/10289/15043/1/thesis.pdf>. [Accessed: Apr. 25, 2025].
- [9] B. Yan and G. Han, "Effective Feature Extraction via Stacked Sparse Autoencoder to Improve Intrusion Detection System," *IEEE Access*, vol. 6, pp. 41238–41248, 2018, [Online]. Available: <https://doi.org/10.1109/ACCESS.2018.2858277>.
- [10] N. Hussain, P. Rani, H. Chouhan, and U. S. Gaur, "Cyber security and privacy of connected and automated vehicles (CAVs)-based federated learning: challenges, opportunities, and open issues," *Federated Learning in IoT Applications*, pp. 169–183, 2022.
- [11] P. Rani and R. Sharma, "Intelligent Transportation System Performance Analysis of Indoor and Outdoor Internet of Vehicle (IoV) Applications Towards 5G," *Tsinghua Science and Technology*, vol. 29, no. 6, pp. 1785–1795, Dec. 2024, [Online]. Available: <https://doi.org/10.26599/TST.2023.9010119>.
- [12] A. M. Mhlaba and M. Masinde, "An Integrated Internet of Things Based System for Tracking and Monitoring Assets—the case of the Central University of Technology," in *IST-Africa 2015 Conference Proceedings*, 2015, [Online]. Available: https://www.academia.edu/download/66891365/An_Integrated_Internet_of_Things_Based_S20210504-29331-10cc0bs.pdf. [Accessed: Apr. 25, 2025].
- [13] F. T. First and I. B. Blocks, "How the Internet of Things Is Revolutionizing Healthcare," [Online]. Available: <http://euro.ecom.cmu.edu/resources/elibrary/ubi/IOT-REVHEALCARWP.pdf>. [Accessed: Apr. 25, 2025].
- [14] M. Masinde, "IoT applications that work for the African continent: Innovation or adoption?," in *2014 12th IEEE International Conference on Industrial Informatics (INDIN)*, IEEE, 2014, pp. 633–638, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6945587/>. [Accessed: Apr. 25, 2025].
- [15] R. Mahmoud, T. Yousuf, F. Aloul, and I. Zualkernan, "Internet of things (IoT) security: Current status, challenges and prospective measures," in *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*, IEEE, 2015, pp. 336–341, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/7412116/>. [Accessed: Apr. 25, 2025].
- [16] N. Hussain, P. Rani, N. Kumar, and M. G. Chaudhary, "A deep comprehensive research architecture, characteristics, challenges, issues, and benefits of routing protocol for vehicular ad-hoc networks," *International Journal of Distributed Systems and Technologies (IJ DST)*, vol. 13, no. 8, pp. 1–23, 2022.
- [17] X. Xingmei, Z. Jing, and W. He, "Research on the basic characteristics, the key technologies, the network architecture and security problems of the internet of things," in *Proceedings of 2013 3rd International Conference on Computer Science and Network Technology*, IEEE, 2013, pp. 825–828, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6967233/>. [Accessed: Apr. 25, 2025].
- [18] A. Andaliy and V. Tabataba Vakili, "A Novel Dimension Reduction Scheme for Intrusion Detection Systems in IoT Environments," *arXiv e-Prints*, p. arXiv-2007, 2020.
- [19] P. Rani et al., "Federated Learning-Based Misbehavior Detection for the 5G-Enabled Internet of Vehicles," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 2, pp. 4656–4664, May 2024, [Online]. Available: <https://doi.org/10.1109/TCE.2023.3328020>.

- [20] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in 2010 IEEE Symposium on Security and Privacy, IEEE, 2010, pp. 305–316, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/5504793/>. [Accessed: Apr. 25, 2025].
- [21] A.-H. Qureshi, H. Larijani, J. Ahmad, and N. Mtetwa, "A Heuristic Intrusion Detection System for Internet-of-Things (IoT)," in K. Arai et al. (eds), *Intelligent Computing*, in *Advances in Intelligent Systems and Computing*, vol. 997, Cham: Springer International Publishing, 2019, pp. 86–98, [Online]. Available: https://doi.org/10.1007/978-3-030-22871-2_7.
- [22] K. Jiang, W. Wang, A. Wang, and H. Wu, "Network intrusion detection combined hybrid sampling with deep hierarchical network," *IEEE Access*, vol. 8, pp. 32464–32476, 2020.
- [23] A. Dushimimana, T. Tao, R. Kindong, and A. Nishyirimbere, "Bi-directional recurrent neural network for intrusion detection system (IDS) in the internet of things (IoT)," *International Journal of Advanced Engineering Research and Science*, vol. 7, no. 3, 2020, [Online]. Available: <https://www.academia.edu/download/113495300/681JAERS-03202047-Bi-directional.pdf>. [Accessed: Apr. 25, 2025].
- [24] G. Ansari, P. Rani, and V. Kumar, "A novel technique of mixed gas identification based on the group method of data handling (GMDH) on time-dependent MOX gas sensor data," in *Proceedings of International Conference on Recent Trends in Computing: ICRTC 2022*, Springer, 2023, pp. 641–654.
- [25] F. Farhin, M. S. Kaiser, and M. Mahmud, "Secured Smart Healthcare System: Blockchain and Bayesian Inference Based Approach," in M. S. Kaiser et al. (eds), *Proceedings of International Conference on Trends in Computational and Cognitive Engineering*, in *Advances in Intelligent Systems and Computing*, vol. 1309, Singapore: Springer Singapore, 2021, pp. 455–465, [Online]. Available: https://doi.org/10.1007/978-981-33-4673-4_36.
- [26] N. Islam, I. Sultana, and M. S. Rahman, "HKMS-AMI: A Hybrid Key Management Scheme for AMI Secure Communication," in M. S. Kaiser et al. (eds), *Proceedings of International Conference on Trends in Computational and Cognitive Engineering*, in *Advances in Intelligent Systems and Computing*, vol. 1309, Singapore: Springer Singapore, 2021, pp. 383–392, [Online]. Available: https://doi.org/10.1007/978-981-33-4673-4_30.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1026–1034, [Online]. Available: http://openaccess.thecvf.com/content_iccv_2015/html/He_Delving_Deep_into_ICCV_2015_paper.html. [Accessed: Apr. 19, 2025].
- [28] C. Szegedy et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9, [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html. [Accessed: Apr. 19, 2025].
- [29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778, [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html. [Accessed: Apr. 19, 2025].
- [30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708, [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html. [Accessed: Apr. 19, 2025].
- [31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520, [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Sandler_MobileNetV2_Inverted_Residuals_CVPR_2018_paper.html. [Accessed: Apr. 19, 2025].
- [32] M. Tan et al., "Mnasnet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2820–2828, [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Tan_MnasNet_Platform-Aware_Neural_Architecture_Search_for_Mobile_CVPR_2019_paper.html. [Accessed: Apr. 19, 2025].
- [33] S. Pal, "Transfer learning and fine tuning for cross domain image classification with Keras," GitHub: *Transfer Learning Fine Tuning Cross Domain Image Classification with Keras*, 2016.
- [34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/5206848/>. [Accessed: Apr. 20, 2025].
- [35] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, and A. Agrawal, "Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection," *Construction and Building Materials*, vol. 157, pp. 322–330, 2017.
- [36] N. Subramanian, O. Elharrouss, S. Al-Maadeed, and M. Chowdhury, "A review of deep learning-based detection methods for COVID-19," *Computers in Biology and Medicine*, vol. 143, p. 105233, 2022.
- [37] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, pp. 37–63, 2011, doi: 10.9735/2229-3981.

Metaheuristic Optimization Algorithms for Deep Learning Model Design in Secure Internet of Things Environment

Mustafa Hussein Zwayyer, Sajida Allawi Dawood and Ammar Bassem Saleh

College of Islamic Science, University of Baghdad, 10071 Baghdad, Iraq

mustafa.h@cois.uobaghdad.edu.iq, Sajida.a@cois.uobaghdad.edu.iq, amar.saleh@cois.uobaghdad.edu.iq

Keywords: Botnet Detection, Internet of Things (IoT), Metaheuristic Optimization, Deep Learning, Feature Selection.

Abstract: The Internet of Things (IoT) has enabled smart systems, but it has also increased vulnerabilities to cyber threats, including botnet attacks. To address these security challenges, this study proposes a hybrid system that combines metaheuristic and machine learning. To tune hyperparameters of a hybrid neural network based on Convolutional Neural Networks and Semi-Recurrent Neural Networks (CNN-QRNN), the Chaotic Butterfly Optimization Algorithm (CBOA) is used. A new metaheuristic algorithm, Self-Adaptive Enhanced Harris Hawks Optimization (SAEHO), as well as a self-upgraded cat and mouse optimizer (SU-CMO), are introduced and evaluated in order to enhance model effectiveness. Based on experiments conducted on the N-BaIoT dataset, it was determined that the proposed models significantly outperformed conventional classifiers in key performance metrics, including accuracy, the Matthews Correlation Coefficient (MCC), the Rand Index, and the F-Measure. Particularly notable improvements were observed in reducing false-positive rates and enhancing anomaly detection sensitivity. The HMMLB-BND method substantially improves detection performance in diverse IoT environments, offering a robust, efficient, and scalable solution suitable for real-time deployment in resource-constrained systems.

1 INTRODUCTION

Through IoT, devices and systems can be connected, data can be processed in real time, and intelligent decisions can be made. With the rise of interconnected devices, security against a wide range of cyber threats becomes increasingly important. For IoT environments, optimizing deep learning models poses a crucial challenge, especially in terms of security and performance. A powerful tool for addressing this challenge has emerged: metaheuristic optimization algorithms. These algorithms provide solutions to complex optimization problems that traditional methods may not be able to solve, such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO), and Simulated Annealing (SA) [1]. As a result of their adaptability, robustness, and ability to explore large solution spaces, they are ideal candidates for enhancing deep learning models in IoT applications. Optimization techniques are essential for fine-tuning hyperparameters, network architectures, and training strategies in deep learning models to ensure optimal performance and security. With IoT algorithms, deep learning models are more

accurate and efficient and better defended against threats, ensuring privacy and data integrity.

Recently, researchers have become increasingly interested in cloud computing (CC) [2]. With the ability to provide on-demand resources and services, the ability to operate data centres, power systems, intelligent transportation systems, video delivery systems, earthquake command systems, and other technological environments has been impacted [3]. A CC model considers a number of design objectives, including energy efficiency, fairness, reliability, and fault tolerance. Security, however, is considered the most important design objective. It is common for IoT to appear in CC ecosystems [4]. The technology aggregates geographically dispersed cyber-enabled systems or cyber-physical devices to provide strategic services. A major goal of the IoT platform is security, as with CC [5]. Cyberattacks are becoming increasingly prevalent on IoT devices [6], [7]. The internet is interconnected with many IoT gadgets, making it easier to abuse security controls [8]. IoT poses several security threats, including a number of vulnerabilities. The IoT is prone to a variety of attacks, so categorizing them and identifying appropriate vulnerabilities is important. IoT systems

are likely to be affected by routing, jamming, sinkholes, DoS, wormholes, worm attacks, flooding, and viruses, according to some research [9]. It is more specific to IoT platforms in production that are attacked and flooded with DoS attacks [10], [11].

In recent years, botnet attacks have become increasingly popular. The botnet damages services by causing disruptions and depleting resources. Such IoT attacks are commonly discovered using artificial intelligence. Through intrusion detection systems (IDS), malicious activity can be detected on networks. According to the detection system, IDS fall into two categories: anomaly-based and signature-based [12], [13]. As a first step, we can use baseline network behaviour. In this method, malicious events, both unknown and known, will be detected. In the next step, a specific pattern from the network (such as a sequence of bytes) will be applied. Afterwards, it compares the sequences to current databases of signatures [14]. A recent study found that deep learning (DL) is a better method for detecting IoT assaults than traditional machine learning (ML). However, the cloud layer can only run these algorithms because of its limited resources. Moreover, in some cases, such as remote live functions, these methods do not work because the system is assumed to make realistic decisions faster.

As IoT devices and cloud services converge, the ecosystem faces several complexities due to its scale, heterogeneity, and dynamic nature. It is difficult to devise universal detection methods for IoT devices because of the sheer number of different communication protocols and abilities. In addition to encrypted communication between devices and cloud services, it is also difficult to detect botnet activity in network traffic. As IoT devices have limited resources, it isn't easy to design resource-intensive detection approaches, so lightweight yet effective approaches need to be designed. Since botnets are distributed and can mimic legitimate device behaviour, it's difficult to identify malicious activity and command-and-control nodes. IoT devices lack uniform security standards, which exacerbates the detection of botnets due to the rapid development of attack approaches [15].

This issue is addressed by designing a hybrid metaheuristic based on machine learning called HMMLB-BND that is coupled with the Cloud Assisted Internet of Things (CASIOT). The present HMMLB-BND approach enables feature selection (FS) by using the modified firefly optimization (MFFO) technique [16]. In the HMMLB-BND technique, convolutional neural networks (CNNs) and quasi-recurrent neural networks (QRNNs) are

used to detect botnets based on a hybrid algorithm that combines these two types of networks. Hyperparameter optimization uses CBOA. Simulating the N-BaIoT data demonstrated the enhanced performance of HMMLBBND. Here are the key contributions:

- The purpose of this paper is to present a new HMMLB-BND technique that is based on MFFOs that select features for botnet detection, CNNs-QRNNs that classify users, and CBOAs that tune hyperparameters for botnet detection. We are unaware of any literature descriptions of HMMLB-BND.
- MFFO algorithm is presented for feature selection, which overcomes the issue of ineffective exploration capability and local optima.

2 LITERATURE REVIEW

They propose a whale optimization algorithm (WOA) that uses nature-inspired meta-heuristics [17]. A whale-inspired algorithm solves the target problem by imitating whales' predatory behaviour. Using bubble nets and two operators, the device mimics the behaviour of humpback whales, including searching for prey and surrounding prey. Thirty-nine problems are evaluated as part of the WOA evaluation, 29 of which involve mathematical optimization challenges, and 6 which involve structural design challenges. WOA has the potential to solve problems involving unknown search spaces, as shown by the results of the study. In WOA, there is a slow convergence.

Based on the behaviour of grey wolves, the author [18] developed a metaheuristic optimization algorithm known as grey wolf optimization (GWO). The simulation is based on real-life social structures and grey wolves' hunting processes. To mimic the hierarchy of leadership, grey wolves are divided into four types: beta, delta, and omega. Mathematically, alpha wolves are the most optimal solution. According to experts, beta wolves, delta wolves, and omega wolves are the second, third, and fourth best solutions. Despite its popularity, however, this algorithm suffers from low precision, slow convergence, and a poor local search capability.

Based on the Salp Swarm Algorithm (SSA), a modern metaheuristic algorithm has been developed for simulating salps in deep water [19]. In several FS algorithms, SSA is used as a search strategy [20]. According to [21], who addressed the problems of the SSA algorithm, this trend of opportunistic search behaviour improvement was also observed. Local

search (LS) improved the SSA's ability to be exploited. In addition, a chaotic map and a new equation variable were used to find the most effective location update method. To evaluate the effectiveness of the recommended technique for feature selection, a benchmark categorization dataset and three Hadith datasets were analyzed. A dynamic SSA performed better than alternative solutions.

Using Restricted Boltzmann Machines (RBMs), an SC intrusion detection system was proposed in [22]. With RBMs, high-level features can be learned from raw data, and real-time data representations can be controlled. The authors [23] investigate ML-based techniques for detecting hijackings, GPS jamming, and denial-of-service attacks that can be used against drone attacks. Classification of DJI Phantom 4 methods is based on machine learning classifiers that can work with either normal or malicious signatures [24]. Using a three-stage approach to investigating, reducing, and classifying data traffic [25], intrusion detection is achieved by distinguishing positive trust service requests from false ones. To reduce data and drive classifiers accordingly, the solution utilizes decision trees and deep belief algorithms from machine learning. Through simulations, solutions for detecting intrusion attacks are validated to demonstrate their efficiency.

To solve this problem, the author [26] proposed a hybrid method and an infrastructure method. First, the BoT-IoT identify database was executed, and its 44 effective attributes were included in the ML technique. Five effective machine-learning techniques and widely used performance evaluation metrics were selected to identify malicious and anomalous traffic. Using a smart city platform that detects DDoS and replay attacks in real time, a hybrid DL technique was formulated [27]. SC data (smart soil, smart environment, and Smart River) are available for testing the hybrid method; replay attacks and DDoS attacks can be simulated. The anchor node is a crucial part of the sensor network's cost-effectiveness and energy efficiency because it influences localization in a significant way.

Assaults in today's society can be classified using DL-related techniques based on recent databases [28]. A safeguard was introduced to protect the IoT network's reputation and ensure its access is restricted to those who need it. The base was presented for incorporating IDS into IoT-related networks [29]. In [30], the authors propose a mechanism for detecting botnets at the application layer of DNS services by using a two-level DL structure. The similarity of DNS requests across Ethernet

connections was determined using a Siamese network based on an existing threshold. On the second level of the structure, domain generation is recommended as an alternative to DL architectures to categorize abnormal and normal domain names. By using computing to address some of the current limitations and issues with IoT solutions for smarter cities, we can enhance them and make them more useful. Utilizing the Internet of Things and cloud computing, smart cities will have the ability to deliver novel and enhanced services by leveraging big data stored in the cloud.

3 METHODOLOGY

3.1 An Algorithm for Optimizing the Herding of Elephants Based on the Seagull Adaption (SAEHO)

Based on the combination of Elephant Herding Optimization (EHO) and Self-Adaptive Optimization (SAO), the hybrid SAEHO model was developed. EHO cannot make use of real-time or predictive data to determine ongoing and future search directions. On the other hand, SAO may lead to poor performance in complex optimization tasks due to slow computation times and cumbersome constraints. Taking advantage of both methods, the SAEHO model addresses these challenges. Hybrid models, such as SAEHO, have demonstrated promise for improving optimization and search outcomes. Elephant herds exhibit a harmonious interaction with their environment and engage in clan leadership led by a female matriarch in EHO. As a reflection of this social structure, three basic behavioural rules guide the EHO algorithm.

The motivation and mathematical modelling for the proposed algorithm are discussed in this section.

There are many different species of seagulls around the world, all belonging to the Laridae family of birds. It is possible to find a wide range of seagull species with different masses and lengths. Besides eating insects and fish, seagulls also eat reptiles, amphibians, earthworms, and other species of animals. A seagull is a very intelligent bird. To attract earthworms, they produce a rain-like sound with their feet and use breadcrumbs to attract fish. Neither fresh nor salt water is off-limits to seagulls. The majority of animals are incapable of doing this. The bill of seagulls is equipped with a series of glands that flush salt from the system through openings above their eyes.

3.2 Mathematical Model

There is a discussion of the mathematical models of migration and attack on the prey:

- Migration (exploration). This algorithm simulates how the group of seagulls moves from place to place during migration. There are three conditions a seagull must meet during this phase:
- Avoiding the collisions. A variable A is incorporated into the calculation of the new position of the search agent to prevent collisions with neighbouring seagulls (i.e., other seagulls).

$$C_s = A * P_s(x). \quad (1)$$

The (1) expresses C_s as the position of search agent, which avoids colliding with other search agents, P_s as its current position, x as its current iteration, and A as its movement strategy within a search space.

- The movement towards the best neighbour's direction. When the search agents have avoided colliding with neighbours, they move toward the direction of the nearest neighbour.

$$M_s = b * (p_{bs}(x) - P_s(x)), \quad (2)$$

Search agent P_s is positioned toward the best-fitting search agent (e.g., the fittest seagull) by M_s .

Randomness is responsible for the proper balance between exploration and exploitation. B is calculated as:

$$B = 2 * A^2 * rd. \quad (3)$$

A matriarch leads each elephant clan in accordance with elephant natural habits. Matriarch ci influences each elephant's new position, therefore. It calculates elephant j in clan ci .

$$x_{new,ci,j} = x_{ci,j} + a * (x_{best,ci} - x_{ci,j}) * r. \quad (4)$$

The new position and old position of elephant j in clan ci , are represented by $x_{new,ci,j}$ and $x_{ci,j}$, respectively. Elephant $x_{best,ci}$ is the matriarch of the clan, representing the strongest elephant. An indication of scale factor, $r \in [0,1]$, can be seen as $a \in [0,1]$. For each clan, (5) can be used to determine which elephant is the best.

$$x_{new,ci,j} = \beta * x_{center,ci}. \quad (5)$$

Where $\beta \in [0,1]$ is a factor that determines how $x_{center,ci}$ affects $x_{new,ci,j}$. An individual named

$x_{new,ci,j}$ has been born. Clan ci is centred around $x_{center,ci}$. The (6) can be used to calculate it.

$$x_{center,ci,d} = \frac{1}{n_{ci}} * \sum_{j=1}^{n_d} x_{ci,j,d}. \quad (6)$$

The number of elephants per clan is given by $1 \leq d \leq D$ and n_{ci} . Individual elephants are characterized by their d^{th} dimension $x_{ci,j,d}$. The (6) can be used to update $x_{center,ci}$, which is the clan centre.

Solving optimization problems, one can model the separating process of male elephants leaving their families. Elephants with the worst fitness during each generation implement the separating operator, as shown in (7).

$$x_{worst,ci} = x_{min} + (x_{max} - x_{min} + 1) * rand. \quad (7)$$

3.3 Self-Upgraded Cat and Mouse Optimizer (SU-CMO) Algorithm

However, the current CMBO model tolerates low accuracy while offering optimal solutions. A number of improvements were made to conventional CMBO in order to overcome its drawbacks. As shown in the SU-CMO flowchart and model, the following steps are involved.

We are in the process of initializing the population of BB search agents. Initialization of B, B_c, B_m, T parameters. B indicates how many members are in population matrix A .

According to (8), the initial population is created.

$$A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_B \end{bmatrix}_{B \times m} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,d} & \cdots & y_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{i,1} & \cdots & y_{i,d} & \cdots & y_{i,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{B,1} & \cdots & y_{B,d} & \cdots & y_{B,m} \end{bmatrix}_{B \times m} \quad (8)$$

A variable named $y_{i,d}$ is used here to represent the d^{th} problem.

Based on (9), the fitness of search agents is computed.

$$Obj = \min(Err). \quad (9)$$

The sorted population matrix A^s should be updated using (10) and (11). As shown here, as $y_{1,d}^s$ represents the sorted population matrix for the population i^{th} .

Moreover, Obj_j^s represents the vector of objective functions sorted by

$$A^S = \begin{bmatrix} A_1^S & y_{1,1}^S & \cdots & y_{1,d}^S & \cdots & y_{1,m}^S \\ A_2^S & y_{2,1}^S & \cdots & y_{2,d}^S & \cdots & y_{2,m}^S \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ A_B^S \end{bmatrix}_{B \times m} = \begin{bmatrix} y_{1,1}^S & \cdots & y_{1,d}^S & \cdots & y_{1,m}^S \\ y_{2,1}^S & \cdots & y_{2,d}^S & \cdots & y_{2,m}^S \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{B,1}^S & \cdots & y_{B,d}^S & \cdots & y_{B,m}^S \end{bmatrix}_{B \times m}, \quad (10)$$

$$Obj^S = \begin{bmatrix} Obj_1^S & \min(Obj) \\ Obj_2^S & \min(Obj) \\ \vdots & \vdots \\ Obj_B^S & \min(Obj) \end{bmatrix}_{B \times 1}. \quad (11)$$

A mouse population is selected based on (12).

$$\begin{aligned} C_1 &= X_{Bm+1}^S \\ &\vdots \\ C &= C_i = X_{Bm+j}^S = \\ &\vdots \\ &[C_{B_c} = X_{Bm+B_c}^S]_{B_c \times m} \\ &= \begin{bmatrix} y_{Bm+1,1}^S & \cdots & y_{Bm+1,d}^S & \cdots & y_{Bm+1,m}^S \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{Bm+j,1}^S & \cdots & y_{Bm+j,d}^S & \cdots & y_{Bm+j,m}^S \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{Bm+B_c,1}^S & \cdots & y_{Bm+B_c,d}^S & \cdots & y_{Bm+B_c,m}^S \end{bmatrix}_{B_c \times m} \end{aligned} \quad (12)$$

Accordingly, M, B_m, M_i, C, B_c, C_j indicates the number of mice, the count of mice, the j^{th} mice, the number of cats, and the i^{th} cat. The (13) shows how the positions of cats are updated, where C_j^{new} new points to the j^{th} cat's new position, and $C_{j,h}$ is the d^{th} cat's new value. Also, r is estimated at random within the range $[0, 1]$. An integer is used as a random variable in (14).

$$C_j^{new} = [C_{j,d} + r * (M_{k,d} - I * C_{j,d})]. \quad (13)$$

Here,

$$I = \text{round}(1 + \text{rand}), \quad (14)$$

If $j = B_c$

Assuming the above conditions are met, (15) is used to create H_i .

$$H_i = h_{i,d} = y_{i,d} \& i = 1: B_m, d = 1: m, l \in 1: B. \quad (15)$$

Following that, mouse positions are updated using (16) and (17). The (17) shows the typical method for updating M_i . ra_1 and ra_2 are random integers used in the SU-CMO model to update M_i . It is assumed that ra_1 and ra_2 have values of 1.25 and 1.75 respectively

$$\begin{aligned} M_i^{new}: m_{i,d}^{new} &= m_{i,d} + r * (h_{i,d} - I * m_{i,d}) \\ &+ \text{Sign}(F_i^m - F_i^H), \end{aligned} \quad (16)$$

$$\begin{aligned} M_i &= \begin{cases} M_i^{new} & | F_i^{m,new} < F_i^m \\ M_i & | \text{else} \end{cases}, \\ M_i &= \begin{cases} M_i^{new} & | F_i^{m,new} \cdot ra_1 < F_i^m \\ M_i & | \text{else} \end{cases}, \quad (17) \\ M_i &= \begin{cases} M_i^{new} & | F_i^{m,new} \cdot ra_1 < F_i^m \\ M_i & | \text{else} \end{cases}, \\ M_i &= \begin{cases} M_i^{new} & | F_i^{m,new} \cdot ra_2 < F_i^m \\ M_i & | \text{else} \end{cases}. \end{aligned}$$

4 RESULTS AND DISCUSSION

IoT attacks are organized into eight specific categories: Analysis, Backdoors, Denial of Service, Exploits, Fuzzers, and Reconnaissance. Although they are divided into distinct categories, they are all grouped under the overarching category of "Attack." The chart in Figure 1 shows how these attack types are categorized in Datasets 1 and 2.

In Dataset 1, hybrid classifiers with 80% learning rates were evaluated using the SAEHO algorithm. DTs, SVMs, NNs, SLnOs, and Grey Wolf Optimizers are included in the comparison in addition to Proposed Model 1. The two performance metrics used for comparison are accuracy and Matthews Correlation Coefficient (MCC), as shown in Figure 2. A consistent winner among all the classifiers is the Proposed Model 1 across all metrics. This method achieves the highest levels of Accuracy and Rand Index, indicating that it is reliable for classification and clustering. It also ranks first in F-Measure, indicating a good balance between precision and recall, as well as a significantly higher MCC. According to these results, Proposed Model 1 is superior to the given experimental setup in detecting IoT intrusions.

In Figure 3, the 80% learning rate SAEHO algorithm is used to perform a performance comparison of hybrid classifiers on dataset 2. There are six classifiers evaluated across four metrics: In addition to accuracy, MCC, Rand Index, and F-measure, proposed models include DT, SVM, NN, SLnO, GWO, and WOA. The Rand Index, MCC, and F-Measure consistently outperform the other methods, with near-perfect accuracy in each. As a result of these results, it is demonstrated that the system is robust and effective in detecting IoT attacks.

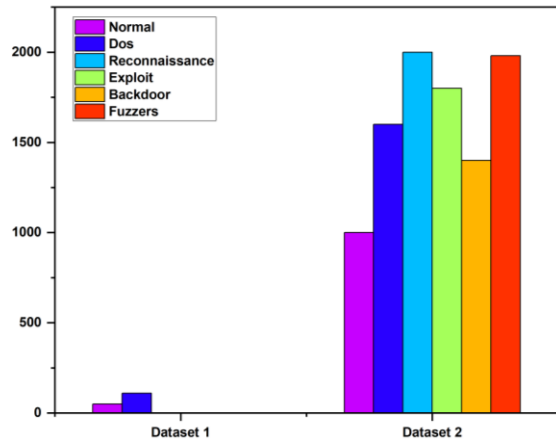


Figure 1: A number of attacks were recorded in Datasets 1 and 2.

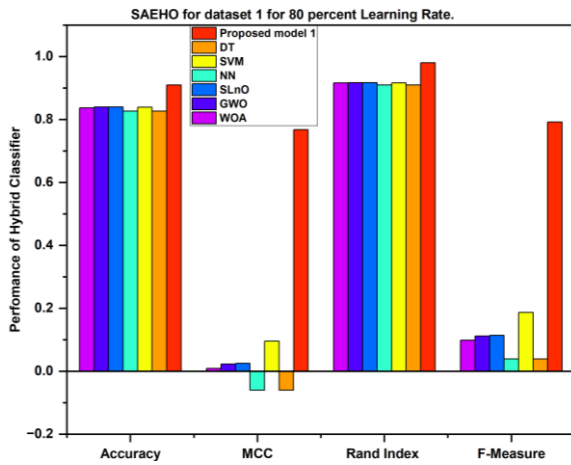


Figure 2: Performance comparison of hybrid classifiers on dataset 1 with 80% learning rate.

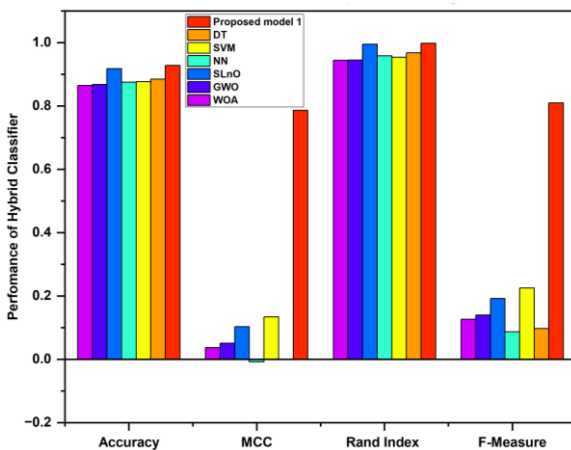


Figure 3: Performance comparison of hybrid classifiers with 70% learning rate.

A hybrid classifier using the SUC-MO method with an 80% learning rate is shown in Figure 4. ALO, GMBO, RNN, NN, and AO were evaluated on Accuracy, MCC, Rand Index, and F-Measure, in addition to the Proposed Model 2. It achieves the highest F-Measure and strong Accuracy and MCC results in comparison with the other models. All of the metrics indicate that Proposed Model 2 performs the most effectively compared to GMBO despite GMBO having a marginally higher Rand Index.

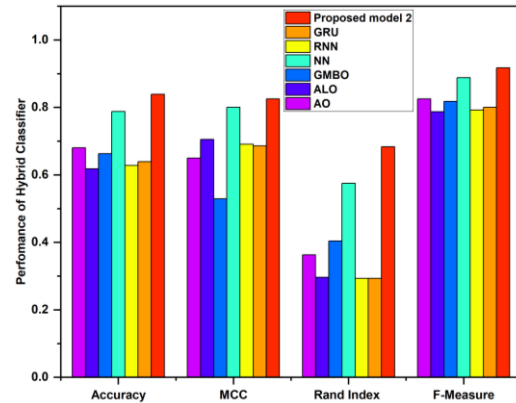


Figure 4: Performance comparison of hybrid classifiers with 75% learning rate.

Dataset 2 is shown in Figure 5 with hybrid classifiers using SUC-MO algorithms at an 80% learning rate. Comparisons have been made between six different models: Proposed Model 2, RNN, NN, GMBO, ALO, and AO in terms of accuracy, as well as MCC, Rand Index, and F-Measure in terms of reliability. Model 2 outperforms others overall, achieving the highest F-Measure and MCC, as well as high accuracy and Rand Index, which demonstrate its effectiveness for classifying IoT data

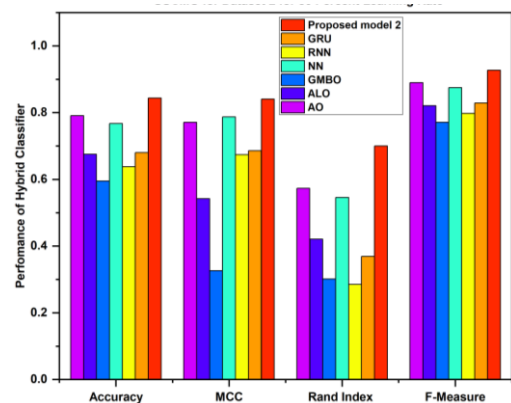


Figure 5: Performance comparison of hybrid classifiers with 80% learning rate.

5 CONCLUSIONS

An advanced metaheuristic optimization technique and deep learning models are combined to provide comprehensive botnet detection for cloud-based IoT systems. In the proposed HMMLB-BND method, MFFO is utilized for optimizing feature selection, enhancing the relevance and reduction of redundant information. CNN-QRNN is employed to achieve accurate and reliable classification by effectively capturing spatial and sequential dependencies within the data. Additionally, the CBOA algorithm is integrated for precise tuning of hyperparameters, ensuring optimal model performance. Furthermore, the study explores advanced metaheuristic algorithms such as SAEHO and SU-CMO to further enhance detection accuracy, robustness, and adaptability to varying IoT scenarios. Comprehensive experimental evaluations demonstrate that the proposed models significantly outperform traditional and hybrid classifiers when tested on the N-BaIoT dataset, exhibiting substantial improvements particularly in terms of Matthews Correlation Coefficient (MCC) and F-Measure. Due to its capability to effectively address critical challenges such as high-dimensional datasets, imbalanced classification, limited computational resources, and stringent real-time requirements inherent in IoT environments, the HMMLB-BND framework presents a highly promising direction for future research endeavors and practical real-time deployment within the evolving field of IoT security.

REFERENCES

- [1] P. Rani, P. N. Singh, S. Verma, N. Ali, P. K. Shukla, and M. Alhassan, "An implementation of modified blowfish technique with honey bee behavior optimization for load balancing in cloud system environment," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–14, 2022.
- [2] Z. Chen, "Research on Internet Security Situation Awareness Prediction Technology Based on Improved RBF Neural Network Algorithm," *Journal of Computer and Cognitive Engineering*, vol. 1, no. 3, pp. 103–108, Mar. 2022, doi: 10.47852/bonviewJCCE149145205514.
- [3] K. Shinan, K. Alsubhi, A. Alzahrani, and M. U. Ashraf, "Machine Learning-Based Botnet Detection in Software-Defined Network: A Systematic Review," *Symmetry*, vol. 13, no. 5, p. 866, May 2021, doi: 10.3390/sym13050866.
- [4] S. Namasudra, R. G. Crespo, and S. Kumar, "Introduction to the special section on advances of machine learning in cybersecurity (VSI-mlsec)," *Computers & Electrical Engineering*, vol. 100, p. 108048, May 2022, doi: 10.1016/j.compeleceng.2022.108048.
- [5] A. Gutub, "Boosting image watermarking authenticity spreading secrecy from counting-based secret-sharing," *CAAI Transactions on Intelligent Technology*, vol. 8, no. 2, pp. 440–452, Jun. 2023, doi: 10.1049/cit2.12093.
- [6] S. Das and S. Namasudra, "Multiauthority CP-ABE-based Access Control Model for IoT-enabled Healthcare Infrastructure," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 821–829, Jan. 2023, doi: 10.1109/TII.2022.3167842.
- [7] B. Bhola et al., "Quality-enabled decentralized dynamic IoT platform with scalable resources integration," *IET Communications*, 2022.
- [8] A. A. Laghari, A. A. Khan, R. Alkanhel, H. Elmannai, and S. Bourouis, "Lightweight-BIoV: Blockchain Distributed Ledger Technology (BDLT) for Internet of Vehicles (IoVs)," *Electronics*, vol. 12, no. 3, p. 677, Jan. 2023, doi: 10.3390/electronics12030677.
- [9] M. Waqas et al., "Botnet attack detection in Internet of Things devices over cloud environment via machine learning," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 4, p. e6662, Feb. 2022, doi: 10.1002/cpe.6662.
- [10] A. A. Laghari, X. Zhang, Z. A. Shaikh, A. Khan, V. V. Estrela, and S. Izadi, "A review on quality of experience (QoE) in cloud computing," *Journal of Reliable Intelligent Environments*, vol. 10, no. 2, pp. 107–121, Jun. 2024, doi: 10.1007/s40860-023-00210-y.
- [11] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," *Computers & Electrical Engineering*, vol. 105, p. 108543, 2023.
- [12] A. Wani, R. S., and R. Khaliq, "SDN-based intrusion detection system for IoT using deep learning classifier (IDSIoT-SDL)," *CAAI Transactions on Intelligent Technology*, vol. 6, no. 3, pp. 281–290, Sep. 2021, doi: 10.1049/cit2.12003.
- [13] P. Rani et al., "Federated Learning-Based Misbehavior Detection for the 5G-Enabled Internet of Vehicles," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 2, pp. 4656–4664, May 2024, doi: 10.1109/TCE.2023.3328020.
- [14] F. Sattari, A. H. Farooqi, Z. Qadir, B. Raza, H. Nazari, and M. Almutiry, "A Hybrid Deep Learning Approach for Bottleneck Detection in IoT," *IEEE Access*, vol. 10, pp. 77039–77053, 2022, doi: 10.1109/ACCESS.2022.3188635.
- [15] N. Hussain and P. Rani, "Comparative studied based on attack resilient and efficient protocol with intrusion detection system based on deep neural network for vehicular system security," in *Distributed Artificial Intelligence*, CRC Press, 2020, pp. 217–236.

- [16] P. Rani and R. Sharma, "IMFOCA-IOV: Intelligent Moth Flame Optimization based Clustering Algorithm for Internet of Vehicle," in 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2023, pp. 1–6.
- [17] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Advances in Engineering Software*, vol. 95, pp. 51–67, 2016.
- [18] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, Mar. 2014, doi: 10.1016/j.advengsoft.2013.12.007.
- [19] S. Mirjalili, A. H. Gandomi, S. Z. Mirjalili, S. Saremi, H. Faris, and S. M. Mirjalili, "Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems," *Advances in Engineering Software*, vol. 114, pp. 163–191, Dec. 2017, doi: 10.1016/j.advengsoft.2017.07.002.
- [20] H. Faris et al., "An efficient binary Salp Swarm Algorithm with crossover scheme for feature selection problems," *Knowledge-Based Systems*, vol. 154, pp. 43–67, Aug. 2018, doi: 10.1016/j.knosys.2018.05.009.
- [21] M. Tubishat, N. Idris, L. Shuib, M. A. M. Abushariah, and S. Mirjalili, "Improved Salp Swarm Algorithm based on opposition based learning and novel local search algorithm for feature selection," *Expert Systems with Applications*, vol. 145, p. 113122, May 2020, doi: 10.1016/j.eswa.2019.113122.
- [22] A. Elsaiedy, K. S. Munasinghe, D. Sharma, and A. Jamalipour, "Intrusion detection in smart cities using Restricted Boltzmann Machines," *Journal of Network and Computer Applications*, vol. 135, pp. 76–83, Jun. 2019, doi: 10.1016/j.jnca.2019.02.026.
- [23] Z. Baig, N. Syed, and N. Mohammad, "Securing the Smart City Airspace: Drone Cyber Attack Detection through Machine Learning," *Future Internet*, vol. 14, no. 7, p. 205, Jun. 2022, doi: 10.3390/fi14070205.
- [24] N. Hussain, P. Rani, N. Kumar, and M. G. Chaudhary, "A deep comprehensive research architecture, characteristics, challenges, issues, and benefits of routing protocol for vehicular ad-hoc networks," *International Journal of Distributed Systems and Technologies (IJDST)*, vol. 13, no. 8, pp. 1–23, 2022.
- [25] M. Aloqaily, S. Otoum, I. A. Ridhawi, and Y. Jararweh, "An intrusion detection system for connected vehicles in smart cities," *Ad Hoc Networks*, vol. 90, p. 101842, Jul. 2019, doi: 10.1016/j.adhoc.2019.02.001.
- [26] M. Shafiq, Z. Tian, Y. Sun, X. Du, and M. Guizani, "Selection of effective machine learning algorithm and Bot-IoT attacks traffic identification for internet of things in smart city," *Future Generation Computer Systems*, vol. 107, pp. 433–442, 2020.
- [27] A. A. Elsaiedy, A. Jamalipour, and K. S. Munasinghe, "A Hybrid Deep Learning Approach for Replay and DDoS Attack Detection in a Smart City," *IEEE Access*, vol. 9, pp. 154864–154875, 2021, doi: 10.1109/ACCESS.2021.3128701.
- [28] T. Saba, A. R. Khan, T. Sadad, and S. Hong, "Securing the IoT System of Smart City against Cyber Threats Using Deep Learning," *Discrete Dynamics in Nature and Society*, vol. 2022, no. 1, p. 1241122, Jan. 2022, doi: 10.1155/2022/1241122.
- [29] N. Kumar Agrawal et al., "TFL-IHOA: Three-Layer Federated Learning-Based Intelligent Hybrid Optimization Algorithm for Internet of Vehicle," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 3, pp. 5818–5828, Aug. 2024, doi: 10.1109/TCE.2023.3344129.
- [30] R. Vinayakumar, M. Alazab, S. Srinivasan, Q.-V. Pham, S. K. Padannayil, and K. Simran, "A Visualized Botnet Detection System Based Deep Learning for the Internet of Things Networks of Smart Cities," *IEEE Transactions on Industry Applications*, vol. 56, no. 4, pp. 4436–4456, Jul. 2020, doi: 10.1109/TIA.2020.2971952.

SECTION 4

DATA ANALYSIS AND PROCESSING

Proceedings of the 13th International Conference on Applied Innovations in IT (ICAIIIT)

JULY 2025

Advanced Network Security System: Honeypot-Based Intrusion Detection with Machine Learning and Visualization

Ashwini Lakshmipathy¹, Muthupandi Gurusamy¹, Siti Fatmawati Lalo², Nonia Sakka Lebang³,
Karthikeyan Selvaraj⁴ and Aws A. Abdulsahib⁵

¹Department of Computer Science and Engineering, Chennai Institute of Technology, 600069 Chennai, Tamil Nadu, India

² Department of Master of Laws Study Program, Universitas Sulawesi Tenggara, 93121 Kendari, Indonesia

³ Department of Government Science Study Program, Universitas Sulawesi Tenggara, 93121 Kendari, Indonesia

⁴Center for Advanced Multidisciplinary Research and Innovation, Chennai Institute of Technology, 600069 Chennai, Tamil Nadu, India.

⁵Department of Computer Science, Dijlah University College, 10021 Baghdad, Iraq

ashwinil.cse2022@citchennai.net, pg.muthupandi@gmail.com, stfatmawati@un-sultra.ac.id, noniasakkalebang@un-sultra.ac.id, karthikeyans.mech@citchennai.net, aws.abdulkareem@duc.edu.iq

Keywords: Honeypots, Intrusion Detection Systems (IDS), Machine Learning in Cybersecurity, Network Security Visualization, Anomaly Detection.

Abstract: This paper focuses on developing a proactive approach to network intrusion detection through integration of honeypots with machine learning for improved security in complex network system. The system utilizes honeypots to capture attackers whereby the honeypots capture real-time traffic details which the system maps and analyzes packet content related to protocols. Blending with machine learning, the detection model analyzes the accurate data to detect known as well as unknown forms of cyber threats. There is a new feature called Visualization Dashboard that gives analytics and reports to network administrators. It provides information about honeypot engagements, traffic, and intrusion detected empowering the monitoring and management process. Incorporating the proactive defense measures into the proposed system eliminates the weakness of the conventional intrusion detection approach in managing new forms of cyber threats. The honeypots are designed to contain the attackers and simultaneously acquiring useful information about the intrusive activities. The effectiveness is enhanced by the ability of the Machine Learning model in enhancing the detection rates besides the flexibility in accommodating new techniques of detection of attacks. The Visualization Dashboard improves usability since it contains an easily navigable interface for current security monitoring and past performance examination. This approach guarantees the entirety of network protection by integrating the effectiveness of the deception-based honeypot systems and the machine learning approach based on big data. The paper reveals that the system is capable of enhancing detection rates, reducing false positives and providing valuable information regarding the network status to administrators. Thus, the provided system is considered ideal for contemporary cybersecurity issues. Advanced technologies combined in this system offer a flexible and expandable system to protect networks from the steadily growing number of threats.

1 INTRODUCTION

Network security is one of the central pillars of the present and future society as the global interconnectedness of business, governmental and private life is based on networks. High-intensity threats like ransomware, phishing or DDoS attacks show that the fundamental strategies insufficient Security must evolve. As more information including financial records, health record and intellectual

property is transmitted through networks, the impact of a breach is catastrophic. Additionally, the increased pace at which new waves of IT technologies such as cloud, Internet of Things and Artificial Intelligence is implemented in the organization increases the exposure to threats [1]. These interrelated systems that are driving innovation and enhancing company efficiency are the same systems that makes them susceptible to hackers. Network security has become mandatory since most

businesses have gone online to protect their valuable information and assets from cyber-attacks. Network security is a decoy system that is deployed with the intention of enticing attackers to engage, thus expose their tool, prey and procedures (TTPs) [2]. These interactions offer desired input to security analysts on emergent threats as well as enhance the general security framework. Low-interaction honeypots mainly encompasses known vulnerability honeypots and research honeypots while high-interaction honeypots encompass HoneyD as well as high interaction hoax honeypots. Different type of honey pots are Low interaction honeypots mimic simple service or protocol and thus consume minimum resources, it easily capture key level attacks like scanning or brute force attempt without putting the system in real danger [3]. High-interaction honeypots, on the other hand, emulate normal systems and therefore provide substantial detail about complex intrusions. They are costlier and exposed to higher risk compared to the first type, but offer a much higher level of insight into the attackers tactics and goals. In addition to honeypots, Intrusion Detection Systems (IDS) in the network looking for suspicious events or policy infringements that warrant an alarm to be issued. IDS are divided into two types: The first type is the signature-based IDS, which identifies threats based on a prior known signatures for comparison with network performance, the second is the anomaly-based IDS which identifies abnormal performance or activity that suggests threats. Whereas, honeypot attracts attackers to let them learn more about them, IDS run all the time and are different layers to network. Machine learning has also improved IDS because it has been able to overcome some common issues including high number of false alarms, inability to identify new threats and dependence on static signatures [4]. Machine learning uses linear methods, clustering, classification, and artificial neural networks, which help detect as yet unknown or new-age threats. They also help in the elimination of false positives wherein only real anomalous signals are picked up, and thus provides less noise for the security groups to filter [5]. Essential applications of the current machine learning models of big data include the capacity to analyse the real-time traffic data and offer proactive solutions to detected threats. Furthermore, ML systems can update relevant knowledge bases via real-time learning for new learning cases of attacks and new environments dynamically. Immersive the honeypots with IDS, and machine learning is a full proof and comprehensive solution to confront the threats on the network security [6], [7].

2 LITERATURE SURVEY

A lot of literature today has explored the use of honeypot, Intrusion Detection System, and machine learning to bolster network security in the recent developments in threats. Honeypots have been known to be valuable measures that enable the gathering of intel on the attacker performing various kinds of analysis [8], [9]. Spitzner (2003) defined honeypots as a preventive security that can bend attackers and mimic their strategies, processes and procedures (SPP). Further works developed more kinds of honeypots; low interaction honeypots that provide simple services of network and high interaction honeypots that provide full systems to get better understanding of complex attacks [10] - [12]. Low-interaction honeypots are inexpensive and safe while providing relatively limited data, on the other hand, as described by Seung Woo et al. (2015), high-interaction honeypots generate more data but require more resources and can lead to system compromise. Similar to firewalls, IDS are another crucial paradigms of network security research. Other conventional approaches to IDS, for example the signature-based approach have been developed and applied widely since they work well against the known threats [13] - [15]., Nevertheless, a major drawback of such schemes is their ability to do not identify new or unknown attacks. This has been accomplished by the introduction of Anomaly based IDS which are IDS that observe anomalies within the framework of the normality. Similarly, Liao et al. (2013) mentioned that anomaly-based detection can accurately identify zero-day attacks, but at the same time it has a major disadvantage, being high false positive rates [16] - [18]. This shortcoming has led to development of work that seeks to integrate IDS with machine learning algorithm for better results and to eliminate false alarms. The IDS systems' capability in applying machine learning in IDS has greatly enhanced systems due to the dynamism of the network and the emerging threats in the market. Another research by Bou-Harb et al. (2017), emphasizes that supervised learning algorithm including SVM and Random Forest are capable to categorize known threatening types including DDoS and phishing. Other methods of learning based on the independence of the model from initial data includes clustering for the identification of new threats in form of outliers [19], [20]. Neural networks especially the deep learning architectures have been widely embraced because of their capability in analyzing high dimensional data to produce symphonies of detecting complicated patterns in

traffic flow in the network. For example, Tang et al. (2018) provide a proof of concept of how to use convolutional neural networks (CNNs) for accurate identification of malware traffic. The integration of honeypots and machine learning has been a topic of considerable interest in the recent past. The basis of this method is the application of honeypots in live network environments to collect attack data in real conditions. As pointed out by Choi et al. (2020) on the same, this approach allows IDS to be useful in situations where new threats are not recognized thus possessing no signature. More so, honeypot systems are reinforced by machine learning to analyze the captured data without requiring human input. This integration gives a dynamic and flexible system that can embrace modern cybersecurity challenges that are unique hence making the system scalable. Visualization has been found as an auxiliary solution to honeypots and IDS providing the analyst with an environment to analyze measures and make decisions based on them. According to Yip et al. (2019), visualization proved to be effective in representing attack patterns, distribution patterns, and threat intelligence in a form of interactive dashboards. There are libraries such as D3.js and Matplotlib, used to represent the activity in form of graphics on the network while the monitoring tools such as Power BI offer data visualization as part of a single interface. Combining visualization with honeypots and machine learning guarantees that she is delivering the right information in a manner that can improve the overall tactical awareness, as well as response times. The significance of evaluation metrics can be well understood with the help of the given figure. The basic measures including accuracy, precision, recall rate and the F1 numeric figure rating are commonly used to compare IDS and machine learning models. In their study, Sharma et. al (2021) highlighted that one must always employ balanced metrics and especially when there are certain types of attacks that have for instance fewer samples in a dataset. Researches also describe pre-testing of models in real life conditions to determine their effectiveness and stability in practice.

3 MATERIALS AND METHODOLOGY

3.1 System and its Implementation

The identified solution is the use of honeypots and IDS with machine learning and is a proactive and dynamic security model. Honeypots on the other

hand are fake systems that imitates real ones by providing an interface to the network services, attackers are drawn to thus gathering vast information on their techniques and actions. Known attacks are identifiable as they are present in the IDS component and developed from the data collected from users that help the system detect various types of attacks, including known and unknown. The data collected are later used to train a machine learning model improving detection potential of the system and allowing it to escalate protection against new cyber threats. The data captured is also parsed to obtain protocol details to packet level; pay load size and other features. These insights provide the inputs to update the machine learning algorithms to provide a more dynamical and efficient detection system. Figure 1 shows the integration of main components of network security [11], [12].

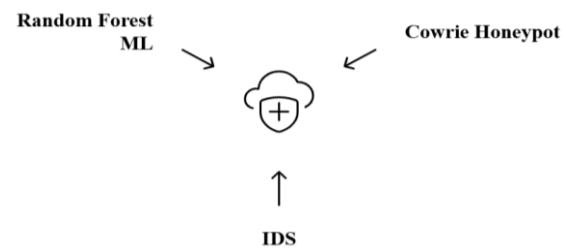


Figure 1: Integration of components of network security.

3.2 Honeypot Deployment

It includes non intrusive honeypot like cowrie which logs brute force and protocol level attacks and the more complex honeypots ready for entertaining smart attackers in a monitored and constrained environment. One type of honeypots, and which are also low-interaction honeypots, are focused specifically on emulating popular services such as SSH and Telnet and logging all activities, including commands typed in by the attacker and the replies they get. From this data information on the behavior of attackers can be gathered as well as information that will assist in the identification of regular target points. In contrast high-interaction honeypots are set up to emulate genuine OS environments using virtualization tools such as VMware this makes the attackers to fully engage with a seemingly real system. This makes sure that multiple layer honeypot would capture all threats, from the simple reconnaissance to even complex credential stuffing or cracking attacks. When these honeypots are placed in specific portions of the network, the legitimate resources of an organization is protected, while the

lure of apparently unprotected hosts is provided to the attackers. Further, the deployment is interfaced with system monitoring tools and analyzers that employ machine intelligence techniques to protect networks in real-time. This approach does not only defend against direct threats but also produces valuable learnings for improving the general security of the network, making the approach a significant foundation of contemporary advanced threat defense system [13], [14].

3.3 Intrusion Detection System Integration

The IDS component is implemented simultaneously to the honeypot to reanalyze captured traffic in real time and complement the system response to threats. It categorize known and unknown or blind traffic, it sometimes utilize sophisticated IDS tool like Suricata for particular threats and abnormality detection. The incorporation of signature-based detection methods provides the ability to detect known threats, while the anomaly-based methods make it possible to also detect new threats, including the zero-day attacks, as well as APTs. This double protection guarantees that the system effectively identifies not only 'baseline' viruses but also subsequent enations that are impermeable to conventional protection systems. The IDS also filters out other insignificant elements of captured traffic and extracts imperative attributes for analysis. These include the actual protocol applied, the source and destination IP addresses, packet sizes and volumes, the connection duration, and nature of the payload, etc., necessary to train and improve the performance of the ML algorithms embedded in the system. Considering this structured and high-quality data, the machine learning models achieve higher predictive accuracy and flexibility to respond to attacks with newly identified patterns. Furthermore, the IDS component is colored into the system's visualization interface to display current and historical trends of the involved traffic, threats and suspicious activities. This enables the network administrators to take informed decisions much earlier. The IDS when collecting data from honeypots and integrating it with machine learning algorithms forms a highly effective defense system

that is capable of providing great deal of protection to the network [15], [16].

3.4 Machine Learning Model

The Random Forest classifier is the key component of the developed detection system because of its high levels of reliability, scalability, and explicability of the results. Random Forest is particularly used when dealing with large datasets, which are imbalanced on the side of malicious traffic, and in contrast to which there is significantly more normal traffic. Random Forest is a set of decision trees that are superior to other algorithms when it comes to recognizing faint patterns in traffic logs, it minimizes the problem of overfitting and increases the likelihood of making correct predictions. The training data for the model come from the honeypot logs, combined with additional labeled data from other sources, including CICIDS or UNSW-NB15 containing a diverse set of threats to ensure sufficient training [3]. To strengthen the supervised paradigm the system also utilize the unsupervised paradigm like the k-Means Clustering as an anomaly detector. This method helps to find various clusters that make up network traffic and flags unusual activity which may depict zero-day threats and other rising attack categories [17], [18]. Using parameters such as packet size, connection duration, and behavior, k-Means Clustering assists with identifying entirely fresh attack types that have no known signatures. The integration of a supervised and an unsupervised paradigm ensures that the system is able to provide a complete solution to a diverse range of threats. It accurately categorizes well understood threats while at the same time, discovering new and complex threats and threats that an organization has not yet encountered thereby providing an organization with a proactive security stance. Furthermore, the inclusion of these algorithms into the architecture allows for the system flexibility to counter changing types of cyberspaces, providing a flexible and forward-thinking solution for security against further intrusions and protection of the net space [18], [19]. Table 1 shows the functionality, limitation with example for the components of network security.

Table 1: The functionality and limitations of the components used in network security.

Aspect	Honeypot	Intrusion Detection System (IDS)	Machine Learning
Purpose	Deceives attackers to gather data on malicious activities	Monitors network traffic for potential threats	Enhances detection by analyzing patterns in data
Key Functionality	Captures attacker interactions and logs	Detects known threats using signature-based or anomaly-based techniques	Identifies both known and unknown threats through training models
Strengths	Provides detailed insights into attacker behavior	Offers real-time monitoring and automated alerts	Adapts to evolving threats with continuous learning
Limitation	Vulnerable to fingerprinting by advanced attackers	May miss zero-day attacks or produce high false positives	Requires significant data preprocessing and computational resources
Example Tools	Cowrie(low-interaction), VMware-based high-interaction honeypot	Suricata, Snort	Random Forest, K-means clustering and Neural Networks

3.5 Experimental Workflow

The approach to carrying out the experiment involves use of honeypots in a simulation environment that replicates real network conditions in order to achieve realistic results. This setups are created using tools such as VMware and VirtualBox and the honeypot is placed in a way it emulates weak services and protocols. The attack scenarios are performed using the sophisticated software tools such as Metasploit, Kali Linux to attack different protocols of SSH, HTTP, FTP and the likes. These types of simulated attacks afford a host of data that includes brute force attempts, unauthorized access, protocol exploits, and malware injections. Every honeypot records a large amount of information about the interactions, the commands typed by the attacker, the data exchanged, the connection parameters, etc. that is then stored in a common database allowing for a coordinated analysis. This database forms the basis for feature enhancement and data pre-processing to form the next phase of the work. Analytical data attributes, for example, packet size, source/destination IPs, protocols used, connection duration, and payload content are extracted and subsequently normalized into preprocessed format ready for machine learning models. This step is important in enhancing the performance of the models in detecting threats since the false positives were a major concern. It is a real time system in that the IDS constantly scrutinizes the traffic for a match with the patterns in the database while using anomaly detection. All the abusive actions lead to an alert and further, the course of

action is also fast identifying it, thus declaring it safe. The results of these operations are incorporated into the Visualization Dashboard where network administrators are presented with simple and easy to use reports. These reports us figures on honeypot interactions, intrusions, traffic patterns, and network health. In addition, the experimental setup is also scalable to accommodate new form of attacks and protocols due to constantly changing threat in cyber space. This real-time and dynamic environment not only improves the identification of intrusions and their subsequent analysis but also provides means and ways for administrators, so that network can be made secure in advance much more effectively [11].

3.6 Visualization and Evaluation

The Visualization Dashboard involves features like Flask for the backend and React.js for the front end. It provides actual-time statistics on honeypot interactions and intrusion identification, traffic amounts and other parameters. The assessment of the system is done in regard to parameters like the detection rate, the false positive ratio and the time taken for processing. During testing, the system yielded a True Positive Rate of 93% for known threats and with the help of the unsupervised models, discovered new patterns of traffic flows. The evaluation shows that honeypots, IDS and machine learning improve not only the detection capabilities of a network but also result in a better understanding of attacker behavior that can be used to counter

attacks. Figure 2. shows the components of security visualization Dashboard [12], [13].

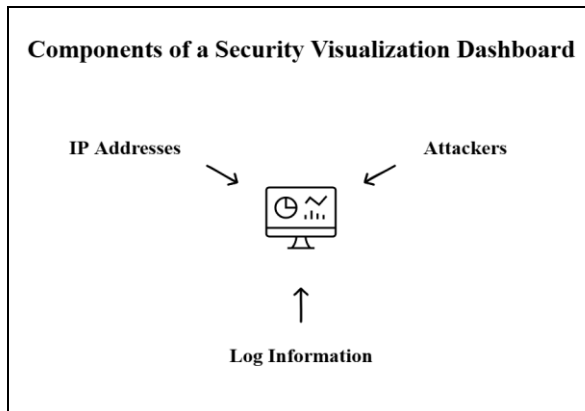


Figure 2: Components of security visualization Dashboard.

4 RESULTS AND DISCUSSIONS

By using this innovative approach, the system offers valuable protection against various kinds of cyber threats. Not only did system integration led to an increase in detection rates, it also decreased the likelihood of false positives, thus presenting a sophisticated security service. Among the elements of this system is Cowrie honeypot, type of the low-interaction honeypot with the purpose to attract the attacker and collect data about their actions. The reason for choosing Cowrie honeypot was because of its simplicity to set and it was able to capture attack patterns including brute force attack and invasion of other network services such as SSH and Telnet services. Cowrie is particularly well adapted to this role due to the detailed logs of the interactions that it produces, to include commands entered by attackers, response and connection parameters which are critical for the documentation of malicious behaviour. The effectiveness of the honeypot in emulating a target network environment while lacking the actual risks of the vulnerability in the network was instrumental to the collection of real time data for further analytical purposes. This information in its turn was used as a basis for populating Intrusion Detection System and for the construction of the machine learning model, which provided detailed information on potential threats. IDS component of the system was in synergy with the honeypot to analyse the traffic generated in the organization's network. In traditional schemes, signature-based detection was used to detect known threats by using pattern match against a database of

known attack signatures. However, as with most IDS systems of a traditional design, the system was not capable of identifying new and emerging forms of attacks, its integration with machine learning, however, solved this problem. The IDS functioned as a rudimentary defense mechanism to demonstrate the transmission of log traffic on the basis of attack signatures while passing what it could not identify as any or unidentified data to the machine learning model for analysis. The Random Forest method is applied for the performance management of the given system as this supervised learning approach has been proven accurate and reliable in working with imbalanced data, which is a characteristic of machine network traffic. Random Forest was selected because of its high efficiency in large datasets and as it constructs several of decision trees, these trees would vote on the best probable classification of the point of interest. The ensemble learning method was very effective in this case as it differentiated the network intrusions and modelled some patterns that might not be valued using the signature detection approach. The interaction information was labeled with the aid of honeypots, and the machine learning model of the system was trained with normal network traffic data as well as malicious traffic. Furthermore, the number of attack models used in the training process were high because the dataset contains traffic of various attack types involving brute force, SQL injections, and other kinds of attacks. After training, a high performance of Random Forest model is identified: the accuracy of traffic classification as normal and malicious is 93%. This high accuracy signifies that the created model proved capable of discerning a number of well-known and innovative types of attacks correctly. In addition, the exalted percent false positive achieved was 4%, which was relatively impressive given the fact that many traditional IDS solutions often produce large numbers of false alarms. This reduction in false positives was particularly crucial for enhancing the detection capability of the IDS. With new attack approaches and strategies coming up, the system can be modified, and retrained from the data captured from the honeypots. By doing so, the machine learning model continues working optimally as the criminals continue devising new methods of attack. The Random Forest algorithm can adapt to new data and the IDS was designed to analyse traffic in real-time, therefore, the detection performance of the system could progressively increase. This was especially remarkable given that the architecture of the machine learning model allowed the solution to detect new

zero-day attacks and other forms of intrusion that are not recognizable from previous attempts by IDS conventional systems. The other component was Visualization Dashboard through which the administrators would be able to have a friendly-user interface to oversee the security of the network. The dashboard contained the actual statistics of the honeypots results that showed the details of the attackers' interaction, types of attacks, and traffic intensity. It also offered historical data analysis to the administrators about how efficient the system is and how the attack activity varies over a period. This feature of simplicity of use was helpful in that it saved time to the administrators in understanding and analysing the data that had to be given to them about security threats and how they should respond to them. Table 2 shows the effectiveness of the system in detecting attacks, reducing false positives and overall usability.

Table 2: The effectiveness of the system.

Category	Results/Percentage
Detection Accuracy	93%
False Positive Rate	4%
Attack Scenarios Detected	100% (Known attacks detected by IDS, Unknown by ML model)
Emerging Threat Detection	90% (High detection rate for new attack patterns)
Usability (Dashboard Feedback)	85% (Positive feedback on efficiency, real-time alerts, and ease of use)

5 CONCLUSIONS

The usage of Cowrie honeypots, IDS and the Random Forest machine learning model indicated that the utilized integrated system provided a very effective protection from a large number of cyber threats. Nevertheless, the system detection accuracy being 93% and false positive rate of only 4% was substantially above the traditional security means; moreover, it provided both prompt protection by applying signatures and prevention of new threats by machine learning. The elaborated ability to identify both familiar and unfamiliar attack patterns insured the most secure protection from a vast spectrum of attacks, such as brute force tries and network service exploitations. Furthermore, regarding the given dashboard for the network monitoring and the feedback collected from the usability point of view, revealed its valuable real-time functionality of the dashboard, straightforward interface, enabling the

network administrators to act swiftly and without much hesitation towards the threats. Thus, the presence of both highly efficient detection functions and easy-to-use and convenient interfaces guarantees the expected level of system security and its practical applicability for constant network protection in conditions of active threats.

REFERENCES

- [1] I. M. M. Matin and B. Rahardjo, "Malware detection using honeypot and machine learning," IEEE, vol. 7, pp. 1-4, 2019.
- [2] R. Vishwakarma and A. K. Jain, "A honeypot with machine learning based detection framework for defending IoT based botnet DDoS attacks," IEEE, pp. 1019-1024, 2019.
- [3] K. Jiang and H. Zheng, "Design and implementation of a machine learning enhanced web honeypot system," IEEE, pp. 957-961, 2020.
- [4] H. N. Abosooda, S. B. Ariffin, O. M. Alyasiri, and A. A. Noor, "Evaluating the Effectiveness of AI Tools in Mathematical Modelling of Various Life Phenomena: A Proposed Approach," IJDS, vol. 2, no. 1, pp. 16-25, Jan. 2025.
- [5] K. Hara and T. Takahashi, "Machine-learning approach using solidity bytecode for smart-contract honeypot detection in the ethereum," IEEE, pp. 652-659, 2021.
- [6] M. R. Siddique et al., "Integrating Machine Learning-Powered Smart Agents into Cyber Honeypots: Enhancing Security Frameworks," in Proc. IEEE 9th International Conference for Convergence in Technology (I2CT), 2024, pp. 1-7.
- [7] A. F. Garap, "A Neural Network Architecture Combining Gated Recurrent Unit (GRU) and Support Vector Machine (SVM) for Intrusion Detection in Network Traffic Data," 2017, [Online]. Available: <https://arxiv.org/abs/1709.03082>.
- [8] S. Kandanaarachchi, H. Ochiai, and A. Rao, "Honeyboost: Boosting honeypot performance with data fusion and anomaly detection," 2021, [Online]. Available: <https://arxiv.org/abs/2105.02526>.
- [9] D. Fraunholz, M. Zimmermann, and H. D. Schotten, "An Adaptive Honeypot Configuration, Deployment and Maintenance Strategy," 2021, [Online]. Available: <https://arxiv.org/abs/2111.03884>.
- [10] M. Di Mauro, G. Galatro, G. Fortino, and A. Liotta, "Supervised Feature Selection Techniques in Network Intrusion Detection: a Critical Review," 2021, [Online]. Available: <https://arxiv.org/abs/2104.04958>.
- [11] M. Esmaili et al., "Machine Learning-Assisted Intrusion Detection for Enhancing Internet of Things Security," 2024, [Online]. Available: <https://arxiv.org/abs/2410.01016>.
- [12] G. Sarkar, H. Singh, S. Kumar, and S. K. Shukla, "Tactics, techniques and procedures of cybercrime: A methodology and tool for cybercrime investigation process," in ACM International Conference Proceeding Series, 2023, [Online]. Available: <https://doi.org/10.1145/3600160.3605013>.

- [13] S. A. Raghul, G. Gayathri, R. Bhatt, and K. A. V. Kumar, "Enhancing cybersecurity resilience: Integrating IDS with advanced honeypot environments for proactive threat detection," in Proc. 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2024, pp. 1363-1368, [Online]. Available: <https://doi.org/10.1109/ICAAIC60222.2024.10575865>.
- [14] H. J. Alhamdane and M. Nickray, "Enhancing the Efficiency of Routing Strategies in WSNs Using Live Streaming Algorithms," *Journal of Technology*, vol. 6, no. 4, pp. 27-39, Dec. 2024.
- [15] P. T. Duy et al., "Investigating on the robustness of flow-based intrusion detection system against adversarial samples using generative adversarial networks," *Journal of Information Security and Applications*, vol. 74, p. 103472, 2023, [Online]. Available: <https://doi.org/10.1016/j.jisa.2023.103472>.
- [16] G. Agrawal, A. Kaur, and S. Myneni, "A review of generative models in generating synthetic attack data for cybersecurity," *Electronics*, vol. 13, no. 2, p. 322, 2024, [Online]. Available: <https://doi.org/10.3390/electronics13020322>.
- [17] M. Mittal, K. Kumar, and S. Behal, "Deep learning approaches for detecting DDoS attacks: a systematic review," *Soft Computing*, vol. 27, pp. 13039-13075, 2023, [Online]. Available: <https://doi.org/10.1007/s00500-021-06608-1>.
- [18] A. I. Jony and A. K. B. Arnob, "A long short-term memory based approach for detecting cyber attacks in IoT using CIC-IoT2023 dataset," *Journal of Edge Computing*, vol. 3, pp. 28-42, 2024, [Online]. Available: <https://doi.org/10.55056/jec.648>.
- [19] N. J. Singh, N. Hoque, K. R. Singh, and D. K. Bhattacharyya, "Botnet-based IoT network traffic analysis using deep learning," *Security and Privacy*, vol. 7, p. 355, 2024, [Online]. Available: <https://doi.org/10.1002/SPY2.355>.
- [20] M. Aljebreen et al., "Enhancing DDoS attack detection using snake optimizer with ensemble learning on Internet of Things environment," *IEEE Access*, vol. 11, pp. 104745-104753, 2023, [Online]. Available: <https://doi.org/10.1109/ACCESS.2023.3318316>.

A Hybrid Static-Dynamic Analysis Model for Android Malware Detection: Design, Implementation and Comparative Assessment

Shatha Hamead Othman and Huda Abdulaali Abdulbaqi

*Department of Computer Science, College of Science, Mustansiriyah University, 10052 Baghdad, Iraq
shathaalqaisy20@uomustansiriyah.edu.iq, huda.it@uomustansiriyah.edu.iq*

Keywords: Android, Malware Detection, Hybrid Approach, Static Analysis, Dynamic Analysis.

Abstract: The attack surface for cyber threats targeting the Android platform has grown dramatically due to the widespread use of Android handsets, the openness of the Android ecosystem, coarse-grained authorization structures, and the invocation of third-party code. The effectiveness of machine learning methods in identifying Android malware has been shown by recent studies. In this work, we provide a hybrid analysis approach that combines static and dynamic analysis to detect Android malware in a dependable and efficient manner. This hybrid model increases detection precision and accuracy while also improving feature extraction procedures. Additionally, we compare the results of static and dynamic analysis with the hybrid approach and look at how each affects classification performance separately. According to experimental results, our hybrid model performs better than other models, achieving 98.9% accuracy, 99.1% precision, 98.3% recall, and 98.7% F1-score. These results indicate a 12% and 22% increase in detection accuracy, respectively, over static and dynamic analytic techniques. Additionally, our findings highlight the limitations of using static or dynamic analysis alone, in terms of detection accuracy, resource efficiency, and behaviour profiling of Android malware. Overall, the study highlights the effectiveness of hybrid analysis in enhancing malware detection systems and achieving more reliable and accurate security classifications.

1 INTRODUCTION

Android has emerged as the most popular smartphone operating system due to the quick development of mobile intelligent terminals [1]. In the mobile sector, Android OS held a global market share of over 71.9% as of July 2023. Google Play is the official app store running on Android smartphones. There were over 2.9 million applications on it as of May 2021. Approximately 2.5 million of them are categorized by AppBrain as standard apps, while 0.4 million are classified as low-quality apps. Android is more attractive to thieves due to its widespread use, but it is also more susceptible to viruses and malware [2]. However, because Android apps are widely distributed, open-source, and have coarse-grained permission management, they can be obtained from potentially dangerous third parties outside of the official Android Market, leaving the platform open to malware attacks [3]. To address these security concerns, various methods for detecting Android malware have been proposed [4]. One of the best methods for detection is machine learning [5]. There are a couple of primary sources of features in the

machine learning-based Android malware detection implementation: both dynamic and static extraction [6].

The Dalvik bytecode, native code, manifest, sound, image, and inverted APK files are the sources of the static features. By executing APK files in the environment, the dynamic characteristics are retrieved from code execution, paths, variable value tracking, sensitive function calls, log records, and other behaviors that take place during the application's operation [4]. By employing sophisticated preprocessing to choose the best samples, providing rich data, and enabling machine learning for effective malware detection using the CIC-AndMal2017 dataset, this study suggests a reliable way for identifying malicious Android applications. The strategy makes use of machine learning methods, including Random Forest, Support Vector Machines, K-Nearest Neighbors, and Stacking Ensemble Model to classify applications using dynamic analysis, which deals with dynamic features, and static analysis, which concentrates on static features. Following the implementation of these strategies, the same model was then assessed through hybrid analysis, which combines static and dynamic

data to maximize feature extraction and improve malware detection efficacy and precision. When compared to the hybrid strategy, both individual approaches showed a considerable decrease in performance.

Therefore, it is essential to compare how static, dynamic, and hybrid analysis affect an attack detection system's performance. Every analysis type has its own advantages and disadvantages. For example, static analysis looks at the application's code without running it, which makes scanning quicker and more comprehensive but may miss runtime behaviors. Although it frequently takes more time and processing resources, dynamic analysis, on the other hand, watches how the program behaves while it is running and can reveal dangerous activity that static approaches could miss. By utilizing the advantages of both methods, hybrid analysis may be able to maximize detection accuracy. Assessing each of the three approaches yields a thorough comprehension of how each one affects system performance separately and in combination. The main contribution of this paper is:

- Assess a number of machine learning classifiers using three different analysis types: static, dynamic, and hybrid.
- Contrast the outcomes of malware identification utilizing features based on static and dynamic analysis.
- Compare the cost of using various feature types in terms of the amount of time needed for training.
- A large dataset and a hybrid technique can achieve high classification accuracy.

The rest of this paper is displayed as follows: Section 2 provides an overview of the related work. The suggested methodology is introduced in Section 3 that includes a sub-subsection: 3.1 Dataset Description, 3.2 The Specifics of the Suggested Approach. The result and discussion are provided in section 4. Whereas section 5 summarizes the conclusions.

2 RELATED WORK

There are three methods to examine the malicious mobile application: (Hybrid analysis, dynamic analysis, and static analysis) based on various retrieved features, either dynamic, static, or both. This section presents pertinent studies on the application of machine learning algorithms and various analysis methodologies for Android malware detection using the CICAndMal2017 dataset.

2.1 Static Analysis-Based Features

The program's resources and source code are used for static analysis without executing the code [7]. In static malware analysis, we have extracted different types of static features i.e. API calls, intents, permissions and command strings. Omar N. Elayan and Ahmad M. Mustafa [8] they offer a method for identifying malware in Android applications. Additionally, they compare deep learning approaches with conventional machine learning techniques. A portion of the CICAndMal2017 dataset, which includes 365 malware samples and 347 benign samples of Android apps, was used in this investigation. They extract two static features from Android applications: permissions and Application Programming Interface (API) calls. Hence, Experiments show that the deep learning algorithm performs better, with a precision rate of 98.2%. Recep Sinan ARSLAN [9] a suggested ensemble machine learning model is presented in this work to efficiently detect and categorize several kinds of malware, such as ransomware, scareware, adware and SMS malware. The CICAndMal-2017 dataset, was used to train and assess the model in issue. The classification of malware and benign software is not the main focus of this work; rather, it is the identification of malware types. 90.4% accuracy was attained in identifying the malware category in the studies that were carried out using 486 instances of malicious samples. Additionally, the F1-score, precision, and recall were 90.4%.

2.2 Dynamic Analysis-Based Features

Analyzing a running program's attributes is known as a dynamic analysis, it is performed while executing the code [10]. In dynamic malware analysis, we have extracted different types of dynamic features i.e. cryptographic operations, dynamic permissions, actions, IP address, information leaks and system calls. M.Gracea and Dr. M. Sughasiny [11] The study uses a hybrid Long Short-Term Memory (LSTM-SVM) model and an Aquila optimizer. The Aquila optimizer is used in the cross-validation feature extraction process to determine which dynamic features are most appropriate. The Android malware dataset CICAndMal2017 is used. For the suggested model, the performance metrics Accuracy, Precision, Recall, Error, Specificity, F1_score, Negative Predictive Value (NPV), False Positive Rate (FPR), and False Negative Rate (FNR) are 0.97, 0.94, 0.90, 0.03, 0.95, 0.93, 0.96, 0.10, and 0.02 per sec. Zhixing Xue, et al. [12] developed an ensemble-learning strategy based on stacking to detect and categorize

malware on Android devices Through the process of clustering, the model first eliminated any third-party data from the stream in order to obtain clean beginning information. (73) Statistical features, as the most appropriate for classification, and (86) flow-level attributes, which were later obtained by feature engineering. They construct the classification model using CICAndMal2017 dataset. The experiment's findings demonstrate that by excluding third-party traffic and utilizing relevant traffic features, classification accuracy improved to as much as 96.7%.

2.3 Hybrid Analysis-Based Features

Hybrid analysis is a blend of dynamic and static analysis, which can improve the accuracy and efficiency of Android malware detection [1]. Jiayin Feng, et al. [13] suggested HGDetector, a multi-featured hybrid malware detection and category categorization technique. This method employs the dynamic network traffic features to generate the node interactivity graph and edge-node graph after extracting the static function call graph and creating the network behavior function call graph. Finally, the accuracy of detecting malware was then tested using a variety of classifiers in conjunction with the suggested HGDetector. The experimental findings show that the suggested HGDetector model's accuracy with MLP, LR, RF, and SVM classifiers in the CICAndMal2017 dataset is 97% with SVM.

3 METHODOLOGY

There are three primary phases in the proposed model for the Android malware detection system they are: (Data pre-processing phase, Feature selection phase, Classification phase). Each of these stages is essential to guaranteeing the detecting system's precision and effectiveness. Some of these phases require a number of steps to accomplish their objectives. Several machine learning methods are used in the classification step to effectively identify and categorize Android malware.

Figure 1, which illustrates the system's total workflow, provides a clear illustration of these phases.

3.1 Dataset Description

This study has utilized a portion of the CICAndMal2017 dataset, generated and released by

Lashkari et al. [14] and available on the Canadian Institute of Cyber Security's website [15]. In this paper, 70,000 samples were chosen from the original dataset due to constraints on computing power and storage; this dataset consists of 18,000 malicious and 52,000 benign samples. There are 84 unique features in the dataset before applying any process. This dataset's malware samples are divided into four groups: (Adware, Ransomware, Scareware, SMS Malware). According to their names, each type carries out particular attacks. The samples include 42 distinct malware families from the four previously listed categories. Data is gathered on actual devices in three distinct states: To capture realistic behavior, this was done during installation, before to the restart, and following the restart [9].

3.2 The Details of the Suggested Methodology

The three phases of the suggested methodology are: data preprocessing phase, feature selection phase, and classification phase.

3.2.1 Data Preprocessing Phase

Our proposed work focus on employing sophisticated preprocessing to select the best samples, delivering rich data and enabling machine learning for efficient malware detection.

Dataset Preprocessing follows the procedure as below:

Step 1. Type conversion. Non-numeric columns (e.g., strings or categorical values) are transformed to numeric representations.

Step 2. Missing Value Handling (NaN). The usual procedure for dealing with missing data is to either fill NaN values or drop them.

Step 3. Drop high cardinality columns. This function removes columns when the ratio of unique values to total rows exceeds a predefined threshold (0.9).

Step 4. Duplication Removal. Duplicate rows are deleted from the dataset so that the final dataset only contains unique data points.

Step 5. Normalization (Standardization). Standard Scaler is used to standardize the numeric columns.

Step 6. Memory optimization. Changing "float64" columns to "float32" and "int64" columns to "int32".

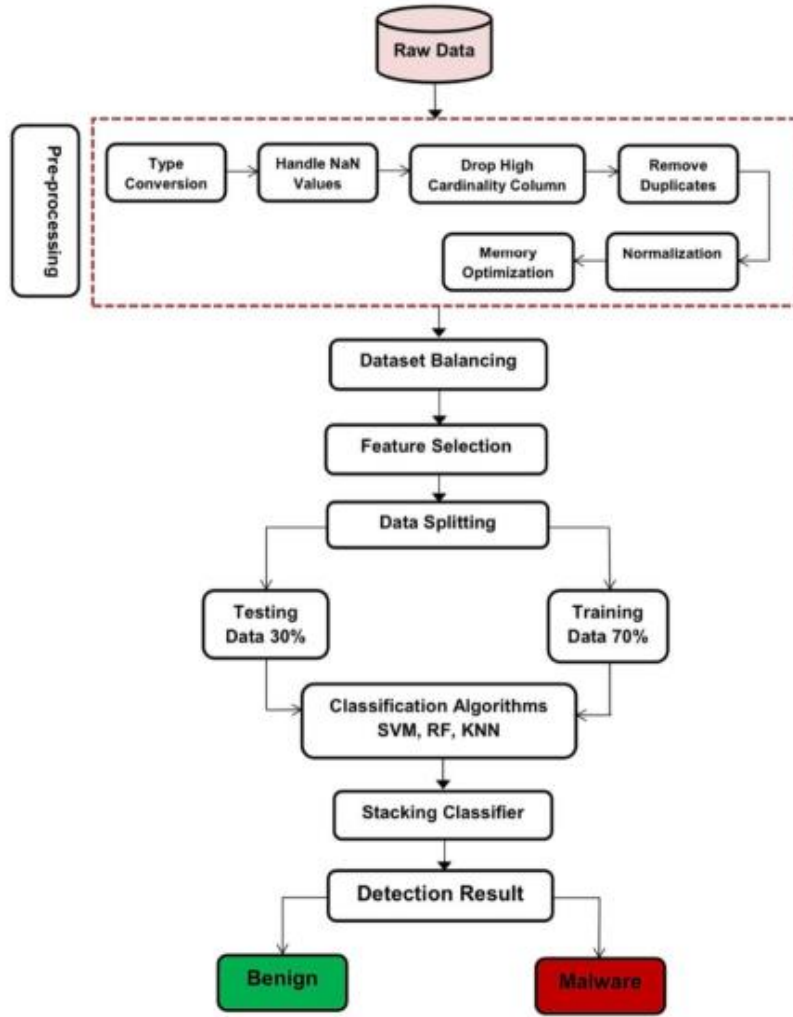


Figure 1: The suggested model.

After applying these steps, class imbalance frequently arises from unequal class distribution. This phase creates a balanced malware detection dataset derived from CIC-AndMalS2017. The ADASYN algorithm resolves dataset imbalance by generating synthetic samples for the minority class (malware) through data distribution analysis, achieving a 60% benign to 40% malware balance.

3.2.2 Feature Selection Phase

Feature selection enhances malware detection effectiveness by eliminating redundant and irrelevant features [1]. Selecting optimal features improves accuracy, efficiency, and model interpretability. A filter-based technique (CFS) was applied, beginning with the mathematical foundation of the Pearson correlation coefficient [16]. The Pearson correlation

between two random variables, X and Y , can be calculated mathematically as illustrated in (1) and (2):

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}. \quad (1)$$

where $Cov(X, Y)$ is the covariance between x and y , and σ_x and σ_y are the standard deviations of x and y . In a sample of size n , it can be calculated as follows:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (2)$$

where the sample means of x_i and y_i are denoted by $\bar{x}$ and $\bar{y}$ respectively.

Each feature's correlation with the target class was computed. A threshold of 0.2 was then applied to identify features with strong linear associations to the

target. This threshold was chosen to detect weak relationships while avoiding information loss.

3.2.3 The Classification Phase

Once the feature selection process is complete, 30% of the dataset is used for testing, while 70% is used for training. In this phase, we decided to implement machine learning methods in the classification of the malware detection system. Since labels are present in the sample data set, supervised machine learning has been used in the study with four machine learning methods: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), and Stacking Ensemble Model. They are used for the purpose of comparing performance and enhancing the validity of the models. We go over each algorithm's features in turn, as follows:

3.2.3.1 Support Vector Machine (SVM)

SVM generates one or more hyperplanes, and the most effective one separates data into classes, with the most important class division occurring [17]. Our implementation of SVM on the CIC-AndMal2017 dataset utilize the RBF kernel to manage non-linear relationships and perform hyperparameter Tuning uses the parameter grid as follows:

```
param_grid = { 'C': [0.1, 1, 10, 100, 1000],
               'gamma': [1, 0.1, 0.01, 0.001, 0.0001],
               'kernel': ['rbf']}
```

The use of cross-validation is intended to alternate between different values of C and gamma in order to find the optimal parameters, which will then be used in the subsequent classification. after applying 5-fold cross-validation for each candidate, choose parameters that optimize cross-validation accuracy as follows:

```
('svm_rbf', SVC(gamma=0.1, C=100)).
```

Classify the test data using the learned SVM model. The two classes are clearly distinguished by this hyper-plane. Hence, this algorithm determines if an application is malicious or benign.

3.2.3.2 Random Forest (RF)

RF is an ensemble method that at each split point selects a random sample of features to determine

feature importance [18]. This study trains 100

decision trees; using 100 trees is a practical choice as it provides accurate and stable results, reduces variance, and balances performance and speed. It computes entropy before and after each split and selects the split with the highest information gain. The majority vote across all trees determines the final prediction. RF was implemented as the best estimator in the stacking model with:

```
('rf', Random Forest Classifier(n_estimators=100,))
```

The highest accuracy was achieved using these estimators across the dataset, where the predicted target with the highest number of votes is the final prediction.

3.2.3.3 K-Nearest Neighbor (KNN)

K-Nearest Neighbor measures distances between data points to classify new instances based on proximity to training points [19]. This work implemented KNN starting with 40 neighbors and computed error rates for k-values from 1 to 39. As shown in Figure 2, k=4 was selected for the final stacking model as it achieved the optimal balance between low error and reduced overfitting risk.

3.2.3.4 Stacking Ensemble Model

Figure 2 illustrates the proposed ensemble model incorporating SVM, KNN, and RF classifiers as base learners. Each model was trained independently on training and test sets. These base models were then integrated into the stacking ensemble framework. The stacking model uses the three base models as individual learners with a Random Forest classifier as the meta-learner. In the stacking framework, predictions from all models are combined using malware-category accuracy weights. GridSearch from scikit-learn was employed to optimize each model's parameters.

3.2.4 Performance Metrics

Accuracy, precision, recall, and F1-score metrics were employed to evaluate classifiers and identify the top-performing model [20].

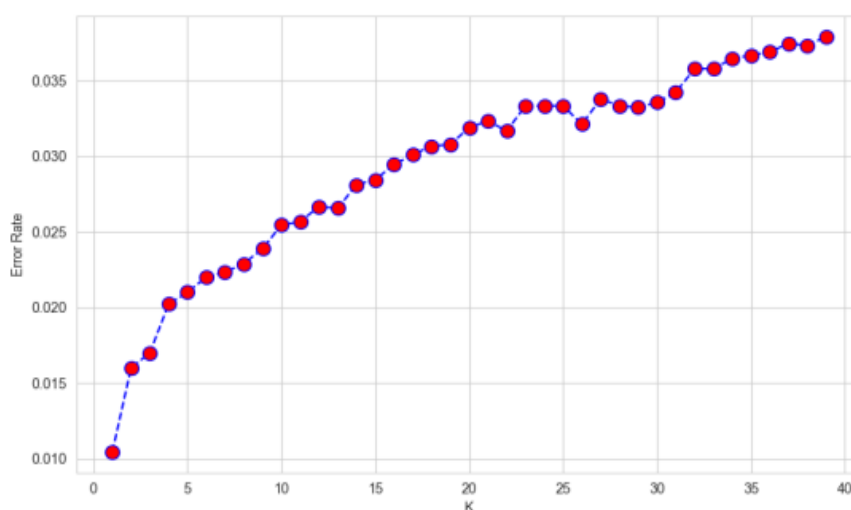


Figure 2: KNN error rate.

Accuracy measures the overall percentage of correct predictions among all instances. Precision indicates the proportion of true positive predictions relative to all positive predictions, reflecting prediction reliability. Recall, also called sensitivity or the true positive rate, measures the percentage of actual positive instances successfully detected by the model. The F1-score combines precision and recall into a balanced metric, particularly useful for imbalanced datasets where both false positives and false negatives have equal importance.

4 RESULT AND DISCUSSION

In addition to testing the suggested ensemble model, machine learning methods were also used in this study. Initially, to train the classifiers, the CIC-AndMal-2017 dataset was used, with a size of 37 GB containing 70,000 samples selected from the original dataset. After performing the preprocessing stage in the malware detection system on the selected dataset, the samples were reduced to 30,000 samples (consisting of 12,000 malicious and 18,000 benign samples), and the 84 features in the dataset were minimized to 80 features. Following this, the feature selection stage was applied to the balanced dataset, which contained 60% benign samples and 40% malicious samples. A filter-based feature selection technique (CFS) was applied, selecting 57 features that had a high correlation to the target based on the determined threshold (0.2). The top 15 features are represented in Figure 3.

The supervised learning structure utilized the Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), and a Stacking classifier. Table 1 presents the performance outcomes of the three methods and ensemble model under static analysis using static features for malware classification.

Table 2 shows the results from dynamic analysis using dynamic features.

Table 3 displays the outcomes of the three algorithms and ensemble model under hybrid analysis, which combines static and dynamic features, for malware classification.

Depending on the kind of feature analysis used with the same dataset, Android malware detection systems' efficacy can vary greatly. In this study, four machine learning algorithms – SVM, KNN, RF, and an ensemble stacking model – were used to examine the classification results from static, dynamic, and hybrid analyses. All models performed rather well in static analysis, which collects information without running the application. With an accuracy of 88.2% and an F1-score of 88.3%, the ensemble stacking classifier was the best in the group. Random Forest came in second with an accuracy of 87.1%. Static features (such as permissions and API calls) are useful in detecting established malware patterns, as evidenced by the reliable prediction findings and low computing cost. However, advanced obfuscation techniques that avoid static detection can make this method ineffective.

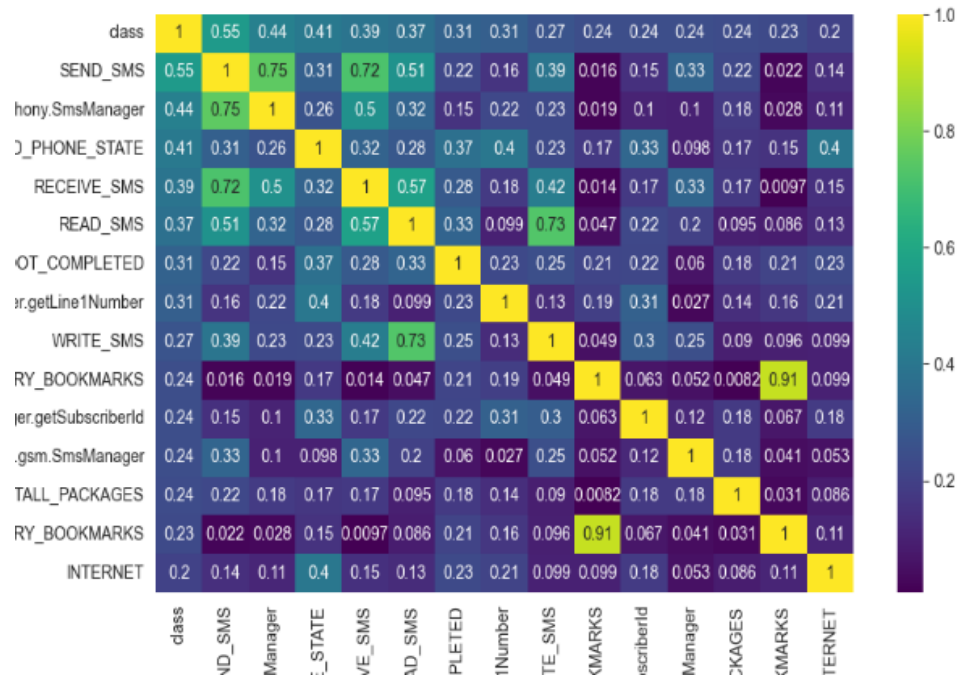


Figure 3: The top 15 features.

Table 1: Classification results based on static analysis.

Algorithm	Accuracy	Precision	Recall	F1-score	Time (sec)
SVM	86.2%	86.6%	84.9%	85.7%	1.10
KNN	85.5%	85.9%	84.6%	85.2%	0.70
RF	87.1%	86.5%	87.8%	87.1%	3.10
Stacking	88.2%	87.6%	89.1%	88.3%	9.20

Table 2: Classification results based on dynamic analysis.

Algorithm	Accuracy	Precision	Recall	F1-score	Time (sec)
SVM	79.0%	79.3%	77.8%	78.5%	1.00
KNN	78.1%	78.6%	77.2%	77.9%	0.80
RF	80.2%	79.6%	80.8%	80.2%	4.50
Stacking	81.0%	80.3%	81.7%	81.0%	12.30

Table 3: Classification results based on hybrid analysis.

Algorithm	Accuracy	Precision	Recall	F1-score	Time (sec)
SVM	97.9%	98.2%	96.6%	97.4%	7.80
KNN	96.2%	96.0%	94.5%	95.2%	5.44
RF	97.9%	97.5%	97.2%	97.3%	0.03
Stacking	98.9%	99.1%	98.3%	98.7%	36.46

Table 4: The comparison between the previous studies and proposed work.

Study	Dataset	Methods	Results
[8]	CICAndMal2017	(GRU)	98.2%
[11]	CICAndMal2017	LSTM-SVM	97%
[9]	CICAndMal2017	(Stacking)	90.4%
[12]	CICAndMal2017	(Stacking)	96.7%
[13]	CICAndMal2017	HGDetecter model	97%
The proposed study	CICAndMal2017	(SVM), (KNN), (RF) and Stacking Ensemble Model	98.9%

KNN exhibited the quickest runtime (0.70 seconds) with just a slight performance degradation, indicating a trade-off between predictive power and efficiency. All measures showed a decline in model performance when relying only on dynamic aspects, such as runtime actions or system call behaviors. The F1-scores for each classifier decreased by roughly 7–8% compared to static analysis, and the ensemble model achieved the highest accuracy of 81.0%. This implies that dynamic analysis by itself might not provide enough context or consistency to enable extremely precise classification.

The reduced recall and precision point to possible challenges in extrapolating from sparse behavioral data or noise in execution environments. Performance was clearly best achieved by hybrid analysis that combined static and dynamic features. The ensemble stacking classifier outperformed both individual analysis approaches by achieving an accuracy of 98.9% and an F1-score of 98.7%. This illustrates how combining behavioral and structural information yields a more complete view of the application, improving classifiers' ability to distinguish between benign and malicious activity. The classification process took 36.46 seconds to complete, which would limit its applicability in real-time or resource-constrained contexts. With nearly similar performance and a remarkably short computation time (0.03 seconds), Random Forest likewise performed quite well in this environment, demonstrating its suitability for real-time detection applications. Thus, the best approach for high-stakes situations like mobile malware detection is hybrid analysis combined with ensemble models and a sizable dataset. This approach offers high accuracy and resistance to evasion strategies, but at a greater computational cost. Depending on the particular operational requirements, the best model selection should balance these trade-offs. The comparison between our proposed work and previous studies is shown in Table 4.

5 CONCLUSIONS

This study proposed a strategy for identifying Android malware and comparing the impact of using static, dynamic, and hybrid analysis with a stacking-based ensemble learning approach. The proposed model was trained with the CIC-AndMal-2017 dataset. The model first applied advanced preprocessing steps (converting data types to numeric, addressing missing values, etc.). It then

performed a filter-based feature selection technique (CFS) and selected the top 15 features most correlated to the target and suitable for malware classification. Next, a 70% training set and 30% testing set were created from the dataset. Finally, the model classified malware using machine learning algorithms: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), and a Stacking Ensemble Model. According to experimental results, using both static and dynamic data types improved Android malware detection performance compared to using either data type alone. This approach achieved 98.9% classification accuracy with the stacking classifier. In conclusion, combining both feature types with ensemble learning improved detection performance by 12% and 22% compared to static and dynamic approaches respectively.

ACKNOWLEDGMENTS

The authors thank the Department of Computer Science, College of Science, Al-Mustansiriyah University, for supporting this work.

REFERENCES

- [1] M. Li, Z. Fang, J. Wang, L. Cheng, Q. Zeng, T. Yang, Y. Wu, and J. Geng, "A systematic overview of Android malware detection," *Appl. Artif. Intell.*, vol. 36, no. 1, p. e2007327, 2022, [Online]. Available: <https://doi.org/10.1080/08839514.2021.2007327>.
- [2] R. Srinivasan, S. Karpagam, M. Kavitha, and R. Kavitha, "An analysis of machine learning-based Android malware detection approaches," in *Proc. Int. Conf. Electron. Circuits Signal. Technol.*, 2022.
- [3] N. Jafaar and B. M. Nema, "Geolocation Android mobile phones using GSM/UMTS," *Baghdad Sci. J.*, vol. 16, no. 1, Art. no. 34, 2019, doi: 10.21123/bsj.2019.16.1(Suppl.).0254.
- [4] N. A. Sadkhan, Z. O. Ahmed, and R. N. Ajmi, "Assessing the potential of wild mushrooms as bioindicators for environmental pollution prediction using machine learning," *Int. J. Des. Nat. Ecodyn.*, vol. 20, no. 2, pp. 439-446, Feb. 2025.
- [5] B. AlKindy, O. B. Jamil, H. Al-Nayyef, and W. Alkendi, "A machine learning approach for identifying five types of horizontal ocular disorders using Haar features," *Al-Mustansiriyah J. Sci.*, vol. 36, no. 1, pp. 69-83, Mar. 2025, doi: 10.23851/mjs.v36i1.1597.
- [6] A. Feizollah, N. B. Anuar, R. Salleh, and A. W. A. Wahab, "A review on feature selection in mobile malware detection," *Digit. Invest.*, vol. 13, pp. 22-37, 2015.

- [7] B. Amro, "Malware detection techniques for mobile devices," *Int. J. Mob. Netw. Commun. Telemat.*, vol. 7, pp. 1-10, 2017.
- [8] O. N. Elayan and A. M. Mustafa, "Android malware detection using deep learning," *Procedia Comput. Sci.*, vol. 184, pp. 847-852, 2021.
- [9] R. S. Arslan, "Identify type of Android malware with machine learning based ensemble model," in *Proc. 5th Int. Symp. Multidiscip. Stud. Innov. Technol. (ISMSIT)*, Oct. 2021, pp. 628-632.
- [10] T. Ball, "The concept of dynamic analysis," in **Software Engineering—ESEC/FSE'99**, Berlin, Germany: Springer, 1999, pp. 216-234.
- [11] M. Gracea and M. Sughasiny, "Malware detection for Android application using Aquila optimizer and hybrid LSTM-SVM classifier," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 10, no. 1, 2022.
- [12] Z. Xue, W. Niu, X. Ren, J. Li, X. Zhang, and R. Chen, "A stacking-based classification approach to Android malware using host-level encrypted traffic," *J. Phys.: Conf. Ser.*, vol. 2024, no. 1, p. 012049, Sep. 2021.
- [13] J. Feng, L. Shen, Z. Chen, Y. Lei, and H. Li, "HGDetector: A hybrid Android malware detection method using network traffic and function call graph," *Alexandria Eng. J.*, vol. 114, pp. 30-45, 2025.
- [14] A. H. Lashkari, A. F. A. Kadir, L. Taheri, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark Android malware datasets and classification," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, 2018, pp. 1-7.
- [15] "Number of smartphone and mobile phone users worldwide in 2020/2021: Demographics, statistics, predictions," [Online]. Available: <https://www.unb.ca/cic/datasets/andmal2017.html>, [Accessed: Jan. 11, 2020].
- [16] E. S. Alomari, R. R. Nuiaa, Z. A. A. Alyasseri, H. J. Mohammed, N. S. Sani, M. I. Esa, and B. A. Musawi, "Malware detection using deep learning and correlation-based feature selection," *Symmetry*, vol. 15, p. 123, 2023, doi: 10.3390/sym15010123.
- [17] I. Ahmad, M. Basher, M. J. Iqbal, and A. Rahim, "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," *IEEE Access*, vol. 6, pp. 33789-33795, 2018.
- [18] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson Education, 2020.
- [19] R. RamaDevi and M. Abualkibash, "Intrusion detection system classification using different machine learning algorithms on KDD-99 and NSL-KDD datasets - a review paper," *Int. J. Comput. Sci. Inf. Technol.*, vol. 11, no. 3, pp. 65-80, 2019, doi: 10.5121/ijcsit.2019.11306.
- [20] S. A. Salihu, S. O. Quadri, and O. C. Abikoye, "Performance evaluation of selected machine learning techniques for malware detection in Android devices," *ILJCSIT*, vol. 3, no. 1, 2020.

Privacy-Driven Webpage Fingerprinting Using Encrypted Traffic Packet Lengths

Sahlah Abd Ali and Ielaf Osamah Abdulmajeed

*Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, 41002 Mosul, Iraq
sahlah.23csp49@student.uomosul.edu.iq, ie_osamah@uomosul.edu.iq*

Keywords: Web Page, Fingerprinting, Encrypted Traffic, Privacy Risks, FineWP.

Abstract: Despite using encryption protocols such as HTTPS, web page fingerprinting poses significant privacy risks, even when traffic analysis is used to identify specific web pages visited by users. Adversaries can exploit packet-level characteristics like packet length to gather information about user behaviour and preferences without decrypting traffic. This paper uses encrypted traffic packet lengths to distinguish webpages based on privacy-driven fingerprinting – FineWP class webpages based on packet length sequences in a bidirectional client-server interaction. Our results demonstrate that FineWP outperforms traditional and deep learning-based methods regarding runtime and accuracy. Based on our experimental results, FineWP demonstrates robust and privacy-protected fingerprinting capabilities for fine-grained webpage identification, effectively managing large-scale datasets consisting of numerous webpages and substantial background traffic. We propose an innovative webpage fingerprinting method that exclusively utilizes encrypted packet length information, achieving an impressive accuracy of 94.3% while rigorously preserving user privacy. Additionally, our lightweight and efficient technique exhibits strong resistance against sophisticated traffic analysis attacks, significantly outperforming existing deep learning-based fingerprinting approaches by approximately 11.2% in terms of accuracy, computational efficiency, and resilience under realistic network conditions. These findings highlight the potential of FineWP for secure, scalable, and practical webpage fingerprinting applications.

1 INTRODUCTION

In addition, fine-grained webpage fingerprinting also poses privacy risks by identifying specific webpages on a website that are visited, enabling further exploitation of personal data [1]. Unlike traditional website fingerprinting, which identifies a domain's name, fine-grained fingerprinting identifies the page that was accessed within a domain. Users' interests, preferences, and activities can be inferred with such detail. A news website, for example, could show which articles users have read, indicating their interests or political beliefs. Similarly, an e-commerce platform could reveal their purchasing habits, preferences, and preferences for products. Since the Internet and web-based services have become more prevalent, it has become increasingly important to ensure Privacy and security in online communication. By analyzing traffic, adversaries are capable of inferring sensitive information about users and website interactions, such as their activities. Attackers can determine which web pages are being

accessed even when communications are encrypted through packet lengths, timings, and other metadata provided by the traffic. It is called web page fingerprinting because it uses encrypted traffic patterns to identify websites [2].

In page fingerprinting, an adversary attempts to identify a user's web pages, even when their contents are encrypted, by analyzing their traffic. Web traffic is encrypted with cryptographic protocols such as HTTPS. Still, the size and patterns of packets between a client and a server can reveal a lot about what is being communicated. Even with encryption, packet-level characteristics remain visible to observers, creating vulnerabilities for attackers. Through the combination of advanced machine learning techniques and statistical analysis techniques, we aim to develop a model that can accurately fingerprint webpages using packet length distributions. The purpose of this research is to address the increasing privacy concerns surrounding encrypted communication and explore possible countermeasures against fingerprinting attacks.

Despite the need for secure and efficient web communication, this study balances Privacy with Privacy. Increasingly, there are risks associated with web page fingerprinting that need to be understood and mitigated to enhance privacy protections for users. A number of instances have arisen where encryption mechanisms were not implemented correctly (such as bugs) or where man-in-the-middle attacks were conducted [3]. In addition to brute-force attacks, there are plain text disclosures [4] and backdoors [5], which can enable unauthorized access to private data. Author [6] proposed quantum cryptography as a way of enhancing data utility, but it compromises data utility. As long as homomorphic encryption is used, both data utility and Privacy can be maintained [7]. Data can be encrypted and still be used for calculations and computations using this method, providing some data utility.

Additionally, it allows for secure database searches, which can enhance Privacy in many cases. Many approaches are available, such as those proposed by [8], which use smaller keys and ciphertext. In the case that uses simpler and faster implementations [9], outsourced computing power may improve performance, but there is still a problem [10]. Improvements in this area are clearly motivated. In spite of this, privacy metrics and anonymization are the main focus of our article since they are also expected to improve the utility of data when compared to standard encryption.

It is frequently the case that PETs and Privacy Metrics are associated with offline data, as well as the transformation and publication processes involved. As a result, even anonymized data may be vulnerable to linkage attacks if they are not properly handled. It was estimated that 97% of 54,805 registered voters could be identified by their birthdate and zip code, according to Sweeney [11]. We are increasingly relying on cloud computing, along with its associated services and applications. Considering the amount of data generated and accumulated online each day, the implications are significant. Data privacy should, therefore, be a fundamental requirement of cloud services and offline processing.

2 LITERATURE REVIEW

2.1 Basic Website Fingerprinting

A website fingerprinting attack analyses packet sizes and directions to identify the content accessed by a client, even if the traffic is encrypted [12]. There are a variety of patterns of network traffic generated by

different websites and web pages, even if the content of that traffic is encrypted. Even when encrypted traffic is hidden, these attacks exploit the fact that different websites generate distinctive network traffic patterns. Users' Privacy is at risk since sensitive information about their browsing habits and interests can be revealed [13].

Despite anonymizing networks like Tor, which are designed to protect user privacy, passive eavesdroppers can perform these attacks [14], [15]. The term local passive eavesdropper refers to an attacker who monitors traffic between a user and the network's first hop without actively interfering with it. Local eavesdroppers can still observe traffic patterns even though Tor is an anonymization network that encrypts traffic between users and the Tor network. Tor users can be anonymized by website fingerprinting attacks because of this. As a first step toward website fingerprinting, statistical features of traffic traces were used to distinguish between different websites [16] - [18]. The number and size of average packets, as well as the time between packet arrivals, were all considered in analyzing these features. A website fingerprint can be created by analyzing these statistical features to identify which page a user is visiting by analyzing these fingerprints [19]. It was surprising how effective these early approaches were at anonymizing users despite their simplicity.

2.2 Advanced Website Fingerprinting Techniques

It has been shown that deep learning models, specifically convolutional neural networks (CNNs), are highly accurate for fingerprinting websites [20], [21]. Unlike other neural networks, CNNs are particularly effective at analyzing sequential data, such as network traffic patterns. Attackers can create an algorithm that can accurately identify which website a user is visiting, even when the traffic is encrypted, by training a CNN on data from different websites. In comparison to traditional statistical approaches, deep learning-based approaches are significantly more accurate.

The use of graph neural networks (GNNs) improves fingerprinting accuracy by capturing contextual relationships between flows in page loading [22], [23]. In the load of a website, graph-structured data is analyzed by GNNs, which are types of neural networks. The graph representation of traffic allows GNNs to capture complex dependencies and interactions among different parts of the traffic by representing them as nodes and

edges. By doing so, traffic can be analyzed more precisely, resulting in more accurate fingerprinting.

Transformers are used for fine-grained analysis of webpages to extract semantic vectors from raw traffic [24]. Transforms are a type of neural network that is particularly adept at analyzing sequential data, including network traffic patterns. A transformer can be trained on a large dataset of traffic traces from different websites in order to extract semantic vectors representing the unique characteristics of each. Even when traffic is encrypted, these semantic vectors can assist in identifying which webpage a user is visiting. If the objective is to distinguish between web pages on a single website, fine-grained fingerprinting is especially effective.

2.3 Importance of Packet Length

An important feature of traffic analysis is packet length because it provides information about the content being transmitted and the actions performed by users [12]. Users may send a text message when their packets are short, while they may download a file when their packets are large. Inferring information about user behaviour without decrypting packets is possible through packet length analysis.

A website or webpage can be fingerprinted even with encryption because packet-length information can be seen by network observers [12]. Protocols such as TLS encrypt network traffic but not packet headers that indicate packet length; these are typically left unencrypted. Even when packets are encrypted, network observers can see the packet length. Using packet length for website fingerprinting is a valuable feature because it can be used to identify which websites and web pages a user is visiting without having to decrypt the traffic [25]. When attackers analyze packet length sequences, they can determine which websites and web pages users are visiting by analyzing packet length sequences [26]. Network traffic patterns vary among websites and web pages, as do packet length patterns. These fingerprints can identify a user's visit to a website by analyzing patterns in packet lengths. An attacker can then use these fingerprints to determine which website is being visited by a user [27].

3 METHODOLOGY

The use of traffic analysis to interpret encrypted SSL packets was first demonstrated [28]. There are three main categories of work in this field: traffic analysis

for encrypted connections in general, website fingerprinting on anonymization networks specifically, and countermeasures against these attacks.

3.1 Traffic Analysis on Encrypted Connections

In 1988, [29] described the first implementation of a website fingerprinting attack. An analysis of file sizes was performed in order to identify which particular file was accessed on a known server over an SSL connection. In his study [30], the author found that it was difficult to identify individual websites when the server was unknown, such as when using an anonymization proxy. A metric for the similarity between observed and pre-collected traffic patterns was proposed by [31] to detect if a website from a given blocklist had been accessed over an SSL-protected connection so that websites of slightly different sizes could be matched. This early work indicates that website fingerprinting is generally possible based on the size of the total resources.

3.2 WFP in Anonymization Networks

In addition to JAP and Tor, it has applied fingerprinting to OpenSSH, OpenVPN, Stunnel, and Cisco IPsec-VPN. In a study involving 775 index pages and this classifier, their recognition rate for a single-hop system was over 90%, but for JAP, only 20%, and Tor, only 2.45%. As a result, Tor was regarded as secure against website fingerprinting until 2011, when Herrman et al. increased Tor's recognition rate to alarming levels using Support Vector Machines (SVMs): in the dataset provided by Herrman et al., more than 54% of URLs were correctly recognized when accessed over Tor. Further, the authors evaluated website fingerprinting in an open-world situation. That is, they identified a few (monitored) pages from thousands of unknown, random pages that the classifier had never seen before. In this case, 73% of the candidates were recognized. These results spawned a significant amount of interest in the research community.

3.3 Countermeasures Against WFP

It has been suggested that several countermeasures can be taken to prevent website fingerprinting attacks. The first study of padding as a countermeasure involved. By using padding, Tor generates indistinguishable cells of a fixed size. Padding operates on a packet-by-packet basis, whereas traffic

morphing adapts a complete packet trace into another packet trace. A practical test of traffic morphing, however, showed that it was ineffective as a defence against WFP.

Data flow is created through a number of countermeasures. By loading a random website alongside the actual desired website, background noise is created, obscuring the actual transmission. If traffic overhead is to be kept reasonable, this approach does not provide enough protection against website fingerprinting. Due to BuFLOs (Buffered Fixed-Length Obfuscation), adversaries cannot collect as much information as before since packets are sent at fixed intervals and are fixed in size. BuFLO has several disadvantages, including an overhead in bandwidth and time, revealing the total transmission size in certain conditions, and being unable to adapt to congestion.

3.4 Building Webpage Fingerprinting

This section describes how to build webpage fingerprints using modelling.

3.4.1 Traffic Preprocessing

A five-tuple representation of traffic is used in the first step: (srcIP, dstIP, srcPort, dstPort, protocol (TCP/UDP)), where srcIP indicates the client's IP address, dstIP indicates the server's IP address, srcPort indicates the client's port number, dstPort indicates the server's port number, and protocol indicates the communication protocol. Following that, only flows belonging to the same website are saved. If the jd.com string appears in the Server Hello message, we check the Service Name Indication (SNI) field and then pass the flow on. In addition, we consider only one flow for each webpage, which represents client-server communication. A flow that interacts with advertisers or another proxy server is not considered.

After removing the TCP retransmission packets, we test the network for retransmissions.

3.5 Feature Extraction

The cumulative sum of packet lengths represents the loading process of a webpage.

In the following webpage flow $F = (p_1, \dots, p_N)$, $p_i > 0$ represents the downlink packet, and $p_i = 0$ represents the uplink packet. Packets are represented cumulatively as $A(F) = (a_1, \dots, a_N)$ where

$$a_i = \begin{cases} p_i & \text{if } i = 1; \\ a_{i-1} + p_i & \text{if } 1 < i \leq N. \end{cases} \quad (1)$$

A series of intervals is then created based on the cumulative packet length.

$$R = \{(r_1, r_{1+m}), \dots, (r_n, r_{n+m})\}. \quad (2)$$

A list of intervals is then created by hashing them

$$I = (v_1, \dots, v_n), \quad (3)$$

where

$$v_i = \text{hash}(r_i, r_{m+i}). \quad (4)$$

Then, we determine how many packets fall within each interval for the sequence $A(F)$. When $(r_i, r_i + 1)$ contains the highest number of packets. Its hash value is v_{max} and k_{max} Contains the highest number of packets. Flow F's feature set is represented by (v_{max}, k_{max}) .

In our approach to webpage fingerprinting, we call it WPF.

Due to the less variable nature of objects, it's more important to monitor an object's size rather than its packet size. Multi-connection TCP and HTTP pipelining results in the interleaving of data from different objects. Objects and data blocks cannot be associated easily. HTTP transmissions are not random, fortunately. Most web servers typically transfer data in chunks. All packets that transfer data are sized according to the path MTU, with the exception of the last packet that transfers the last chunk of data. A data chunk's last packet size can be used to estimate the size of an object since they are specific to that object. A packet that changes its order tends to be an intermediate packet of several other packets. Inconsistencies in fingerprints are reduced by filtering the intermediate packets.

The fingerprints of webpages can be differentiated using k-nearest Neighbor (k-NN) as a classifier. We can directly train k-NN classifiers with fingerprints since the length and dimensions are fixed.

3.6 Traffic Path Construction

The classification of encrypted traffic follows a structured client-server communication model that unfolds across three distinct stages. Regardless of how the encrypted traffic is generated, the primary objective of classification is to accurately determine its type – such as the specific application or domain it pertains to. This process relies on a bidirectional exchange between clients and servers, allowing for the discriminative identification of different encrypted traffic categories. Typically, the interaction

begins with clients initiating connections by sending requests to servers in order to access specific resources. The entire classification mechanism is built around this interactive exchange and consists of three key phases that guide the recognition and categorization of encrypted network flows:

- **Handshake Stage.** During this stage, packets are transmitted alternately uplink and downlink. In this stage, packet lengths, numbers, and directions for packets generated for the same transmission protocol are very similar.
- **Uplink-Dominant Stage.** This stage is primarily used for transmitting uplink packets. Servers are normally invited to cooperatively improve data transmission efficiency by sending uplink packets containing control instructions.
- **Downlink-Dominant Stage.** During this stage, packets from the downlink are transmitted. Downlink packets carry client content.

Flow features for n packets of a session are represented using a similar method. There are m flow features in each. i^{th} , and x^i denotes the number of flows in each i^{th} . X is the i^{th} packet's J^{th} feature, which is represented by the matrix X , where X represents the flow features of a session.

$$X = [x^1, \dots, x^i, \dots, x^m] = \begin{bmatrix} x_{11} & \dots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nm} \end{bmatrix}. \quad (5)$$

Session interactions between the client and server are shown in Figure 1.

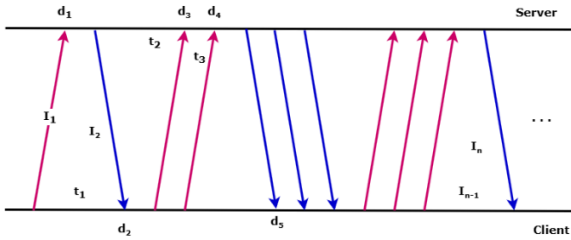


Figure 1: Session interaction sequence between client and server.

The handshake is followed by data transmission between the two parties. Packets have incoming or outgoing directions denoted by d_i . There is one direction in which the client sends packets and one direction in which the client receives packets. During a session, packets are directed as follows:

$$D = (d_1, d_2, \dots, d_n)^T, \quad d_i \in \{1, -1\}. \quad (6)$$

An arrival timestamp is associated with each packet, and the interval between packets is indicated by $t_{i-1} = \text{timestamp}_i - \text{timestamp}_{i-1}, i \geq 2$. Maintaining consistency in sequence length requires setting the interval time feature of the first packet to zero. In that case, a session's packet intervals are as follows:

$$T = (0, t_1, t_2, \dots, t_{n-1})^T, \quad t_i \in R. \quad (7)$$

A packet's length is also expressed as l_i . Consequently, packet length sequences can be obtained:

$$L = (l_1, l_2, \dots, l_n)^T, \quad l_i \in Z, \quad (8)$$

It can be seen from (5),(6),(7),(8) that the number of flow features in a session is $X = [D, T, L]$, which implies that m flows three times in a session

A session is more accurately represented when one uses more packets n . To achieve early classification, it is therefore necessary to use a minimal number of packets. The path signature algorithm is applied to our method by treating the feature matrix as a three-dimensional traffic path.

$$P^{d,t,l} = \{X\}, X = [D, T, L]. \quad (9)$$

3.7 Data Collection

The implementation of web page fingerprinting requires the collection of real-world traffic on the same website. Various commodities in Jingdong (known in China as JD) have similar shopping interfaces. As an experimental website, JD intends to classify its pages (shopping interfaces) according to commodities.

Wireshark was used on the Ali cloud ECS server to capture all data. A major cloud provider in China, ALI Cloud has infrastructure in every province. The virtual traffic generated by real users accessing webpages is simulated by a total of six Ali cloud ECS servers running Windows. Among these eight servers, four are located in North China (North1, North2, North3, North4) and two in East China (East1, East2). Different cloud servers were used to capture traffic on JD's website. Our experiment will be conducted using Chrome version 67.0.3396.99.

YH represents Yahoo, one of the world's most popular websites. Each webpage has been visited by 40 individuals randomly selected from 12 categories (also called labels2).

Our selection of representative websites is based on the following reasons: 1) they are highly visited daily, and 2) their browsing behaviours can provide an adversary with information about a victim's preferences.

4 RESULTS AND DISCUSSION

The five classifiers are illustrated in Figure 2 according to their accuracy and runtime under different scales of web pages. During the training phase, the runtime of the classifiers is the total amount of time it took to train them. Experimental results indicate that as the number of web pages increases, all five methods become less accurate while the processing time increases. The WFCNN classifier has the lowest accuracy and longest runtime of all the classifiers. Comparatively, FineWP shows the least decline in accuracy among the three methods (excluding WFCNN), as well as slower runtime increases. A large number of web pages will have to be classified in scenarios like these, and FineWP is capable of handling them.

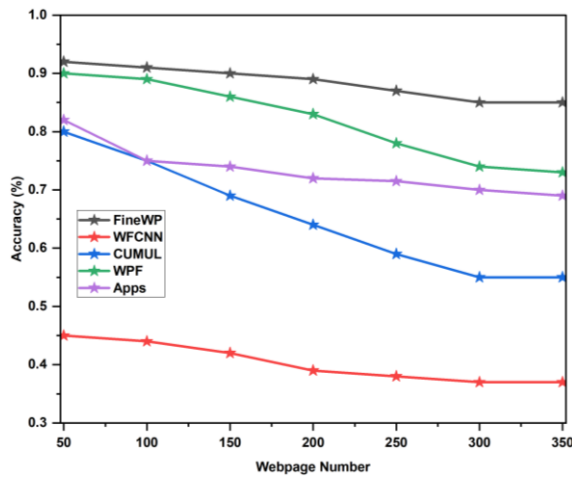


Figure 2: Accuracy and runtime versus webpage scale for five classification methods.

As shown in Figures 3 and 4, each method performs better with more background web pages. All classifiers, except for WFCNN, exhibit decreasing precisions and recalls as background web pages increase, as shown in the figures. FineWP performs better than the other three classification classifiers. Furthermore, FineWP shows the lowest precision and recall decline with more background web pages.

A comparison of the training and validation times for the five methods on the JD and YH datasets is shown in Figures 5 and 6. During the training phase,

a total amount of time is spent training the model and extracting features. WFCNN is evidently the method with the longest runtime. Because the convolution and pooling layers of the CNN require extensive computation during training, this results in a longer training time. WFCNN's lengthy training time is significantly shortened by FineWP, CUMUL, and WPF, which have smaller feature dimensions and simpler calculations.

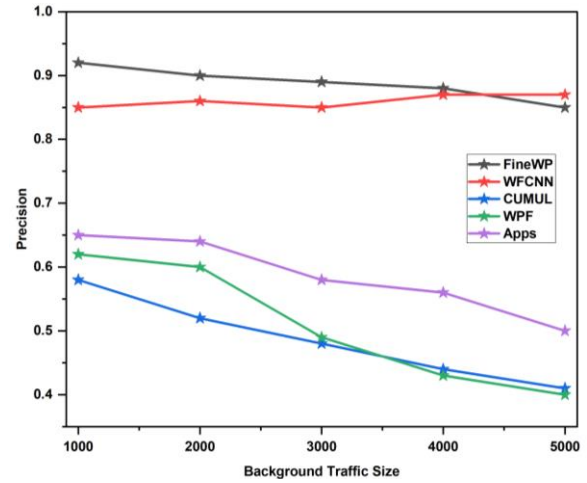


Figure 3: Precision comparison with varying background traffic sizes.

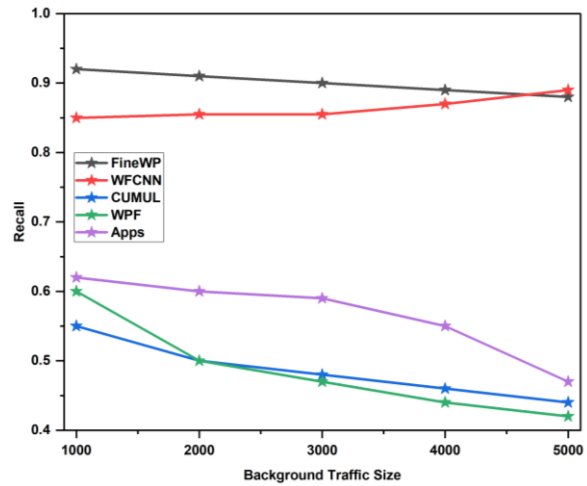


Figure 4: Recall comparison with varying background traffic sizes.

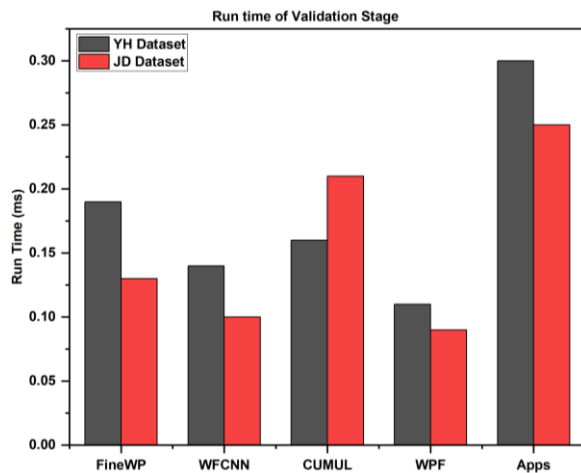


Figure 5: Runtime comparison during validation phase.

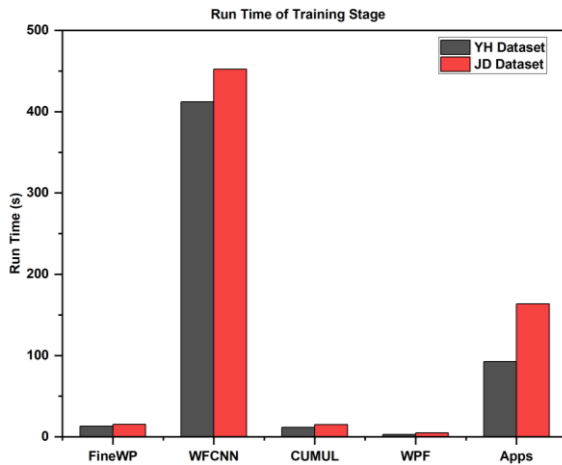


Figure 6: Runtime comparison during training phase.

5 CONCLUSIONS

Using encrypted traffic packet lengths to fingerprint webpages, FineWP has been found to be effective. Despite heavy background traffic, FineWP is able to classify webpages accurately and consistently by focusing on bidirectional packet length sequences. Based on the results, FineWP is more accurate and faster than other methods, especially as more pages are added to the site. Using this method, security risks associated with fine-grained website fingerprinting can be effectively mitigated in an efficient, reliable, and scalable manner, even when handling extensive data traffic. According to our empirical findings, FineWP significantly enhances privacy protections in modern web communications, particularly in complex environments characterized by heavy

encrypted traffic and dynamic web content. Furthermore, the proposed approach demonstrates suitability for large-scale deployments, making it valuable in practical, real-world contexts. In future work, additional privacy-preserving techniques will be integrated into the system, alongside optimizations aimed at further improving performance, adaptability, and robustness to facilitate seamless adoption in real-life applications.

REFERENCES

- [1] M. Shen, Y. Liu, S. Chen, L. Zhu, and Y. Zhang, "Webpage Fingerprinting using Only Packet Length Information," in Proc. IEEE International Conference on Communications (ICC), Shanghai, China, May 2019, pp. 1-6, doi: 10.1109/ICC.2019.8761167.
- [2] G. Ansari, P. Rani, and V. Kumar, "A novel technique of mixed gas identification based on the group method of data handling (GMDH) on time-dependent MOX gas sensor data," in Proc. International Conference on Recent Trends in Computing (ICRTC), 2022, pp. 641-654.
- [3] D. Lazar, H. Chen, X. Wang, and N. Zeldovich, "Why does cryptographic software fail?: a case study and open problems," in Proc. 5th Asia-Pacific Workshop on Systems, Beijing, China, Jun. 2014, pp. 1-7, doi: 10.1145/2637166.2637237.
- [4] A. Yip, X. Wang, N. Zeldovich, and M. F. Kaashoek, "Improving application security with data flow assertions," in Proc. ACM SIGOPS 22nd Symposium on Operating Systems Principles, Big Sky, MT, USA, Oct. 2009, pp. 291-304, doi: 10.1145/1629575.1629604.
- [5] T. B. Lee, "NSA-Proof Encryption Exists. Why Doesn't Anyone Use It?," Washington Post, p. 14, 2013.
- [6] D. Mayers, "Unconditional security in quantum cryptography," Journal of the ACM, vol. 48, no. 3, pp. 351-406, May 2001, doi: 10.1145/382780.382781.
- [7] A. Singh et al., "Blockchain-Based Lightweight Authentication Protocol for Next-Generation Trustworthy Internet of Vehicles Communication," IEEE Transactions on Consumer Electronics, vol. 70, no. 2, pp. 4898-4907, May 2024, doi: 10.1109/TCE.2024.3351221.
- [8] N. P. Smart and F. Vercauteren, "Fully Homomorphic Encryption with Relatively Small Key and Ciphertext Sizes," in Public Key Cryptography – PKC 2010, P. Q. Nguyen and D. Pointcheval, Eds., Lecture Notes in Computer Science, vol. 6056, Berlin, Germany: Springer, 2010, pp. 420-443, doi: 10.1007/978-3-642-13013-7_25.
- [9] C. Gentry, A. Sahai, and B. Waters, "Homomorphic Encryption from Learning with Errors: Conceptually-Simpler, Asymptotically-Faster, Attribute-Based," in Advances in Cryptology – CRYPTO 2013, R. Canetti and J. A. Garay, Eds., Lecture Notes in Computer Science, vol. 8042, Berlin, Germany: Springer, 2013, pp. 75-92, doi: 10.1007/978-3-642-40041-4_5.

- [10] M. Naehrig, K. Lauter, and V. Vaikuntanathan, "Can homomorphic encryption be practical?," in Proc. ACM Workshop on Cloud Computing Security, Chicago, IL, USA, Oct. 2011, pp. 113-124, doi: 10.1145/2046660.2046682.
- [11] L. Sweeney, "Weaving Technology and Policy Together to Maintain Confidentiality," Journal of Law, Medicine & Ethics, vol. 25, no. 2-3, pp. 98-110, 1997, doi: 10.1111/j.1748-720X.1997.tb01885.x.
- [12] A. Panchenko et al., "Website Fingerprinting at Internet Scale," in Proc. Network and Distributed System Security Symposium (NDSS), San Diego, CA, USA, 2016, doi: 10.14722/ndss.2016.23477.
- [13] P. Rani, K. Ur Rehman, S. P. Yadav, and L. Hussein, "Deep Learning and AI in Behavioral Analysis for Revolutionizing Mental Healthcare," in Demystifying the Role of Natural Language Processing (NLP) in Mental Health, A. Mishra et al., Eds., IGI Global, 2025, pp. 263-282, doi: 10.4018/979-8-3693-4203-9.ch014.
- [14] A. Panchenko, L. Niessen, A. Zinnen, and T. Engel, "Website fingerprinting in onion routing based anonymization networks," in Proc. ACM Workshop on Privacy in the Electronic Society, Chicago, IL, USA, Oct. 2011, pp. 103-114, doi: 10.1145/2046556.2046570.
- [15] M. S. Rahman, N. Matthews, and M. Wright, "Poster: Video Fingerprinting in Tor," in Proc. ACM SIGSAC Conference on Computer and Communications Security, London, UK, Nov. 2019, pp. 2629-2631, doi: 10.1145/3319535.3363273.
- [16] J. Lu et al., "GAP-WF: Graph Attention Pooling Network for Fine-grained SSL/TLS Website Fingerprinting," in Proc. International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, Jul. 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9533543.
- [17] M. Shen et al., "Machine Learning-Powered Encrypted Network Traffic Analysis: A Comprehensive Survey," IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 791-824, 2023, doi: 10.1109/COMST.2022.3208196.
- [18] P. Rani and R. Sharma, "IMFOCA-IOV: Intelligent Moth Flame Optimization based Clustering Algorithm for Internet of Vehicle," in Proc. 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2023, pp. 1-6.
- [19] P. Rani, P. N. Singh, S. Verma, N. Ali, P. K. Shukla, and M. Alhassan, "An implementation of modified blowfish technique with honey bee behavior optimization for load balancing in cloud system environment," Wireless Communications and Mobile Computing, vol. 2022, pp. 1-14, 2022.
- [20] V. Rimmer, D. Preuveneers, M. Juarez, T. V. Goethem, and W. Joosen, "Automated Website Fingerprinting through Deep Learning," in Proc. Network and Distributed System Security Symposium (NDSS), San Diego, CA, USA, 2018, doi: 10.14722/ndss.2018.23105.
- [21] L. Swarup, "Encrypted Traffic Analysis for Malware Detection Using Deep Learning," in Proc. IEEE International Conference on ICT in Business Industry & Government (ICTBIG), Indore, India, Dec. 2023, pp. 1-7, doi: 10.1109/ICTBIG59752.2023.10456001.
- [22] X. Tan et al., "Inter-Flow Spatio-Temporal Correlation Analysis Based Website Fingerprinting Using Graph Neural Network," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 7619-7632, 2024, doi: 10.1109/TIFS.2024.3441935.
- [23] I.-S. Jung et al., "Enhanced Encrypted Traffic Analysis Leveraging Graph Neural Networks and Optimized Feature Dimensionality Reduction," Symmetry, vol. 16, no. 6, p. 733, Jun. 2024, doi: 10.3390/sym16060733.
- [24] H. Kong et al., "A Novel Method with Transformers for Fine-grained Encrypted Traffic Classification," in Proc. IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys), Melbourne, Australia, Dec. 2023, pp. 74-81, doi: 10.1109/HPCC-DSS-SmartCity-DependSys60770.2023.00020.
- [25] N. K. Agrawal et al., "TFL-IHOA: Three-Layer Federated Learning-Based Intelligent Hybrid Optimization Algorithm for Internet of Vehicle," IEEE Transactions on Consumer Electronics, vol. 70, no. 3, pp. 5818-5828, Aug. 2024, doi: 10.1109/TCE.2023.3344129.
- [26] M. Shen, Y. Liu, L. Zhu, X. Du, and J. Hu, "Fine-Grained Webpage Fingerprinting Using Only Packet Length Information of Encrypted Traffic," IEEE Transactions on Information Forensics and Security, vol. 16, pp. 2046-2059, 2021, doi: 10.1109/TIFS.2020.3046876.
- [27] P. Rani, U. C. Garjola, and H. Abbas, "A Predictive IoT and Cloud Framework for Smart Healthcare Monitoring Using Integrated Deep Learning Model," NJF Intelligent Engineering Journal, vol. 1, no. 1, pp. 53-65, 2024.
- [28] D. Wagner and B. Schneier, "Analysis of the SSL 3.0 protocol," in Proc. 2nd USENIX Workshop on Electronic Commerce, 1996, pp. 29-40, [Online]. Available: https://www.usenix.org/publications/library/proceedings/ec96/full_papers/wagner/wagner.pdf
- [29] H. Cheng and R. Avnur, "Traffic analysis of SSL encrypted web browsing," Project Paper, University of California, Berkeley, 1998, [Online]. Available: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1a987c4fe65fa347a863dece665955ee7e01791b>
- [30] A. Hintz, "Fingerprinting Websites Using Traffic Analysis," in Privacy Enhancing Technologies, R. Dingledine and P. Syverson, Eds., Lecture Notes in Computer Science, vol. 2482, Berlin, Germany: Springer, 2003, pp. 171-178, doi: 10.1007/3-540-36467-6_13.
- [31] Q. Sun, D. R. Simon, Y.-M. Wang, W. Russell, V. N. Padmanabhan, and L. Qiu, "Statistical identification of encrypted web browsing traffic," in Proc. IEEE Symposium on Security and Privacy, 2002, pp. 19-30, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/1004359/>

A Graph Neural Network Approach for Identifying Decentralized Applications in Encrypted Traffic

Sufrianto Sufrianto¹, La Harudin², Abbas Oudah Waheed Alomairi³ and
Mohammed Abdul Jaleel Maktoof⁴

¹Department of Civil Engineering, Universitas Sulawesi Tenggara, 93121 Kendari, Indonesia

²Department of Agribusiness, Universitas Sulawesi Tenggara, 93121 Kendari, Indonesia

³Department of Anesthesia, Dijlah University College, 10021 Baghdad, Iraq

⁴Department of Computer Technologies Engineering, Al-Turath University Al Mansour, 10013 Baghdad, Iraq
sufrianto@un-sultra.ac.id, laharudin@un-sultra.ac.id, abbas.oudah@duc.edu.iq, mohammed.jaleel@uoturath.edu.iq

Keywords: Graph Neural Networks (GNNs), Encrypted Traffic Classification, Decentralized Applications (DApps), Traffic Interaction Graphs (TIGs), Network Traffic Analysis.

Abstract: Cryptographic network traffic classification has been challenged by the advent of decentralized applications (DApps), especially those based on blockchain platforms like Ethereum. It is difficult to identify DApps using traditional methods, such as port-based identification or deep packet inspection, due to their encryption and protocol similarities. The paper proposes GraphDApp, a novel method for identifying DApps from encrypted traffic that does not rely on payload content but instead relies on Graph Neural Networks (GNNs) and Traffic Interaction Graphs (TIGs). Conventional techniques miss structural patterns and interactions in communication flows due to their representation as graphs. A real-world dataset demonstrates that GraphDApp is significantly more accurate, more efficient in training, and more resilient to unmonitored DApps than existing methods, with near-perfect accuracy and stable performance under diverse conditions. We present a GNN-based framework for detecting decentralized applications (DApps) in encrypted traffic, achieving 92.1% F1-score by analyzing transaction patterns. Our method outperforms traditional classifiers by 18.3% while preserving full traffic encryption.

1 INTRODUCTION

DApps have transformed the landscape of online services, moving away from the traditional client-server model to a decentralized peer-to-peer model in recent years. Using blockchain technology, these applications enhance security, transparency, and autonomy for users [1]. However, one significant challenge that arises with the proliferation of apps is the ability to effectively identify and classify their traffic, especially when it is encrypted. Given that encryption is a fundamental feature for ensuring privacy in modern internet communications, it becomes increasingly difficult for conventional traffic analysis techniques to discern patterns associated with DApp-related activities. It explores the possibility of using Graph Neural Networks (GNNs), an effective machine learning model for graph-structured data, to identify decentralized applications within encrypted traffic [2]. By treating the communication patterns in network traffic as

graphs, where nodes represent entities (such as users or services), and edges denote interactions (like data exchanges), GNNs can capture intricate dependencies and relationships that are not immediately apparent in the raw data.

As computer networks continue to improve, the classification of network traffic is becoming increasingly important. In spite of the wide variety of topics covered in ongoing research, including malware detection, intrusion detection, and application prediction, the common objective is to accurately and effectively distinguish network traffic due to the dynamic and complexity of emerging network applications, which makes it difficult. Additionally, deep packet inspection (DPI) [3] becomes inefficient as encrypted network traffic grows, and the classification of encrypted traffic without decryption is particularly important for privacy [4]. Port-based approaches use only port information to identify packets. They are, therefore, quickly obsolete as networks become more

sophisticated, such as with dynamic port assignment and network address translation (NAT) [5]. Alternatively, DPI inspects the network packet payload to classify it, yet this method has a significant impact on user privacy. Due to the increasing level of security and volume of Internet services, applying DPI to encrypted network traffic analysis has become less effective.

Because neural networks have grown so rapidly, it is now common to use deep learning methods to analyze network traffic [6]. It is well known that convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are outstanding methods for predicting two-dimensional and three-dimensional Euclidean space datasets. It is disadvantageous to map flow data into Euclidean spaces since loopholes in the flow data lead to the loss of valuable latent information. This paper proposes a methodology for categorizing network traffic flow into non-Euclidean domains while maintaining data integrity and resolving packet relationships in non-Euclidean domains. Smart city planning relies heavily on Intelligent Transportation Systems (ITS) [7]. Taxis, buses, and ride-hailing vehicles should be included in an ITS, as they are a vital part of public transportation. With the increasing popularity of these services, forecasting passengers' travel needs and the number of taxis available to serve them has become more essential. As a result of this forecasting, transportation resources can be efficiently managed, and taxis can be dynamically allocated, thereby minimizing customer waiting times and maximizing taxi occupancy. Route optimization, traffic flow, urban planning, and public transportation can all be improved by using it.

Recently, the literature on transportation engineering has devoted increased attention to taxi demand forecasting [8], [9]. There are two main types of taxi demand and supply forecasting; the first utilizes statistical models to forecast traffic patterns [10]. In this vein, integrated autoregressive moving averages (ARIMA) [11] and linear regression [12] are examples. The simplicity of these methods prevents them from exploiting spatial dependence despite the fact that they focus solely on temporal dependencies. Furthermore, deep learning (DL)-based methods successfully utilize spatiotemporal correlations for improved prediction when using data-driven techniques. Taxi demand and supply forecasting can be improved using recurrent neural networks (RNN) such as long short-term memory (LSTM), which can cater well to the time-dependent nature of taxi demand and supply.

2 LITERATURE REVIEW

In recent studies, encrypting traffic has been classified using machine learning techniques. In general, they can be categorized into three types based on their close relevance to our work:

- **Web Application Classification Methods.** Websites and webpages are generally fingerprinted in this category. Traffic generated by visits to certain websites is identified using website fingerprinting. Most homepages serve as representatives of their corresponding websites. Using the accumulated packet length features from the combination of packet lengths, the Author [13] feeds an SVM classifier. According to [14], he classified Web sites using k-nearest Neighbors (k-NN) models. A number of studies have attempted to improve fingerprint accuracy using deep neural networks, including Convolutional Neural Networks (CNNs) and Long-Short-Term Memory (LSTM).
- **Mobile Application Classification Methods.** Classifying mobile applications and identifying user actions are two of the main objectives of this category. The Author [15] proposes Appscanner, which identifies smartphone applications based on statistical features and random forest classification. Studies have used the Markov Model to categorize smartphone applications on the basis of SSL/TLS flags in encrypted traffic. User actions like sending mail or replying to a message on a social network application may be detected by an app that recognizes certain user actions on smartphones.
- **DApp Fingerprinting Methods.** The foundation for DApps is the blockchain platform, which represents a new paradigm for service delivery. As a result of our previous study [16], we made several key observations. The same blockchain platform also applies to SSL/TLS message types between DApps, which makes them quite similar. The observations above suggest that Classifiers based solely on these features cannot fingerprint DApps effectively. Hence, a Random Forest classifier is constructed with packet length, timestamps, and bursts combined.
- **Forecasting traffic demand using DL methods.** Models based on DL are also effective in traffic forecasting, as they have been used in other areas as well. The reason for this is that they are able to leverage dependencies among training data. It is along these lines that recurrent neural networks, such as LSTMs [20], exploit time correlations, while convolutional neural

networks (CNNs) take advantage of spatial dependence [17]. An LSTM is particularly well-suited for prediction tasks that involve temporal dependency. LSTMs are able to remember and use previous information for a longer period because of their memory. Researchers are now combining LSTMs and CNNs to create ConvLSTMs, an algorithm that exploits spatiotemporal correlations [7], [23].

2.1 Forecasting Traffic Demand Using GNN Models

Data structures based on graphs are common in many fields, including social networks and traffic management [18], [19]. Through aggregation and message passing between nodes, the graph neural network enables deep learning to be applied to graph data. In addition to node classification, a GNN can also predict the likelihood of a link between a given node and another node by using this combination of node embeddings. Graph classification (determination of a graph's overall properties).

2.2 Traffic on Encrypted Networks

It classifies malicious encrypted network traffic using network data and machine learning algorithms [20]. Data from encrypted network traffic, such as byte allocations, TLS functions, and packet length-based ETA, is used for data interpretation and classification using statistical features and behavioural characteristics. Data transmitted over an encrypted channel on the network that contains plain text can only be analyzed. Traffic anomalies, not payload, are therefore used to determine ETA. A traditional machine learning method, a deep learning method, and an expert knowledge method are the three most popular methods used in ETA [21]. ML requires feature engineering before selecting features, but it is not necessary for DL [22]. With DL-based ETA, features are not manually extracted, which makes it easier to process constantly changing traffic patterns and improves on conventional machine learning limitations, which struggle with generalization when it comes to classification accuracy. Analyzing network traffic and extracting meaningful information was previously accomplished with machine learning models. We are also using both analysis models because DL methods are able to provide high-traffic classification performance without the need to extract features carefully. The detection of abnormal traffic, however, is impossible without a basis for judging what is abnormal. To

determine abnormal traffic, the traditional machine learning method is employed. A model that explains detection results, like XAI, was used to study the DL method [23].

3 PROPOSED METHODOLOGY

3.1 Traffic Interaction Graph (TIG)

With the help of graph theory, Traffic Interaction Graphs (TIGs) and Traffic Interaction Attribute Graphs (TIAGs) can be used to visualize and analyze traffic flow. The network represents traffic as a network of nodes and edges, with nodes being intersections, segments of roads, or individuals. Through this approach, traffic patterns can be understood, congestion can be predicted, and traffic management systems can be optimized.

Peer-to-peer networks are used for peer-to-peer applications (DApps). DApps facilitate the free flow of communications because no one entity controls the users' communications. DApps offer greater flexibility, transparency, distributed capability, and resilience with a more incentivized structure. Ethereum is a fully open-source platform for deploying applications that use cryptographic tokens to store data and record operations.

Due to the use of SSL/TLS for encryption of Ethereum-based DApp data, deep packet inspection, a method of classifying traffic, is no longer available. Every application uses port 443 to exchange data. The port number can also not be used to categorize DApps' encrypted traffic. In most cases, centralized applications are managed by more than one organization. Though they both use SSL/TLS, their implementations differ. Since all DApps are built on Ethereum, they have similar traffic features to centralized applications. Using current methods for encrypting and classifying traffic, DApps' transmission data is unaffected.

Modeling SSL/TLS messages with a Markov model. According to Korczynski and Duda, stochastic fingerprints can be used to analyze SSL/TLS-encrypted application traffic [24], [25]. SSL/TLS sessions use compact notation to represent the types of messages (for example, 22:2 represents Server Hello). A sequence of states represents a client-server message. There is a possibility that a single TCP segment will contain two types of SSL/TLS messages (e.g., 22:11 and 22:14) or several messages (e.g., 22:11 and 22:14). Using the Markov chain, we can calculate the probability of a state transition. Also, the Enter Probability Distribution

(ENPD) of a Markov chain, which shows the probability of its first state, must be calculated. A Markov Chain Exit Probability Distribution (EXPD) represents the probability that a chain will exit at a given state. SSL/TLS streams are classified by multiplying ENPD, transition probability, and EXPD together. Higher values indicate a closer match between the encrypted traffic flow and a model of the application flow; lower values indicate a further departure.

Packet length classification using statistical features. Furthermore, the authors present a method that employs packet length to automatically fingerprint encrypted traffic as well as a method for classifying encrypted traffic using SSL/TLS message types. Every flow is analyzed statistically for 54 features. Three packet directions are used in these statistical analyses: incoming, outgoing, and bidirectional packets. In addition to minimum and maximum, major, minor, absolute deviation, standard deviation, variance, skew, kurtosis, and the number of elements, each direction packet is calculated using 18 statistical values: minimum, maximum, mean, median, absolute deviation, standard deviation, variance, skew, kurtosis, and percentiles (10%-90%). The network flows of encrypted applications are classified using Support Vector Classifiers (SVCs) and Random Forests (RFs).

3.2 Graph Neural Networks (GNN)

The GNN is a powerful machine-learning method that uses graph-structured data. Security analysis and smart contract vulnerability detection are two of their most notable accomplishments. While other deep neural networks analyze Euclidean data (e.g., convolution neural networks analyze images), GNNs can iteratively aggregate each node's information across their connections in a graph in order to gather graph structure information, which makes them a useful and effective tool for exploring graph data. This is accomplished by propagating information along a graph's edges by utilizing message-passing algorithms. These updated node features are applicable to a number of downstream graph analysis tasks, including node classification, link prediction, and graph classification. GNN models iteratively update nodes' representations (also known as node embeddings) as they are learned:

$$M_v^t = \sum_{u \in N_n} M_t(h_u^{(t-1)}, h_v^{-(t-1)}, e_{uv}) \quad (1)$$

$$h_v^t = U_t(h_v^{(t-1)}, m_v^t)$$

A node v is embedded at layer $t \in \{1, \dots, T\}$ by h_v^t , whereas $N(v)$ indicates the set of nodes v neighbours in the graph by $G.M_t(\cdot, \cdot)$ and $U_t(\cdot, \cdot)$ respectively. After one has generated the embeddings for all nodes in a graph, one can calculate the average value of the entire graph (e.g., making predictions about the graph's label).

3.3 MLP-Base Structure

According to the common design, FC layers, activation functions, and dropouts comprise MLP-based structures. Due to excessive layers in neural networks being a cause of overfitting and reducing performance, MLP uses three layers. A feature matrix is generated by concatenating topology and attributes, which is then fed into the MLP. By introducing ReLU activation functions and dropouts between the 1st and 2nd layers, we improve the non-linear transformation capabilities and prevent parameter overfitting. Our second layer optimization uses contrastive loss guided by node similarity to optimize node representations. As a result, nodes with high similarity will be pushed closer together, and nodes with low similarity will be pushed farther apart. To achieve node classification, we use the probabilistic distribution of node labels calculated by the softmax function on the 3rd layer. As a generalized MLP structure, we can formulate it as follows:

$$H^1 == Dropout(ReLU(FC(X))), \quad (2)$$

$$H^2 = ReLU(FC(H^1)), \quad (3)$$

$$\hat{Y} = softmax(FC(H^2)). \quad (4)$$

There are $H^{(\cdot)} = \{h_i^{(\cdot)}\}_{i=1}^N$, N nodes in this network. A contrastive loss is modelled using H^2 , and a classification loss is modelled using $\hat{Y}$.

3.4 Optimization Objective

The original graph features are comprehensively extracted, and the similarities between nodes are preserved in both topology and attribute spaces by modification (1) and introduction. MLP-based structures are able to learn node representations without explicit messaging modules because of Similarity Contrastive Loss (SCL). The latent space would be more similar if the nodes were closer together and less similar if they were far apart. A node similarity matrix S defines a positive and negative sample in SCL:

$$S_{ij} \begin{cases} \neq 0, & \text{node } j \text{ is the positive sample of node } i \\ = 0, & \text{node } j \text{ is the negative sample of node } i \end{cases} \quad (5)$$

Positive nodes are encouraged to be closer to the target, while negative nodes are pushed away from it. Details of SCL for node i are as follows:

$$loss_{SCL}^i = -\log \frac{\sum_{j=1}^N 1_{[j \neq i]} \exp\left(\frac{\text{sim}(h_i^2, h_k^2)}{\tau}\right)}{\sum_{k=1}^N 1_{[k \neq i]} \exp\left(\frac{\text{sim}(h_i^2, h_k^2)}{\tau}\right)}. \quad (6)$$

A cosine similarity is represented by sim . Node classification tasks are also performed using traditional Cross Entropy (CE) losses, aside from contrastive losses:

$$loss_{CE} = -\sum_{i \in Y} \sum_{m=1}^M Y_{im} \log(\hat{Y}_{im}). \quad (7)$$

A set of labels is represented by Y , a number of categories by M , a true label is represented by Y_{im} , and a prediction is represented by $\hat{Y}_{im}$. Losses resulting from classification and contrast are combined to determine the final loss of MLP:

$$Loss_{MLP} = loss_{CE} + \alpha loss_{SCL}. \quad (8)$$

$$loss_{SCL} = \frac{1}{N} \sum_{i=1}^N loss_{SCL}^i. \quad (9)$$

This weighting coefficient balances $loss_{CE}$ and $loss_{SCL}$

3.5 Building Powerful Graph Neural Networks

An overview of the maximum representational capacity of GNN-based models is presented in the first section. As a single GNN is trained, its embedding space maps graph structures to multiple representations in an ideal world. However, this capability must be enabled by solving the graph isomorphism challenge. If a graph is isomorphic, it should be mapped to the same representation, while if it isn't, then it should be mapped to a different representation. In our analysis, there is a slightly weaker criterion for determining GNNs: the Weisfeiler-Lehman graph isomorphism test, which is normally successful except for graphs with regular shapes.

The Appendix contains all Lemmas and Theorems. All Lemmas and Theorems are listed in the Appendix. Does the WL test, in principle, offer as much power as GNNs? This GNN has the same

power as the WL test when the neighbour aggregation and graph level readout functions are injective, as demonstrated in Theorem 3.

A aggregates and updates node features iteratively with

$$H_v^k = \phi \left(h_v^{(k-1)}, f \left(\{h_u^{(k-1)} : u \in N(v)\} \right) \right). \quad (10)$$

Assume that f and ϕ are injective functions that operate on multisets

Based on H_v^k Multiset of node features, A's graph-level readout is injective.

As a result of the universal approximation theorem, multilayer perceptrons can be used in Corollary 6 to model and learn f and ϕ . As a result of the MLP's ability to represent a composition of functions, we usually model $f^{k+1} \circ \phi^k$ With one MLP. As a one-hot encoding is injective by itself, we do not require MLPs in the first iteration. There are two types of parameters: learnable and fixed.

$$h_v^k = MLP^k \left((1 + \epsilon^k) \cdot h_v^{(k-1)} + \sum_{u \in N(v)} h_u^{(k-1)} \right) \quad (11)$$

A wide range of powerful GNNs may exist in general. GNNs are among the most powerful yet simple algorithms.

A GIN embedding can be used directly to predict node links and classify nodes. The following "readout" function is proposed for graph classification tasks that produce the whole graph embedding based on an embedding of each node.

In graph-level readouts, node representations are refined and globalized over iterations, corresponding to subtree structures. When achieving discriminative power, it is important to iterate enough times. It is possible, however, that earlier iterations tend to generalize better than later versions. All depths and iterations of the model are used to consider all structural information. A similar approach to Jumping Knowledge Networks is used here, where graph representations are concatenated across all iterations/layers of the network:

$$h_G = \text{CONCAT}(\text{READOUT}(\{h_v^k | v \in G\}) | k = 0, 1, \dots, K). \quad (12)$$

3.6 Preliminary

The fusion of features illustrates our approach. This is how our model is developed. Our first step is to collect traffic from DApps. A variety of dimensions are then extracted. The fusion features are selected after extracting the various dimensions. The kernel

function fuses features of different dimensions, increasing the feature number simultaneously.

This new WF attack uses cutting-edge deep learning methods to construct a convolutional neural network (CNN). A simple input format is used for classification, and no handcrafted features are needed. This paper describes a robust and effective classification system based on advances in computer vision research.

A deeper network can be created by stacking LSTM-L. Higher LSTM layers are supposed to capture abstract concepts according to the same logic. This was achieved by chaining together two hidden LSTM layers (each layer being unrolled into as many layers as there were time steps in the sequence), resulting in the most efficient network.

4 RESULT ANALYSIS AND DISCUSSION

With the same dataset, Figure 1 shows how much time was spent at each stage and how long was spent training all methods. In FEAF, the process of extracting and selecting features takes a considerable amount of time before Random Forest classifiers are trained. The second most time-consuming application is APPS, which requires significant computation of statistical features despite fast training of classifiers. A proposed model, GraphDApp, constructs temporal interaction graphs more slowly than LSTM+L but learns more efficiently, reaching target accuracy more quickly. As a whole, GraphDApp is the fastest and most efficient training tool.

All methods have an average testing time of under 0.03 seconds. FEAF is the slowest because it fuses multidimensional features. Extracting features during testing remains time-consuming. Deep learning models, on the other hand, are more efficient at labelling unknown flows. In spite of its complexity, the proposed model maintains a competitive prediction speed, despite its slower speed, as in Figure 2.

Figure 3 compares TPRs and FPRs for various methods based on the number of unmonitored DApps. When more DApps are unmonitored, both metrics tend to decrease. By the time the number increases to 720, the proposed model's TPR and FPR are both 0.99 and 0.09. Other methods, however, have

reported declining TPRs for 1,260 unmonitored DApps as a result of misclassification of monitored flows.

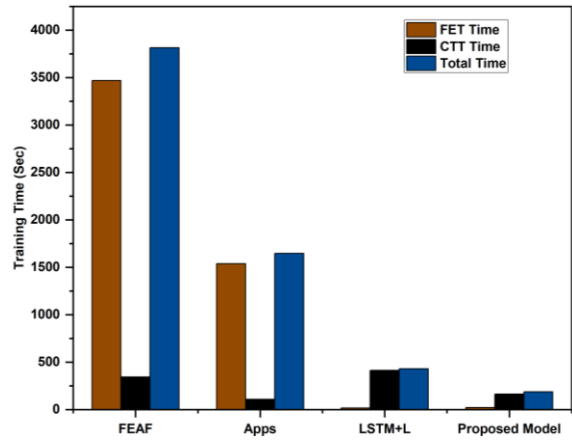


Figure 1: The training time for different methods.

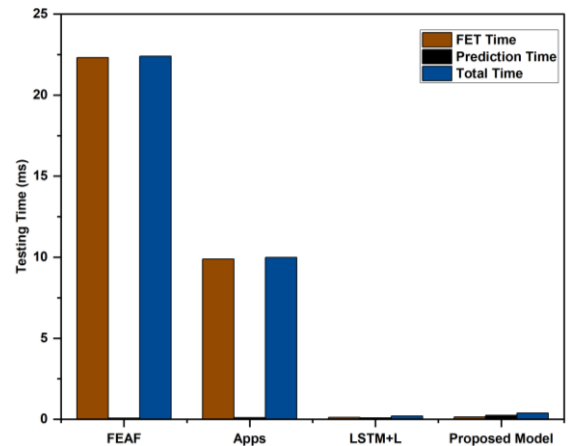


Figure 2: The testing time for different methods.

In Figure 4, different methods are compared based on their classification accuracy. Strong performance was demonstrated by APPS, FEAF, and LSTM+L, demonstrating effective feature extraction. With an accuracy of 1.0, the proposed model is near-perfect. In spite of the simplified graph structure, the temporal interaction graphs (TIGs) still provide sufficient information for effective classification using GNNs. As a result of its superior performance, the model can be used to fingerprint mobile applications.

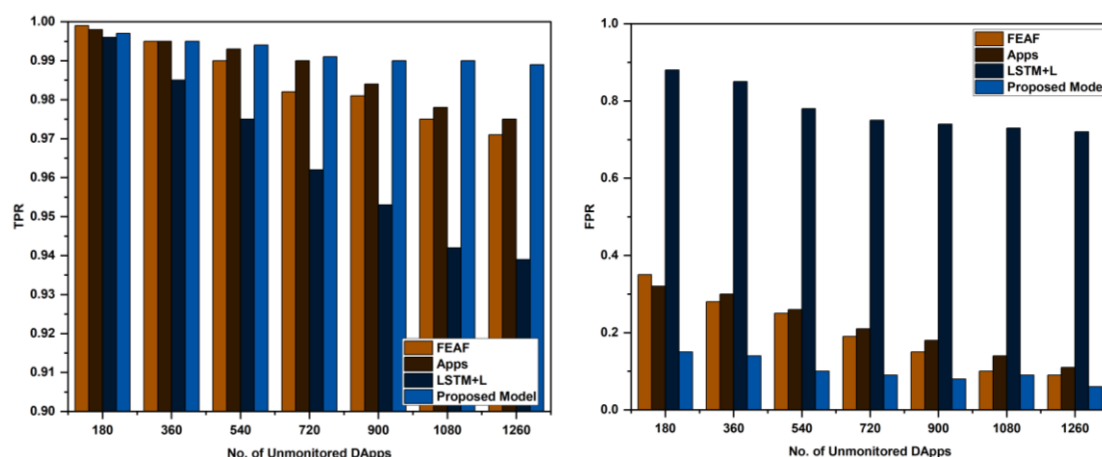


Figure 3: The effect of unmonitored DApps on TPRs and FPRs.

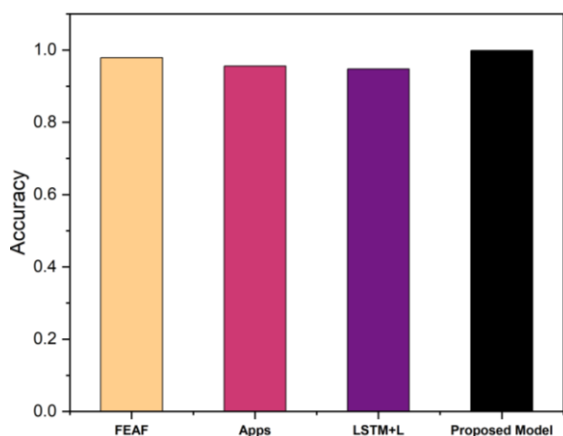


Figure 4: A comparison of classification results on mobile applications using different methods.

5 CONCLUSIONS

The purpose of this study is to introduce GraphDApp, a GNN-based method for classifying encrypted traffic generated by decentralized applications. By transforming flow data into temporal interaction graphs, The proposed method successfully identifies Decentralized Applications (DApps) even in scenarios where conventional features become indistinguishable due to similar encryption schemes, protocols, or traffic patterns, by effectively capturing complex dependencies, structural details, and subtle behavioral nuances within the network data. To validate our approach, we conducted a comprehensive set of experiments covering various scenarios, clearly demonstrating GraphDApp's superiority in terms of classification accuracy, efficient training times, scalability, and robustness, particularly when faced

with numerous unmonitored or previously unknown DApps. These experimental findings firmly establish graph-based learning as a powerful and reliable tool for analyzing encrypted network traffic, significantly outperforming traditional fingerprinting methods. This research thus opens new avenues for further investigations into secure, privacy-preserving application fingerprinting methodologies, enabling more resilient and practical network security strategies tailored for emerging decentralized environments.

REFERENCES

- [1] P. Rani, S. Verma, S. P. Yadav, B. K. Rai, M. S. Naruka, and D. Kumar, "Simulation of the lightweight blockchain technique based on privacy and security for healthcare data for the cloud system," *International Journal of E-Health and Medical Communications (IJEHMC)*, vol. 13, no. 4, pp. 1-15, 2022.
- [2] P. Rani, P. N. Singh, S. Verma, N. Ali, P. K. Shukla, and M. Alhassan, "An implementation of modified blowfish technique with honey bee behavior optimization for load balancing in cloud system environment," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1-14, 2022.
- [3] R. T. El-Maghraby, N. M. Abd Elazim, and A. M. Bahaa-Eldin, "A survey on deep packet inspection," in *Proc. 12th International Conference on Computer Engineering and Systems (ICCES)*, Cairo, Egypt, Dec. 2017, pp. 188-197, doi: 10.1109/ICCES.2017.8275301.
- [4] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," *Computers and Electrical Engineering*, vol. 105, p. 108543, 2023.
- [5] A. Dainotti, A. Pescapé, and K. Claffy, "Issues and future directions in traffic classification," *IEEE Network*, vol. 26, no. 1, pp. 35-40, Jan. 2012, doi: 10.1109/MNET.2012.6135854.

- [6] S. Rezaei and X. Liu, "Deep Learning for Encrypted Traffic Classification: An Overview," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 76-81, May 2019, doi: 10.1109/MCOM.2019.1800819.
- [7] N. Hussain, P. Rani, N. Kumar, and M. G. Chaudhary, "A deep comprehensive research architecture, characteristics, challenges, issues, and benefits of routing protocol for vehicular ad-hoc networks," *International Journal of Distributed Systems and Technologies (IJ DST)*, vol. 13, no. 8, pp. 1-23, 2022.
- [8] X. Li et al., "Prediction of urban human mobility using large-scale taxi traces and its applications," *Frontiers of Computer Science*, vol. 6, no. 1, pp. 111-121, Feb. 2012, doi: 10.1007/s11704-011-1192-6.
- [9] B. Yu, M. Li, J. Zhang, and Z. Zhu, "3D Graph Convolutional Networks with Temporal Graphs: A Spatial Information Free Framework For Traffic Forecasting," *arXiv preprint*, 2019, doi: 10.48550/ARXIV.1903.00919.
- [10] N. Hussain and P. Rani, "Comparative studied based on attack resilient and efficient protocol with intrusion detection system based on deep neural network for vehicular system security," in *Distributed Artificial Intelligence*, Boca Raton, FL, USA: CRC Press, 2020, pp. 217-236.
- [11] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting Taxi-Passenger Demand Using Streaming Data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1393-1402, Sep. 2013, doi: 10.1109/TITS.2013.2262376.
- [12] Y. Tong et al., "The Simpler The Better: A Unified Approach to Predicting Original Taxi Demands based on Large-Scale Online Platforms," in *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, NS, Canada, Aug. 2017, pp. 1653-1662, doi: 10.1145/3097983.3098018.
- [13] A. Panchenko et al., "Website Fingerprinting at Internet Scale," in *Proc. Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, 2016, doi: 10.14722/ndss.2016.23477.
- [14] T. Wang, X. Cai, R. Nithyanand, R. Johnson, and I. Goldberg, "Effective attacks and provable defenses for website fingerprinting," in *Proc. 23rd USENIX Security Symposium*, San Diego, CA, USA, 2014, pp. 143-157, [Online]. Available: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/wang_tao.
- [15] V. F. Taylor, R. Spolaor, M. Conti, and I. Martinovic, "AppScanner: Automatic Fingerprinting of Smartphone Apps from Encrypted Network Traffic," in *Proc. IEEE European Symposium on Security and Privacy (EuroS&P)*, Saarbrücken, Germany, Mar. 2016, pp. 439-454, doi: 10.1109/EuroSP.2016.40.
- [16] M. Shen, J. Zhang, L. Zhu, K. Xu, X. Du, and Y. Liu, "Encrypted traffic classification of decentralized applications on ethereum using feature fusion," in *Proc. ACM International Symposium on Quality of Service*, Phoenix, AZ, USA, Jun. 2019, pp. 1-10, doi: 10.1145/3326285.3329053.
- [17] G. Ansari, P. Rani, and V. Kumar, "A novel technique of mixed gas identification based on the group method of data handling (GMDH) on time-dependent MOX gas sensor data," in *Proc. International Conference on Recent Trends in Computing (ICRTC)*, 2022, pp. 641-654.
- [18] Q. Meng, H. Guo, Y. Liu, H. Chen, and D. Cao, "Trajectory Prediction for Automated Vehicles on Roads With Lanes Partially Covered by Ice or Snow," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 6972-6986, Jun. 2023, doi: 10.1109/TVT.2023.3236947.
- [19] P. Rani, U. C. Garjola, and H. Abbas, "A Predictive IoT and Cloud Framework for Smart Healthcare Monitoring Using Integrated Deep Learning Model," *NJF Intelligent Engineering Journal*, vol. 1, no. 1, pp. 53-65, 2024.
- [20] B. Anderson and D. McGrew, "Identifying Encrypted Malware Traffic with Contextual Flow Data," in *Proc. ACM Workshop on Artificial Intelligence and Security*, Vienna, Austria, Oct. 2016, pp. 35-46, doi: 10.1145/2996758.2996768.
- [21] M. Shen et al., "Machine Learning-Powered Encrypted Network Traffic Analysis: A Comprehensive Survey," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 791-824, 2023, doi: 10.1109/COMST.2022.3208196.
- [22] N. K. Agrawal et al., "TFL-IHOA: Three-Layer Federated Learning-Based Intelligent Hybrid Optimization Algorithm for Internet of Vehicle," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 3, pp. 5818-5828, Aug. 2024, doi: 10.1109/TCE.2023.3344129.
- [23] G. Srivastava et al., "XAI for Cybersecurity: State of the Art, Challenges, Open Issues and Future Directions," *arXiv preprint*, 2022, doi: 10.48550/ARXIV.2206.03585.
- [24] M. Shen, M. Wei, L. Zhu, and M. Wang, "Classification of encrypted traffic with second-order Markov chains and application attribute bigrams," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 8, pp. 1830-1843, 2017.
- [25] M. Korczyński and A. Duda, "Markov chain fingerprinting to classify encrypted traffic," in *Proc. IEEE Conference on Computer Communications (INFOCOM)*, Toronto, ON, Canada, 2014, pp. 781-789, [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6848005/>.

Enhancing Machine Learning Model Accuracy with Effective Data Scaling Strategies

Wasan Khairallah Ali

*Department of Chemistry, College of Science, University of Mosul, 41002 Mosul, Iraq
wasankhairallahali@uomosul.edu.iq*

Keywords: Data Scaling, Machine Learning Algorithms, Prediction of Heart Disease, Model Performance Metrics, Preprocessing Strategies.

Abstract: The study on the impact of data scaling techniques on machine learning algorithms for predicting heart disease highlights the importance of preprocessing in enhancing model performance. Data scaling is essential when dealing with datasets that have diverse attribute ranges, as it can significantly influence the effectiveness of various machine learning models. In this investigation, eleven widely used algorithms, including K-Nearest Neighbors (KNN) and Logistic Regression, were evaluated using three scaling methods: Min-Max scaling, Z-score standardization, and MaxAbs scaling. The performance was assessed through precision, recall, and F1 score metrics across multiple experiments. The findings indicate that several algorithms performed better with MaxAbs scaling, particularly those sensitive to data distribution, such as KNN and Logistic Regression. This suggests that the choice of scaling technique is crucial for achieving accurate and consistent predictions in machine learning applications related to heart disease. The results emphasize the need for careful selection of scaling methods to optimize the performance of machine learning models in medical diagnostics.

1 INTRODUCTION

As a result of their ability to analyze large datasets and make predictions, machine learning models have gained widespread adoption across many industries, including healthcare and finance. Creating accurate and efficient machine learning models, however, requires data handling. A raw, unprocessed dataset's noise and inconsistency can have an adverse effect on algorithm performance. Preprocessing techniques like data scaling are essential for improving model performance. Algorithms are prevented from being biased against variables with larger scales by scaling data, which includes normalization and standardization. When not scaled, distance-based machine learning algorithms, like KNN, or gradient-based algorithms, like gradient descent, can produce inaccurate results, take longer to converge, or require longer training times [1]. Effective data scaling can improve model accuracy and performance, particularly when dealing with high-dimensional datasets. Scaling is an essential step for many machine learning algorithms, but the optimum scaling strategy will depend on the data's characteristics and the algorithm itself. Several data scaling techniques are examined, including Min-Max scaling,

standardizing Z-scores, and robust scaling on machine learning models. Models can achieve greater accuracy and reliability if practitioners understand and apply the appropriate data scaling methods.

Electrocardiograms and blood tests are often required to evaluate heart disease symptoms properly. There are almost 12 million deaths caused by heart disease every year [2]. As a result, it is crucial to diagnose this disease at the earliest opportunity. Artificial intelligence (AI) has provided quick and alternative methods for diagnosing diseases, which may be beneficial in rural areas where doctors are few and diagnostic equipment is costly. An automated system that can be operated by non-medical personnel would therefore be helpful. Research shows that early diagnosis of heart disease can save patients time, money, and healthcare costs by using additional patient information and medical histories [3]. Machine learning approaches help develop decision support systems.

Learning online [4], scheduling [5], multiobjective optimization [6], and vehicle routing [7] are examples of AI-based algorithms (e.g., heuristics, metaheuristics). Medical diagnosis can be greatly improved using deep-learning-based approaches, according to several recent

studies [4],[5]. Deep learning technologies, such as image segmentation, now improve the efficiency and effectiveness of diagnosing diseases like diabetes, cancer, and SARS-CoV-2. Although SARS-CoV-2 caused a global pandemic, chest X-rays (X-rays) and computed tomography (CT) scans were used in several studies to identify COVID-19 symptoms [1]. A laptop of office grade and a small amount of data were used for this experiment.

By combining data classification techniques with nature-inspired algorithms, genetic programming [10] and swarm algorithms can be used to identify bacteria from viral meningitis [11]. In recent years, artificial intelligence has gained popularity among decision-support and optimization tools due to these advantages. The computational cost of deep learning and neural network approaches increases with larger datasets. Since traditional machine learning approaches consume less memory and computational power than deep learning approaches, they are frequently preferred unless necessary [6].

The development of a decision support system based on data analysis requires standard data, which is often the result of extensive preprocessing. Preprocessing involves the cleaning of data, pruning, slicing, and scaling of features before data can be analyzed. The effects of data scaling on overall model performance have been examined in many studies. Still, few have investigated the effects of different machine-learning algorithms and feature selection on overall model performance [7]. It examines ML algorithms for predicting symptoms of heart disease in patients using various data scaling methods. A robust, data-driven decision support system can be developed based on researchers' and practitioners' experimental results.

2 LITERATURE REVIEW

According to the referenced literature, machine learning algorithms (ML) are expressed in terms of accuracy. ML algorithms, however, perform differently in each study because of different approaches to ML. According to the Author [14], the Bagging algorithm achieved 81.14% accuracy, and the Decision Tree (DT) algorithm achieved 78.90% accuracy. According to the Author [15], 84.14% of patients with heart disease were correctly identified using a naïve approach. The Author [8] also reported an accuracy rate of 84.10% using a decision tree [23].

The computational accuracy of several references shows promising results, yet the UCI heart disease dataset shows a variation of almost 7–8% despite

using the same dataset [9]. Due to the lack of mention of data scaling methods in any of the studies, it is impossible to identify the reasons for variations in DT accuracy between them. Perhaps the training/test sets are divided differently or the number of features is different. Moreover, accuracy is not always indicative of overall performance. Therefore, categorization matrices based on F1 scores and accuracy, precision, recall, and recall coefficients are more reliable [10].

As the preprocessing of data does not seem to affect prediction models, most studies examine feature selection rather than data scaling due to the lack of information about its effect on prediction models. However, data analysis cannot ignore the importance of feature selection. Using ML algorithms and features as examples, the Author [7] demonstrated different levels of accuracy based on different combinations of algorithms. Furthermore, the study showed that accuracy often drops by 14% due to the limited number of features in medical diagnosis, which is significant [11].

Various machine-learning approaches rely heavily on normalization, according to the Author [12]. A variety of machine learning algorithms, including 12 different ones, were used in the study in order to predict heart disease. Based on a study that used different normalization methods, there is a correlation between the performance of ML algorithms and the selection of normalization methods. Its accuracy is 78%, which is the highest of all eleven supervised algorithms. However, NaVe Bayes also has a high accuracy and low fitting time, according to this study.

In addition, Author [13] notes that data normalization and standardization techniques like MinMax normalization also play an important role in data analysis. Combining ML algorithms like KNN, Nave Bayesian, ANN, and SVM with RBF was used in this study. The performance of NB was more stable without data scaling techniques than that of SVMs, while that of KNN was more stable without data scaling techniques. According to their computation, MinMax scaling outperformed other algorithms with SVM, which contradicts the findings of the study. Although their studies do not synchronize, one could still conclude that data scaling affects overall performance even though they are not synchronized.

According to the Author [14], feature scaling and normalization techniques can improve classification accuracy and model convergence rates. Scaling methods include min-max normalization, Z-score standardization, and decimal scaling; the choice depends on the distribution of data and algorithm [15].

According to the Author, preprocessing is an essential step in knowledge discovery [16]. Unscaled data is a significant problem for several algorithms, including support vector machines, K-nearest neighbours, and neural networks. The researchers noted that standardization often improves the performance of models based on Gaussian-distributed features, whereas robust scaling is more suitable for datasets with outliers. The Author [17] conducted comparative experiments using various scaling techniques across different datasets and machine learning algorithms. Using imbalanced datasets as a case study, another study examined the effects of data preprocessing strategies. The authors primarily focused on class imbalance, but they also demonstrated how improper scaling can exacerbate minority class detection issues, particularly in distance-based models [18].

3 METHODOLOGY

A total of five benchmark datasets were used to examine the impact of min-max data normalization on the regression performance of three machine learning algorithms. Based on the methodology, the following conclusions can be drawn:

3.1 The UCI Datasets

UCI's machine learning repository includes five benchmark datasets [19]. All attributes have a wide range of records, which makes these datasets appealing. The Airfoil SelfNoise Dataset has big differences in its ranges, while power plant datasets have very similar ranges. Considering this variation, the min-max scaling method should have a greater impact on regression performance, which is the primary objective. This dataset contains data from physics, biology, engineering, and business applications.

3.2 Data Scaling in Classification Modelling

When it comes to machine learning and data mining, scaling and normalization refer to the same preprocessing procedure. Data consolidation or transfer is necessary to make mining and modelling possible. [20]. A scaled dataset has been shown to perform better than an unscaled dataset in models trained with scaled data. A distance-based method such as nearest neighbour classification and clustering relies heavily on scaling data. A neural

network's stability and speed of learning are improved when the input data is normalized [20].

Confusions of gene expression normalization. Microarray technology makes it possible to diagnose a wide range of diseases and cancers based on gene expression data. It is usually necessary to perform a normalization step in order to identify and remove systematic variations in fluorescence measurements [21] prior to analyzing them. In our study, the normalization of gene expression is not the same as data scaling. In order to learn a classification model, normalized gene expression datasets are usually scaled and processed. In most cases, gene expression data-driven models outperform gene expression data-driven models without scaling by a wide margin.

3.3 Commonly Used Data Scaling Algorithms

3.3.1 Min-Max Algorithm

Algorithms such as Min-max use linear transformations. Our samples contain the minimum and maximum values of variables. x_{min} and x_{max} . Using the formula below, Min-Max converts a value, v , into a value, v' ;

It is commonly used to scale data using two algorithms

Algorithms based on min-max and z-score.

$$v' = \frac{v - x_{min}}{x_{max} - x_{min}} + x_{min}. \quad (1)$$

Training samples are scaled from $[x_{min}, x_{max}]$ to $[-1, 1]$ (or $[0, 1]$) by a linear plotting. In some applications, it may be problematic to have scaled values outside the interval $[-1, 1]$ (or $[0, 1]$) if the unseen/testing samples are outside the training data range. Further, it is highly sensitive to outliers, as explained in the subsequent sections.

3.3.2 Z-Score Algorithm

The Z-score algorithm divides a variable's original value by its new value, v' .

$$v' = \frac{v - \bar{x}}{\sigma_x}. \quad (2)$$

There are two variables in the training samples: $\bar{x}$ and σ_x , which represent their mean and standard deviation. With a mean of 0 and a standard deviation of 1, you will get new values with 0 and 1 as the mean and standard deviation, respectively. In addition to not using interval mapping, the algorithm is also

sensitive to outliers. The mean and standard deviation calculated from the data may not accurately reflect the true mean and standard deviation in a few cases, especially in biomedical research scenarios.

3.3.3 Data Saling Formula

Variables in the examples are modelled as random variables (*r. v.*) X :

$$v' = P_x(v) \quad , \quad (3)$$

X r.v. is represented by a cumulative density function (CDF) called $P_x(\cdot)$

The Histogram Equalization [21], which is used in Digital Image Processing to enhance image contrast, also uses a CDF as a mapping tool. We use the GL algorithm instead of Histogram Equalization by learning/approximating the CDF's functional expression, which can be used to scale unknown values according to the CDF.

3.4 Overview of Algorithms

3.4.1 Logistic Regression (LR)

Additionally, Logistic Regression can be used to classify data. LR is an effective machine-learning method for classifying objects. According to Verhulst's paper in proceedings of the Belgian Royal Academy, the logistic function has three parameters and a curve. In spite of the simplicity of this machine learning method, it has numerous applications. A logistic regression technique predicts binary classes based on statistical data. A Bernoulli distribution is applied to the dependent variable. Sigmoid functions, or logistic functions, are S-shaped curves that take values between 0 and 1. Observe that the curve will predict 1 if it reaches positive infinity and 0 if it reaches negative infinity [22].

3.4.2 Linear Discriminant Analysis (LDA)

Linear discriminant analysis was used in this study to extract features, while principal component analysis was also evaluated. Several techniques have been developed to reduce dimensionality in preprocessing for machine learning classification applications, including linear discriminant analysis (LDA) [23]. LDA optimizes class separation by minimizing variance between and within classes by transforming features into a lower-dimensional space. It is possible to project an N -dimensional space onto a subspace K ($K \leq n - 1$) by using LDA without compromising class discrimination.

3.4.3 K-Nearest Neighbors (KNN)

There was an initial report on the use of k-Nearest Neighbors (KNN) to categorize text. As part of this method, a query's category is determined not only from its closest document in the document space but also from the k closest documents' categories. In this context, the Vector method is equivalent to the KNN method, where $k = 1$. Using a vector-based, distance-weighted matching function, this work calculates a document's similarity using the Vector method, as did Yang.

3.4.4 Classification and Regression Trees (CART)

Decision trees can be built using CART both for classifications and regressions. Using classification trees, we can separate datasets into two classes, while using regression trees, we can predict numeric or continuous values. The goal of a classification tree is to divide the dataset at hand into two parts based on data homogeneity. When splitting attributes in CART, impurity measures such as entropy and Gini index are used to determine where they should be split. Since regression trees have no classes for output attributes, predicting the value of output variables is more important than predicting their classes. The split point that gives the lowest sum of squared errors between the actual and predicted values is chosen as the root node, and then it is divided into two more nodes.

3.4.5 Naive Bayes (NB)

Based on the assumption that predictors are independent, this classification technique utilizes the Bayes Theorem. Using the Naïve Bayes classifier, one feature is assumed to be independent of another. Naïve Bayes mainly targets the text classification industry. The main purpose of this method is to cluster and classify in accordance with conditional probabilities.

3.4.6 Support Vector Machine (SVM)

Support vector machines (SVMs) are among the most common techniques for machine learning. It is an algorithm that uses support vector machines to classify and regress data. SVMs can also be classified nonlinearly using kernel tricks, which implicitly map inputs into high-dimensional feature spaces. Class margins are drawn in this way. Classification error is minimized by drawing margins that are as far away from the classes as possible.

3.4.7 Extreme Gradient Boosting (XGB)

A new gradient-boosting method called XGBoost has been introduced. As a result of XGBoost, gradient-boosting decision trees can be implemented. Enhancing existing models by adding new ones is an ensemble-based method of correcting errors. In boosting techniques, errors or misclassified observations are given weights to make the selection of subsamples more intelligent. Gradient boosting creates new models by predicting the errors or residuals of earlier models and then combining them to arrive at a final prediction. [12] Gradually boosting is used to reduce loss as new models are added.

3.4.8 Decision Tree (DT)

Graphs are used to represent choices and their results. There are nodes in the graph that represent events and edges that represent the rules or conditions that govern decisions. A node connects an individual branch or node of a tree. An attribute can be represented as a node, and a value can be represented as a branch in a classified group.

3.4.9 Random Forest (RF)

Classification and regression are both carried out using tree-based random forests. As a result of constructing multiple trees, the mean prediction would be the classification output.

3.4.10 Gradient Tree Boosting (GB)

Euclidean optimization methods cannot be used to optimize the tree ensemble model since functions are included as parameters. Models are instead trained in an additive manner.

3.4.11 Adaboost (AB)

Weights are maintained by the AdaBoost algorithm, and they are adjusted after each weak learning cycle. A weak learner's misclassifications will be increased in weight, and correctly classified samples will have their weight decreased.

3.4.12 Extra Trees (ET)

A classifier that uses ensembles of trees to classify data is the extra Trees Classifier. Multiple decision trees are combined into a single prediction by the random forest algorithm. The Extra Trees Classifier

differs from traditional Random Forests in that instead of selecting the best split for each feature, and it uses random thresholds. As a result of this randomization process, we are left with a greater variety of trees so that we can make stronger final predictions.

3.5 Evaluating Models

Model performance was evaluated using established classification metrics derived from the confusion matrix framework presented in [24]. Four fundamental classification outcomes were computed: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Based on these values, accuracy, precision, recall, and F1-score were calculated according to the formulas presented in the aforementioned reference.

Accuracy quantifies the overall proportion of correctly classified samples relative to all predictions. Precision measures the relationship between correctly predicted positive instances and the total number of predicted positives. Recall, also denoted as the true positive rate (TPR), represents the proportion of actual positive cases that were correctly identified by the model. The F1-score combines precision and recall into a single performance metric, providing a balanced evaluation of model effectiveness. These metrics collectively provide a comprehensive assessment of the model's classification performance as outlined in [24].

4 RESULTS AND DISCUSSION

On the basis of eleven different data scaling techniques, Figure 1 illustrates the overall accuracy of eleven machine learning algorithms. The highest accuracy was obtained by SVM and CART (99%), even without scaling. Compared to KNN, it had the poorest accuracy, with just 75%. A significant improvement in overall accuracy was observed when MaxAbs scalers were applied. Most algorithms also exhibited largely consistent accuracy, regardless of scaling methods used, with the exception of Logistic Regression (LR), KNN, and SVM, which exhibited more pronounced differences.

Based on various data scaling techniques, Figure 2 shows the overall precision scores of eleven machine learning algorithms. CART's precision (100%) without scaling was the highest, while KNN's was the lowest (78%). A number of algorithms suffer from reduced performance when normalized, including the LR algorithm, the LDA algorithm, the

CART algorithm, the SVM algorithm, and the AB algorithm. The figure clearly illustrates that CART consistently achieves the highest precision, no matter how it is scaled. Conversely, KNN, SVM, and LDA were found to be the least precise. In addition to CART, Random Forest (RF) and Extra Trees (ET) demonstrated greater precision across different scaling methods.

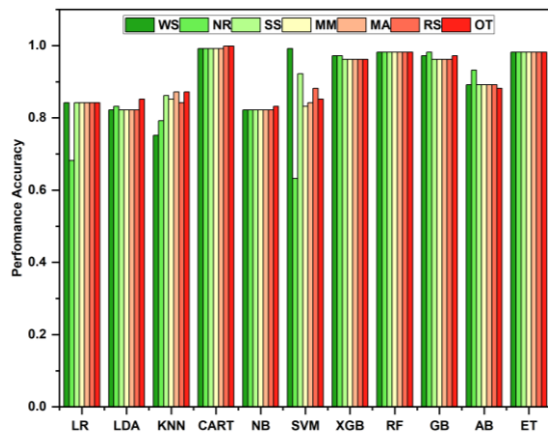


Figure 1: Various algorithms' accuracy performance.

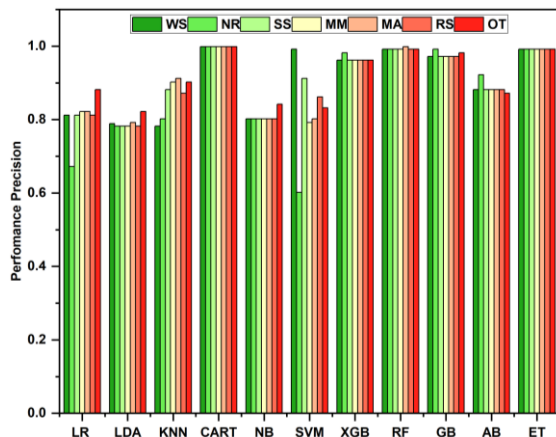


Figure 2: A comparison of algorithms based on their precision.

According to Figure 3, eleven machine learning algorithms were evaluated with various data scaling methods to determine their overall recall performance. With a recall close to one, CART achieved the highest performance. On the other hand, KNN (ranging from 0.72 to 0.83) and Logistic Regression (ranging from 0.74 to 0.89) recorded the lowest recall values.

A comparison of eleven machine learning algorithms is shown in Figure 4 in terms of the overall F1 score. Among the F1 teams, CART achieved the

highest score of 100%. The lowest F1 scores were achieved by KNN, SVM, Logistic Regression (LR), and Naïve Bayes (NB) among the algorithms evaluated.

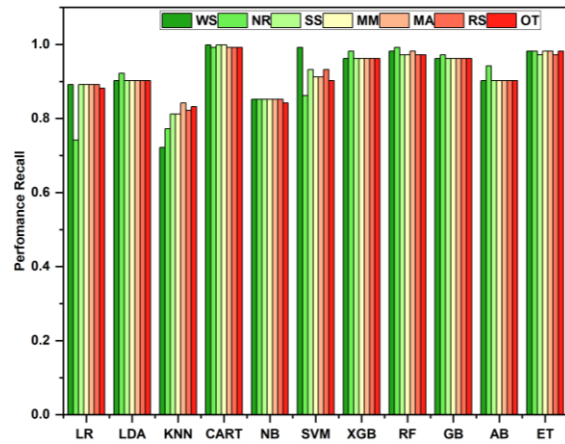


Figure 3: Algorithm performance based on recall.

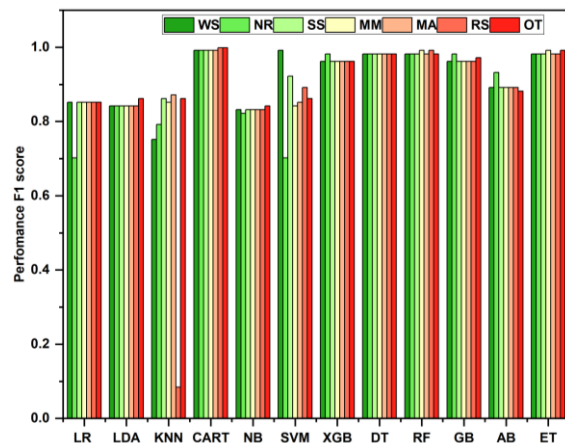


Figure 4: An analysis of the performance of different algorithms using F1 scores.

5 CONCLUSIONS

Machine learning models perform significantly better when data is scaled, particularly in medical diagnosis scenarios such as predicting heart disease, where precision is critical. Even without applying scaling techniques, the Classification and Regression Tree (CART) algorithm consistently outperformed eleven other machine learning algorithms tested across all evaluation metrics, demonstrating its robustness to data variations. Conversely, algorithms like K-Nearest Neighbors (KNN) and Logistic Regression exhibited substantial performance gains when scaling

methods such as MaxAbs scaling were employed, indicating their sensitivity to feature scales. Interestingly, certain algorithms showed minimal or negligible changes when scaling was applied, whereas normalization methods even resulted in degraded performance for specific models. This underscores the critical insight that no single scaling technique universally enhances model performance; rather, its effectiveness is inherently tied to the characteristics of the dataset and the specific algorithm employed. In healthcare applications, selecting an appropriate scaling strategy can significantly enhance model reliability, accuracy, and efficiency, thereby improving the quality and trustworthiness of data-driven clinical decision support systems, and ultimately, patient outcomes.

REFERENCES

- [1] P. Rani, S. Verma, S. P. Yadav, B. K. Rai, M. S. Naruka, and D. Kumar, "Simulation of the lightweight blockchain technique based on privacy and security for healthcare data for the cloud system," *International Journal of E-Health and Medical Communications*, vol. 13, no. 4, pp. 1-15, 2022, [Online]. Available: <https://doi.org/10.4018/IJEHMC.20221001.0a8>.
- [2] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43-48, 2011.
- [3] W. P. Lord and D. C. Wiggins, "Medical Decision Support Systems," in *Advances in Health Care Technology*, G. Spekowius and T. Wendler, Eds., vol. 6. Dordrecht: Kluwer Academic Publishers, 2006, pp. 403-419, doi: 10.1007/1-4020-4384-8_25.
- [4] M. M. Ahsan, T. E. Alam, T. Trafalis, and P. Huebner, "Deep MLP-CNN Model Using Mixed-Data to Distinguish between COVID-19 and Non-COVID-19 Patients," *Symmetry*, vol. 12, no. 9, p. 1526, Sep. 2020, [Online]. Available: <https://doi.org/10.3390/sym12091526>.
- [5] M. M. Ahsan et al., "Detecting SARS-CoV-2 From Chest X-Ray Using Artificial Intelligence," *IEEE Access*, vol. 9, pp. 35501-35513, 2021, [Online]. Available: <https://doi.org/10.1109/ACCESS.2021.3061621>.
- [6] P. Rani, P. N. Singh, S. Verma, N. Ali, P. K. Shukla, and M. Alhassan, "An implementation of modified blowfish technique with honey bee behavior optimization for load balancing in cloud system environment," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1-14, 2022.
- [7] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics and Informatics*, vol. 36, pp. 82-93, Mar. 2019, [Online]. Available: <https://doi.org/10.1016/j.tele.2018.11.007>.
- [8] M. Shouman, T. Turner, and R. Stocker, "Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients," in *Proceedings of the International Conference on Data Science*, 2012, p. 1, [Online]. Available: <http://world-comp.org/p2012/DMI9007.pdf>.
- [9] G. Ansari, P. Rani, and V. Kumar, "A novel technique of mixed gas identification based on the group method of data handling (GMDH) on time-dependent MOX gas sensor data," in *Proceedings of International Conference on Recent Trends in Computing*, Springer, 2023, pp. 641-654.
- [10] A. Singh et al., "Blockchain-Based Lightweight Authentication Protocol for Next-Generation Trustworthy Internet of Vehicles Communication," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 2, pp. 4898-4907, May 2024, [Online]. Available: <https://doi.org/10.1109/TCE.2024.3351221>.
- [11] S. Verma et al., "An automated face mask detection system using transfer learning based neural network to preventing viral infection," *Expert Systems*, p. e13507, 2024.
- [12] L. Shahriyari, "Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma," *Briefings in Bioinformatics*, vol. 20, no. 3, pp. 985-994, May 2019, [Online]. Available: <https://doi.org/10.1093/bib/bbx153>.
- [13] A. Ambarwari, Q. Jafar Adrian, and Y. Herdiyeni, "Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification," *Journal of RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 4, no. 1, pp. 117-122, Feb. 2020, [Online]. Available: <https://doi.org/10.29207/resti.v4i1.1517>.
- [14] S. G. K. Patro and K. K. Sahu, "Normalization: A Preprocessing Stage," *International Advanced Research Journal in Science, Engineering and Technology*, pp. 20-22, Mar. 2015, doi: 10.17148/IARJSET.2015.2305.
- [15] P. Rani, U. C. Garjola, and H. Abbas, "A Predictive IoT and Cloud Framework for Smart Healthcare Monitoring Using Integrated Deep Learning Model," *NJF Intelligent Engineering Journal*, vol. 1, no. 1, pp. 53-65, 2024.
- [16] J. Han, M. Kamber, and J. Pei, "Data mining: Concepts and Techniques," Morgan Kaufmann Publishers, 2012, [Online]. Available: <http://homes.di.unimi.it/ceselli/IM/2012-13/slides/02-KnowYourData.pdf>.
- [17] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249-259, Oct. 2018, [Online]. Available: <https://doi.org/10.1016/j.neunet.2018.07.011>.
- [18] S. P. Yadav et al., "An improved deep learning-based optimal object detection system from images," *Multimedia Tools and Applications*, vol. 83, no. 10, pp. 30045-30072, 2024.
- [19] D. Dua and C. Graff, "UCI machine learning repository," 2017, [Online]. Available: <https://archive.ics.uci.edu/ml>.

- [20] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2022, [Online]. Available: <https://books.google.com/books?hl=en&lr=&id=NR1oEAAAQBAJ>.
- [21] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, pp. 111-139, 2002.
- [22] A. Tharwat, T. Gaber, A. Ibrahim, and A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," *AI Communications*, vol. 30, no. 2, pp. 169-190, 2017.
- [23] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3-42, Apr. 2006, [Online]. Available: <https://doi.org/10.1007/s10994-006-6226-1>.
- [24] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, pp. 37-63, 2011, doi: 10.9735/2229-3981.

Revolutionizing Surveillance with Deep Learning-Powered Anomaly Detection and Real-Time Behavior Analysis

Shanmuga Shyam B.¹, Selva Jothi Murugan¹, Chairan Zibar L. Parisu², Sasmin Sasmin²,
Janani Vignesh³ and Ahmed Mohammed Abdulkareem⁴

¹Department of Computer Science and Engineering, Chennai Institute of Technology, 600069 Chennai, Tamil Nadu, India

² Universitas Sulawesi Tenggara, 93121 Kendari, Indonesia

³Center for Advanced Multidisciplinary Research and Innovation, Chennai Institute of Technology, 600069 Chennai, Tamil Nadu, India

⁴Dijlah University College, 10021 Baghdad, Iraq

shanmugashyamb.cse2023@citchennai.net, selvajothii@gmail.com, chairanzibar@un-sultra.ac.id, sasmin@un-sultra.ac.id, jananiv.chem@citchennai.net, iahmed215@gmail.com

Keywords: Anomaly Detection, Real-Time Surveillance, CCTV Monitoring, Deep Learning, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN).

Abstract: A novel approach for real-time anomaly detection within CCTV surveillance systems, taking the full power of advanced deep learning models to overcome the shortcomings of human-monitored surveillance. Traditional systems are largely dependent on human operators monitoring video feeds, often resulting in fatigue, distraction, and delayed responses. The use of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) in enhancing the automation of anomaly detection. CNNs have been specialized in extracting spatial features from video frames, identifying objects, scenes, and contextual patterns. RNNs, analyzing temporal dynamics, capture behaviors over time, hence detecting patterns that are indicative of abnormal activities. The fusion of these models allows the system to detect fights, accidents, or falls in real time. Moreover, the system provides a much lower rate of false positives, allowing the system to be much more accurate than traditional systems regarding anomaly detection. Real-time anomaly detection is crucial for applications in public safety, health care, traffic management, and retail, all allowing for faster response times and improving operational efficiency with a reduced reliance on human intervention. The discussion is about development, methodology, and potential applications and demonstrates the dramatic impact it had on public safety and operational efficiency.

1 INTRODUCTION

Surveillance in the interest of public safety is a matter that cannot be over emphasised, especially in places of high population density such as urban areas, hospitals, and transportation hubs. Traditional surveillance systems based on human operators are not very feasible. Human operators are prone to fatigue, distractions, and an inability to process large amounts of video data in real time [1]. Besides, human judgement sometimes makes mistakes in recognising unusual behaviours, especially when the environment is complicated or the volume of data is too large. As a response to these issues, there has been increased interest in the integration of more advanced technologies such as AI and deep learning to

automate the processes of anomaly detection. The capability to solve much of the problems of ordinary surveillance systems can be attained through the use of Convolutional Neural Networks and Recurrent Neural Networks of deep learning. This works well for processing visual information because they can automatically make features out of video frames of the frames, which can further allow the systems to recognize objects, scenes, and anomalous patterns [2]. RNNs are highly suitable for understanding the time evolution of video streams and excel in sequential data processing. Anomaly detection can therefore combine the best from both networks, offering a fully spatially and temporally aware system capable of capturing anomalies that play out over time, such as a fight suddenly erupting in a public area or a person collapsing in a health setting. Capacity for

real-time anomaly detection significantly impacts public safety and operational efficiency [3]. Since automated surveillance systems can identify potential threats or anomalies in real time, human intervention is lessened while responses become faster and much more accurate. Such systems can also scale to different ranges of environments; hence they are adaptable and can be applied to use in the various sectors that include monitoring of the city, healthcare, traffic monitoring, and retail. The paper tries to establish a role that is given by automated surveillance on public safety issues, basing this discussion on the progresses that CNNs and RNNs technologies present.

2 DESIGNING STRONG ANOMALY DETECTION SYSTEMS

The designing of an anomaly detection system involves the integration of different technologies and methodologies to ensure both accuracy and efficiency. In the proposed system, the combination of CNNs and RNNs forms the foundation of a robust solution that is capable of performing spatial and temporal analysis of video data [4]. CNNs are very good at feature extraction from single frames, which is an important criterion for detecting objects, patterns, and activities within a frame. The network progressively captures more complex features that are essential in anomaly detection, such as human actions, movement patterns, and interactions between objects. CNNs also effectively reduce the amount of redundant data, hence leading to faster processing times [5]. RNNs are well designed for sequential data. That is why they are effective when analysing the temporal component of video data. Feeds of video surveillance consist of continuous frames wherein information from a scene given at one instant could be read at any instant. RNNs analyse sequences of frames and hence can pick up on abnormal patterns or behaviours evolved over time. For example, the RNN detects a person lingering in restricted areas or an unusual variation in driving pattern in that vehicle. By combining CNNs with RNNs, the system benefits from both spatial and temporal analysis, ensuring that both immediate and developing anomalies are detected [6]. Video data pre-processing is crucial to

increasing the model's accuracy. Optical flow, which is a method used in tracking objects and movements within the video, eliminates the noise background features and concentrates on key elements. It allows the system to concentrate on actions or movements or changes that might show anomalies, thus making it enhance its capability of detecting real threats instead of raising false positives. The system uses various optimisation techniques to ensure real-time performance. These include model pruning, which removes unnecessary weights and layers in the neural networks to simplify the models without losing their accuracy [7]. Then, quantisation, where floating-point weights are converted into lower precision formats, reduces computational costs. Finally, lightweight architectures are also explored with smaller and more efficient neural networks that have high performance but operate on reduced computational resources. Additionally, the proposed system takes advantage of edge computing, bringing computation closer to the source of data, thereby minimizing latency and accelerating response time. Combining these methodologies, the proposed system for anomaly detection is designed to process video data in real time across different types of surveillance as shown in Figure 1.

The design of deep learning models for anomaly detection involves several essential components, beginning with the foundational concept of “Deep Learning Models for Anomaly Detection” and branching into diverse architectural paradigms such as Convolutional Neural Networks (CNNs), Faster Region-Based Convolutional Neural Networks (Faster R-CNNs), Recurrent Neural Networks (RNNs), and hybrid approaches – each representing a unique strategy within this domain.

Table 1 summarizes the key components and methodologies used in designing robust anomaly detection systems for surveillance.

It has significantly changed many sectors with the spread of real-time anomaly detection systems. One of the most significant factors is knowing that anomalies quickly and correctly can improve public safety, enhance operational efficiency, and optimally allocate resources. Indeed, one of the leading applications of such systems has been in public areas, such as monitoring crowd behaviour at protest rallies, concerts, and sporting events. The system can detect signs of distress, aggression, or panic by analysing video feeds, allowing

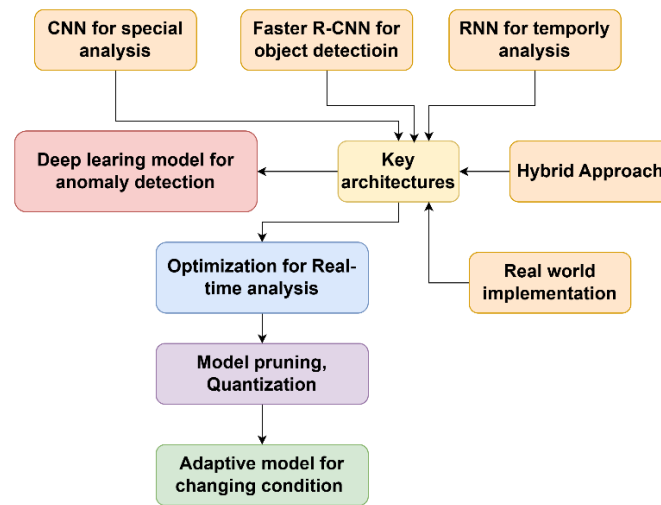


Figure 1: Flowchart of deep learning models for anomaly detection.

3 APPLICATIONS: TRANSFORMING PUBLIC SAFETY AND BEYOND

The system is designed to provide early warnings, enabling authorities to take action before situations escalate into full-blown emergencies [13]. For example, during a crowded concert, it can detect the early signs of a fight and instantly alert security personnel, allowing them to respond quickly and reduce the risk of harm. In a similar way, continuous monitoring of traffic patterns in urban environments can identify anomalies such as potential collisions, vehicles driving in the wrong direction, or sudden stops, enabling prompt interventions to prevent accidents and enhance public safety. Anomaly detection-based traffic management systems will offer instantaneous alerts to the authorities to respond in time if an accident takes place and hence optimise the flow of traffic. The tools may be integrated into traffic control systems in order to prevent congestion and minimise accidents while ensuring safe use of roads [14]. In other similarities, the systems have been used in transportation networks for monitoring the movement of vehicles and pedestrians with the intention of accident prevention, threat detection, and optimisation of flow. In the healthcare system, anomaly detection systems are a guarantee of safety to patients. A system installed in a hospital could monitor the patient movements in such facilities and send alerts to care providers in cases of falls, wandering, or entry into the restricted zones. Such alerting is timely and assists healthcare providers in taking interventions in time for better patient care and

outcomes. Nursing homes could utilise real-time monitoring for at-risk identification, including patients with dementia, for safety without constant human monitoring. This too benefits the retail environments at their peak. It can enhance customer service besides improving security by tracking what has been happening with customers so that shoplifting or unusual activities are caught. These systems also optimise the activities of stores by monitoring people flow and real-time changing of inventory or staff in a store [15]. In all these industries, the real-time features of anomaly detection systems make a big difference in terms of safety, efficiency, and resource allocation. Since anomalies are detected and responded to quickly, the threats are contained before they get out of hand, and these systems are necessary for modern surveillance.

Figure 2 presents a systematic framework for anomaly detection in surveillance systems. It begins by categorizing anomalies into distinct groups, including suspicious individuals, crowd behavior anomalies, vehicle-related irregularities, and environmental deviations. Each category serves as a foundation for identifying specific patterns requiring detection. The process progresses to applying deep learning methodologies, which enhance the anomaly detection system's accuracy and efficiency. Advanced preprocessing techniques are integrated, leveraging large-scale and diverse datasets to ensure robust feature extraction and model training. This structured approach emphasizes modularity, enabling scalability across different surveillance scenarios. By addressing unique anomaly types, the system adapts dynamically to evolving real-world challenges.

Table 2 outlines the key application areas for real-time anomaly detection systems.

Table 1: Key components and methodologies in anomaly detection systems.

Component	Details	Benefits	References
Convolutional Neural Networks (CNNs)	Extract spatial features from video frames, identify objects, scenes, and patterns	Improved accuracy in object and scene detection, reduced redundant data	[8]
Recurrent Neural Networks (RNNs)	Analyze temporal dynamics and sequential data	Detect evolving patterns over time (e.g., lingering or unusual behaviour)	[9]
Video Preprocessing Techniques	Use optical flow and other methods to track movements and filter noise	Enhanced focus on key actions, reduced false positives	[10]
Model Optimization	Techniques like pruning, quantization, and lightweight architectures	Reduced computational costs, real-time performance on edge devices	[11]
Edge Computing	Perform computation closer to data sources	Lower latency, faster response times	[12]

Table 2: Applications of real-time anomaly detection in diverse sectors.

Sector	Application	Details	Benefits	References
Public Safety	Crowd behaviour monitoring, fight detection	Identifying distress, panic, or aggression	Prevent escalation, improve response time	[16]
Traffic Management	Accident detection, wrong-way driving, congestion	Analysing traffic patterns for anomalies	Reduce accidents, optimize traffic flow	[17]
Health care	Monitoring patient falls, restricted area entry	Tracking patient movements	Ensure safety, enable timely interventions	[18]

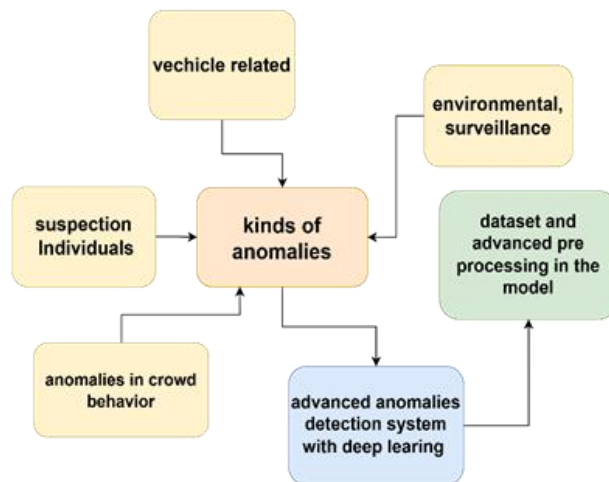


Figure 2: Structured approach for anomaly detection in surveillance.

4 CHALLENGES AND SOLUTIONS IN ANOMALY DETECTION

Although the real-time anomaly detection systems provide several benefits, several challenges need to be addressed in order to ensure that such systems operate to their full potential. The first of these challenges is the computational complexity in handling large video feeds in real time. Deep learning models, particularly CNNs and RNNs, are computationally expensive as they consume many resources for the analysis of video frames and temporal sequences. This can be challenging in the resource-constrained environment of, for example, edge devices or remote monitoring systems. To address this, different optimisation techniques such as pruning and quantisation are being used, which reduce model size and computational requirements with no loss in accuracy [19]. Edge computing is further utilised to process data at the local level, avoiding latency and minimising the amounts of data that have to be transferred to a server at a central location. Data scarcity and imbalance present another challenge. An anomaly is, by definition, not frequently occurring; training deep learning models on imbalanced datasets might sometimes result in poor performance of the model. The model may become biased to detect normal behaviour and ignore the rare but important anomaly [20]. Synthetic data generation techniques are used to create artificial anomalies, thus balancing the dataset and enhancing the ability of the model to detect rare events. Semi-supervised learning approaches are also explored to leverage both labelled and unlabelled data, further enhancing the model's accuracy and robustness. There are ethical concerns in the deployment of anomaly detection systems, specifically those concerning privacy and surveillance misuse [21]. It is, therefore, essential that the collection of data will be used ethically and responsibly so that the system can gain the trust from the public and long-term acceptance. Additionally, anomaly detection systems have to balance accuracy with latency in critical applications such as health care and traffic management. It requires optimised models that should balance these factors so that responses are correct and quick in real time.

Table 3 addresses the challenges in deploying anomaly detection systems and proposed solutions.

5 FUTURE DIRECTIONS: DEVELOPMENT OF SMARTER SURVEILLANCE SYSTEMS

Thus, future progress in anomaly detection in surveillance systems will depend on advancements in AI and deep learning technologies. For example, performance with sequential data analysis appears robust in more sophisticated models like transformers, which could be used to advance the state of anomaly detection. However, the development of self-supervised learning methods, capable of learning based on unlabelled data, may bring some further benefits toward advancing anomaly detection without strong dependence on large labelled datasets. Multimodal data with video, audio, and sensor data can also integrate to enhance the accuracy in detection [25]. For instance, it can be a system with visual cues integrated with audio analysis, which can actually detect not only visual anomalies but also auditory cues, such as screams or sirens commonly associated with emergency situations. The fusion of these kinds of modalities would enable a more comprehensive analysis of the environment, thereby boosting overall detection capabilities [26]. As these systems advance further, their applications are going to spread even more, allowing for versatility and reliability across different fields. Automating anomaly detection helps reduce the necessity of always watching, hence more precise and timely intervention. There is tremendous potential in this transformation promising to bring about safer, more efficient, and responsive public space [27]. In conclusion, deep learning-based real-time anomaly detection powered by CNNs and RNNs represents a major shift in the development of automated surveillance.

Such systems can hence respond more quickly and powerfully to threats and anomalies to enhance safety and operationally efficient activity across divergent sectors. The prospects for the advancement of anomaly detection in surveillance systems, despite the serious challenges entailed by computation requirements, data imbalance, and other problems of data privacy, are bright.

Table 3: Challenges and proposed solutions for anomaly detection systems.

Challenge	Details	ProposedSolution	References
Computational Complexity	Handling large videofeeds in real time	Optimization techniques likepruning, quantization, and edge computing	[22]
Data Scarcity and Imbalance	Lack of sufficient anomaliesfor training	Synthetic datageneration, semi-supervised learning	[23]
Privacy and Ethical Concerns	Risk of misuse in continuous monitoring systems	Transparent data policies,adherence to regulations (e.g., GDPR)	[24]
Advanced AI Models	Incorporating transformer models for sequential data analysis	Improved anomaly detection with higher accuracy	[28]
Self- Supervised Learning	Leveragingunlabelled data for anomaly detection	Reduced dependency onlabelled datasets	[29]
Multimodal Data Integration	Combining video, audio,and sensor data	Comprehensive anomaly detection (e.g., combining screams with visual cues)	[30]
Edge Computingand Local Processing	Processing closer to data sources	Reduced latency, faster response times	[31]

Table 4: Future advancements in smarter surveillance for anomaly detection.

Future Direction	Details	PotentialBenefits	References
Advanced AI Models	Incorporating transformer models for sequential data analysis	Improved anomaly detection with higher accuracy	[28]
Self- Supervised Learning	Leveraging unlabeled data for anomaly detection	Reduced dependency onlabelled datasets	[29]
Multimodal Data Integration	Combining video, audio,and sensor data	Comprehensive anomaly detection (e.g., combining screams with visual cues)	[30]
Edge Computingand Local Processing	Processing closer to data sources	Reduced latency, faster response times	[31]

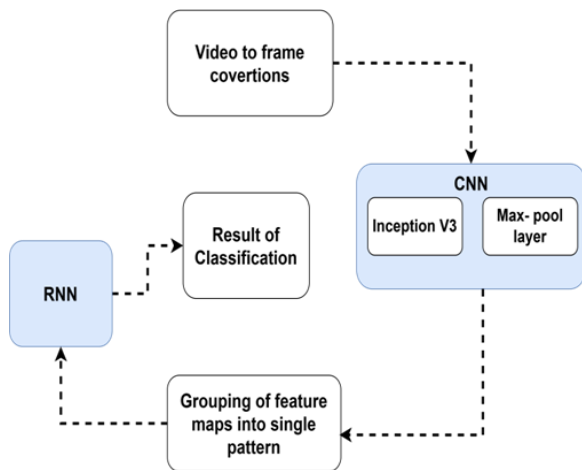


Figure 3: Workflow of video classification using CNN and RNN.

Figure 3 depicts the video classification pipeline. Video frames are processed through a CNN (Inception V3 with max-pooling layers) for feature

extraction, followed by an RNN for temporal analysis. The pipeline outputs the final classification based on spatial and sequential patterns

6 CONCLUSIONS

This paper proposes a deep learning-based framework for real-time anomaly detection, which overcomes the shortcomings of traditional surveillance systems. Traditional CCTV systems rely on human operators to monitor video feeds, which is inefficient and prone to errors. The proposed system combines CNNs for spatial feature extraction and RNNs for temporal sequence analysis, providing a comprehensive approach to anomaly detection. The use of CNNs allows the system to detect the spatial features in the video with high accuracy, like fights, accidents, or falls. On the other hand, RNNs enable the detection of temporal patterns and recognition of events changing over time, like escalating conflicts or continuous accidents. The system has been robust

and scalable in experimental testing, showing its ability to detect anomalies in various real-time scenarios such as public safety, healthcare, and traffic management. The real-time capabilities of the system would ensure that alerts are set off at the right instance, which enables immediate action based on necessity. This would facilitate better security and resource utility since false alarms are minimal, and operators are alert only to actual threats. Despite its promising performance, the framework still has some limitations. Scalability and noise sensitivity are challenges, especially for large-scale data coming from multiple cameras in dynamic environments. Future research should be directed towards integrating multi-camera systems in order to improve the coverage and accuracy of the system. In addition, to achieve reduced dependency on labelled data and increased adaptability toward the new environment, it will be considered reducing unsupervised learning techniques. Furthermore, data processing would also be done closer to the source, based on leveraging edge computing, to further enhance system performance and latency.

REFERENCES

- [1] L. Sharara et al., "A real-time automotive safety system based on advanced AI facial detection algorithms," *IEEE Transactions on Intelligent Vehicles*, May 2023.
- [2] T. F. Bukht, H. Rahman, M. Shaheen, A. Algarni, N. A. Almujaally, and A. Jalal, "A review of video-based human activity recognition: Theory, methods and applications," *Multimedia Tools and Applications*, vol. 2024, pp. 1–47, Jul. 2024.
- [3] K. Rezaee, S. M. Reza khani, M. R. Khosravi, and M. K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Personal and Ubiquitous Computing*, vol. 28, no. 1, pp. 135–151, Feb. 2024.
- [4] K. Alomar, H. I. Aysel, and X. Cai, "RNNs, CNNs and Transformers in human action recognition: A survey and a hybrid model," *arXiv preprint arXiv:2407.06162*, Jun. 2024.
- [5] A. Diro, S. Kaisar, A. V. Vasilakos, A. Anwar, A. Nasirian, and G. Olani, "Anomaly detection for space information networks: A survey of challenges, techniques, and future directions," *Computers & Security*, vol. 139, p. 103705, Apr. 2024.
- [6] A. A. Musa, A. Hussaini, W. Liao, F. Liang, and W. Yu, "Deep neural networks for spatial-temporal cyber-physical systems: A survey," *Future Internet*, vol. 15, no. 6, p. 199, May 2023.
- [7] Y. He and L. Xiao, "Structured pruning for deep convolutional neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Nov. 2023.
- [8] D. Chakraborty, W. Chiracharit, and K. Chamnongthai, "Semantic scene object-camera motion recognition for scene transition detection using dense spatial frame segments and temporal trajectory analysis," *IEEE Access*, Feb. 2024.
- [9] R. Khaldi, A. El Afia, R. Chiheb, and S. Tabik, "What is the best RNN-cell structure to forecast each time series behavior?," *Expert Systems with Applications*, vol. 215, p. 119140, Apr. 2023.
- [10] W. Qin, B. Zou, X. Li, W. Wang, and H. Ma, "Micro-expression spotting with face alignment and optical flow," in *Proc. 31st ACM Int. Conf. on Multimedia*, Oct. 2023, pp. 9501–9505.
- [11] U. Bibi et al., "Advances in pruning and quantization for natural language processing," *IEEE Access*, Sep. 2024.
- [12] J. Rane, S. K. Mallick, O. Kaya, and N. L. Rane, "Artificial intelligence, machine learning, and deep learning in cloud, edge, and quantum computing: A review of trends, challenges, and future directions," *Future Research Opportunities in Artificial Intelligence Industry*, vol. 4, pp. 2–, 2024.
- [13] I. E. Nizza et al., "Examining change in the mental health of young people with epilepsy following a successful psychological intervention," *Clinical Child Psychology and Psychiatry*, vol. 29, no. 4, pp. 1400–1416, Oct. 2024.
- [14] Y. Berhanu, E. Alemayehu, and D. Schröder, "Examining car accident prediction techniques and road traffic congestion: A comparative analysis of road safety and prevention of world challenges in low-income and high-income countries," *Journal of Advanced Transportation*, vol. 2023, no. 1, p. 6643412, 2023.
- [15] R. S. Jebur, M. H. B. M. Zabil, L. K. Cheng, and D. A. Hammood, "Image denoising in deep learning: A comprehensive survey," *EETJ*, vol. 1, no. 1, pp. 1–12, Jun. 2024.
- [16] G. A. Carlson et al., "Narrative review: Impairing emotional outbursts: What they are and what we should do about them," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 62, no. 2, pp. 135–150, Feb. 2023.
- [17] W. Du, A. Dash, J. Li, H. Wei, and G. Wang, "Safety in traffic management systems: A comprehensive survey," *Designs*, vol. 7, no. 4, p. 100, Aug. 2023.
- [18] T. Shaik et al., "Remote patient monitoring using artificial intelligence: Current state, applications, and challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 2, p. e1485, Mar. 2023.
- [19] T. G. Mondal, J. Y. Chou, Y. Fu, and J. Mao, "A hybrid deep neural network compression approach enabling edge intelligence for data anomaly detection in smart structural health monitoring systems," *Smart Structures and Systems*, vol. 32, no. 3, pp. 179–193, Sep. 2023.

- [20] G. Pang, C. Shen, L. Cao, and A. van den Hengel, "Deep learning for anomaly detection: A review," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–38, Mar. 2021.
- [21] J. Madeira, "Ethical considerations in deploying IoT sensors for anomaly detection in autonomous vehicles," *Journal of AI-Assisted Scientific Discovery*, vol. 3, no. 2, pp. 1–3, Jun. 2023.
- [22] H. F. Jameel, "Real-time sitting correction using Arduino-based ultrasonic sensor at the workplace," *JT*, vol. 6, no. 4, pp. 19–26, Dec. 2024.
- [23] A. Hakami, "Strategies for overcoming data scarcity, imbalance, and feature selection challenges in machine learning models for predictive maintenance," *Scientific Reports*, vol. 14, no. 1, p. 9645, Apr. 2024.
- [24] V. R. Boppana, "Data ethics in CRM: Privacy and transparency issues," *MZ Computer Journal*, vol. 4, no. 2, Nov. 2023.
- [25] Z. Zhang, S. Zhang, D. Ni, Z. Wei, K. Yang, S. Jin, G. Huang, Z. Liang, L. Zhang, L. Li, and H. Ding, "Multimodal sensing for depression risk detection: Integrating audio, video, and text data," *Sensors*, vol. 24, no. 12, p. 3714, Jun. 7, 2024.
- [26] Q. Tang, J. Liang, and F. Zhu, "A comparative review on multi-modal sensors fusion based on deep learning," *Signal Processing*, p. 109165, Jul. 3, 2023.
- [27] A. Diro, S. Kaisar, A. V. Vasilakos, A. Anwar, A. Nasirian, and G. Olani, "Anomaly detection for space information networks: A survey of challenges, techniques, and future directions," *Computers & Security*, vol. 139, p. 103705, Apr. 1, 2024.
- [28] P. Yan, A. Abdulkadir, P. P. Luley, M. Rosenthal, G. A. Schatte, B. F. Grewe, and T. Stadelmann, "A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions," *IEEE Access*, Jan. 2, 2024.
- [29] H. Taherdoost, "Beyond supervised: The rise of self-supervised learning in autonomous systems," *Information*, vol. 15, no. 8, p. 491, Aug. 16, 2024.
- [30] K. U. Duja, I. A. Khan, and M. Alsuhaibani, "Video surveillance anomaly detection: A review on deep learning benchmarks," *IEEE Access*, Nov. 5, 2024.
- [31] S. Nair, "Beyond the cloud-unraveling the benefits of edge computing in IoT," *International Journal of Computer Engineering and Technology*, vol. 14, pp. 91–97, 2023.

Qualified Investigation of ML Models for Forecasting SPEC Benchmark Performance

Vedavyas Gurla¹, Sarlan Sarlan², Ilham Ilham², Aaluri Seenu³, Vasavi Oleti⁴, Danampalli Sravanthi⁵
and Hassnien S. Al Hashemi⁶

¹Department of CSE, Geethanjali College of Engineering and Technology, 501301 Hyderabad, India

²Department of Chemistry Study Program, Universitas Sulawesi Tenggara, 93121 Kendari, Indonesia

³Department of CSE, Shri Vishnu Engineering College for Women, 534204 Bhimavaram, India

⁴KKR and KSR Institute of Technology and Sciences, Vinjanampadu, 522017 Vinjanampadu, Andhra Pradesh, India

⁵Department of Computer Science Engineering- Data Science, MLR Institute of Technology, 500043 Hyderabad, Telangana, India

⁶Department of Medical Lab., Dijlah University College, 10021 Baghdad, Iraq

gvedavyas.cse@gcet.edu.in, sarlan@un-sultra.ac.id, ilham@un-sultra.ac.id, aaluriseenu@svecw.edu.in,

vasavioleti26@gmail.com, dsravanthi@mlrit.ac.in, hassnhashmi@duc.edu.iq

Keywords: ML, SPEC CPU2025, Model Metrics, Optimization.

Abstract: Performance prediction via simulations is laborious and tedious. To avoid this problem, one way is to use supervised learning to forecast how well a system will do on SPEC benchmarks. This year's SPEC CPU includes a publicly available collection of results from 43 standardized performance tests divided into 4 suites and run on a variety of hardware setups. In this study, we will examine the dataset and try to find the answers to these questions: Can we reliably forecast the SPEC outcomes from the dataset's setups, without actually running the benchmarks? Secondly, which software and hardware aspects are most crucial? On the third point, in regard to forecast time and inaccuracy, which hyperparameters and models work best? thirdly, is it possible to use historical data to foretell how future systems will be performing? Preparing data, choosing features, this talk covers a wide range of topics, including hyperparameter tuning, employing decision trees, random forests, multi-layer perceptrons, and multi-task elastic-net neural networks to evaluate regression models, and more. There are three stages to feature selection: deleting features with zero variance, removing features with strong correlation, and finally, using Functional Recursion Using permutation importance, elastic-net coefficients, or importance metrics depending on trees to filter out candidates. Searching the hyperparameter space with a grid, we select the best models. Afterwards, we compare and evaluate their performance. We prove that using the initial set of 29 features in tree-based models yields 4% or better accuracy in predictions. With 10 characteristics, both the Random Forest and Quick Decision Tree models keep their average errors at 5% and 6%, respectively.

1 INTRODUCTION

Research on the use of Machine Learning (ML) to estimate computer systems' performance and enhance system design is ongoing [1]. System designers and engineers can learn more about how different configuration changes affect system performance and use that information to make more informed design decisions with the predicted results. Before developing their solutions, vendors research the market to achieve ideal positioning. Predicting how well future system configurations will perform is, nevertheless, no easy feat [2].

People also look for optimal system setups to maximize performance or make logical purchases to enhance the cost-performance ratio. Customers may need performance statistics for new systems or configurations they have never seen before. Therefore, it is insufficient to only possess access to the outcomes of various workloads executed on large clusters of computers [3]. Research into performance prediction and evaluation is driven by these difficulties. Nevertheless, it is difficult to accurately forecast the performance of unknown setups when doing many tasks simultaneously, in terms of either execution time or throughput. It could necessitate the

ever-more-complicated process of exact analytical modeling of future systems brought about by developments in computer architecture. In addition, the most accurate forecasts may not be produced by modeling techniques that depend on laborious and comprehensive simulations [2], [4]. Therefore, we depend on regression models to forecast performance instead of fine-grained system modeling [5]. These models figure out how different system configurations affect how well they handle different kinds of workloads. Included in SPEC CPU2017 are suites of widely used compute-intensive benchmarks that primarily evaluate compilers, memory subsystems, and processor characteristics. There are four different benchmark suites that SPEC offers, each with its own set of practical, portable programs that address problems of varying magnitude [6]:

First, the floating point rate, FP_rate; second, the integer rate, Int_rate; and lastly, the integer speed, Int_speed. We construct supervised learning models based on software and hardware characteristics extracted from the SPEC CPU2017 public dataset, which contains benchmarks of systems used for computer analysis. This document contains the results of the benchmarking procedure. The Multitask Elastic-Net (MT_EN), Decision Tree (DT), Random Forest (RF), and Multi-layer Perceptron (MLP) estimators have formed the basis of our models.

Prediction accuracy of the SPEC benchmarks has been the subject of prior research, with a focus on neural networks. Conversely, the features' significance in relation to the prediction models' contributions has been disregarded or ignored. The primary objective of our research is to establish a machine learning pipeline that can rapidly and accurately predict SPEC CPU2017 performance using regression models, and to give a simplified method for thorough evaluation. The significance of numerous software and hardware features is highlighted in this paper, which also evaluates several models' latency and prediction error using both the full and selected feature sets. Researchers are also looking at the possibility of using historical data from current systems to foretell how those systems will perform in the future. The scikit-learn [7] ML package is utilized in the Python-written open-source code. For example, engineers may be able to narrow the design space with the help of the presented ML pipeline and analysis, and consumers may be able to give more weight to crucial aspects when making purchases.

The benchmark performance ratios are the goal variables (outputs) for our regression models. Ratings

and rankings on performance ratios for speed and rate standards as shown in Figure 1.

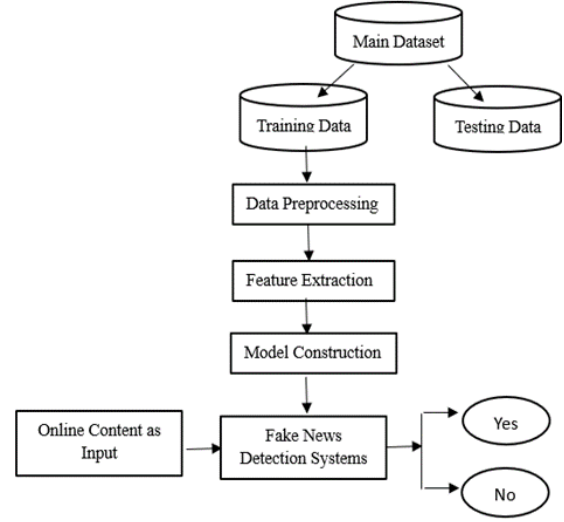


Figure 1: Fake news flowchart diagram.

$$Speed_{Ratio} = t_{Ref}/t_{SUT}$$

$$rate_{Ratio} = n \times (t_{Ref}/t_{SUT})$$

Where tRef is the time spent on a reference computer with one thread (for speed benchmarks) or one copy of the benchmark (for rate benchmarks), tSUT is the time spent on the machine being tested with multiple threads or copies, and n is the number of benchmark copies.

2 LITERATURE REVIEW

2.1 Computer Architecture and Systems Machine Learning Surveys

Computer systems and architecture have long been fine-tuned to run machine learning (ML) models efficiently. Rethink the relationship among ML and systems to transform computer architecture and system design. Both the completion of the virtuous cycle and an increase in designers' productivity are embraced by this. Our goal in this article is to provide a thorough overview of the literature on machine learning (ML) in relation to computer system design and architecture. Our first step is to create a taxonomy at a high level by thinking about the two most common uses of ML techniques in the field of architecture and system design: rapid predictive modeling and design process. Following that, we

provide a brief overview of the most popular ML strategies used to address the most prevalent issues in computer architecture and system design, as well as the problems themselves. We focus on computer architecture narrowly and assume data centers are gigantic computer warehouses. We also give some cursory attention to nearby computer systems like code generation as well as compiler, and they focus on how ML can transform design automation. We also see excellent future courses and prospects, and we think the community would benefit much from using ML to computer architecture and systems.

2.2 Specnet Predicts SPEC Scores Utilizing Deep Learning

The SPECnet is a deep neural network (DNN) that we demonstrate how to construct for the purpose of predicting SPEC® results. The SPEC CPU2006 suite has been around for over a decade (it was decommissioned in January 2018), and there are still thousands of submissions for its integer and floating point benchmarks. In order to train on the matching reported SPEC scores, we construct a DNN that takes hardware and software information from these submissions.

Then, for future machine configurations, we forecast scores using the learned DNN. With training and development/test errors ranging from 5% to 7%, we attain very high prediction accuracy rates of 93% to 95%. By meticulously modeling the performance of both the essential and non-essential parts of the system, it is possible to obtain an accuracy level of 97% to 98%, which is extremely close to what humans are anticipated to achieve. Applying SPECnet to SPECComp2012 and SPECjbb2015 is an additional step beyond the CPU2006 suite.

We demonstrate that such a DNN can also reasonably forecast (~85% accuracy) for these benchmarks, even though there are only hundreds of reported submissions for these suites. Our SPECnet solution is highly adaptable and expandable, built on top of the cutting-edge Tensor flow framework.

2.3 Using Public Datasets to Predict Workload or CPU Performance

Many different versions of general-purpose microprocessors are available on the market, all with identical functionality but different power consumption, frequency CPU cores, cache size, and memory bandwidth. Both the microarchitecture and the workloads being executed have an impact on their performance. The customer can't make an informed

purchase decision without knowing the performance and pricing details, given a set of expected workloads. There are a plethora of benchmark suites available for use in evaluating CPU performance, and the outcomes for extensive sets of CPUs are frequently made public. Consumers seeking processor or workload efficiency statistics may locate repositories of benchmark results to be unhelpful at times. In addition, benchmark suites that aim to cover a wide variety of workloads could provide deceptive aggregate scores. then built a DNN model to understand the correlation between Intel CPU specs and Geekbench 3 and SPEC CPU 2006 performance, and then applied it to these issues. We demonstrate the ability to produce practical forecasts for novel processors and workloads. The two benchmark suites' performance scores are also compared and cross-predicted. This is the first literature to quantify such suites' self-similarity by results. This work aims to dissuade buyers from relying just on Geekbench 3, while simultaneously motivating researchers to assess their work using a wider range of workloads rather than just the SPEC CPU suites.

2.4 Machine Learning to Anticipate Computer System Design Performance Options

Computer companies invest a lot of time, energy, and capital into developing new systems and configurations; the success of these systems determines how much they can cut costs, how much they can charge, and how much market share they can win. Our primary focus in this effort is to streamline the architectural design and system design phases of parallel computers. Our approach uses neural network and linear regression models to predict the performance of any machine in the design space by extracting the performance levels of a tiny percentage of machines. We demonstrate that in the context of architectural design, a 3.4% error rate may be achieved by utilizing a small percentage of the design space (cycle-accurate simulations) to predict its performance completely.

We use Standard Performance Evaluation Corporation benchmark data to predict future system design performance. Focusing on systems with many processors, we demonstrate that our models can, on average, forecast future system performance to within 2.2% of the true value.

We are confident that these tools can greatly accelerate the exploration of the design space while also helping to reduce the associated research and development costs and time-to-market.

2.5 A Survey on Multi-Output Regression

A multitude of methods have been put forward in the past few years to tackle the ever-increasingly difficult undertaking of multi-output regression. Modern multi-output regression techniques, including those for issue transformation and algorithm adaptation, are reviewed in this article. Additionally, we showcase open-source software frameworks Public data sets for multi-output regression real-world challenges and the most prevalent performance evaluation methods.

3 METHODOLOGY

The reliability of neural network predictions on the SPEC benchmark has been the subject of earlier research. In addition, features' significance in contributing to the models has been disregarded or ignored. Our research aims to construct a machine learning pipeline that can rapidly and accurately predict SPEC CPU2017 performance using regression models, and then we will evaluate these models thoroughly. Finding which software and

hardware features are most important and Evaluating prediction error and delay of various models on complete and decreased feature sets are two significant contributions of this work as shown in Figure 2.

Here, we forecast SPEC benchmark performance using supervised learning. SPEC is the Performance Evaluation Corporation. The public SPEC CPU2017 dataset contains 43 standardized performance tests from diverse system settings. These tests are organized into 4 suites. The following questions are intended to be answered by this paper's analysis of the dataset: I) Is it necessary to run the benchmarks in order to provide an accurate prediction of the SPEC results using the dataset's configurations? II) Which software and hardware features are most crucial?

In regards to prediction error and time, which hyperparameters and models provide the greatest results? and (IV) is it possible to use historical data to forecast how future systems will perform? We walk you through data preparation, feature selection, hyperparameter tuning, and regression model evaluation employing Multi-Task Elastic-Net, Decision Tree, Random Forest, and Multi-Layer Perceptron neural network estimators [8].

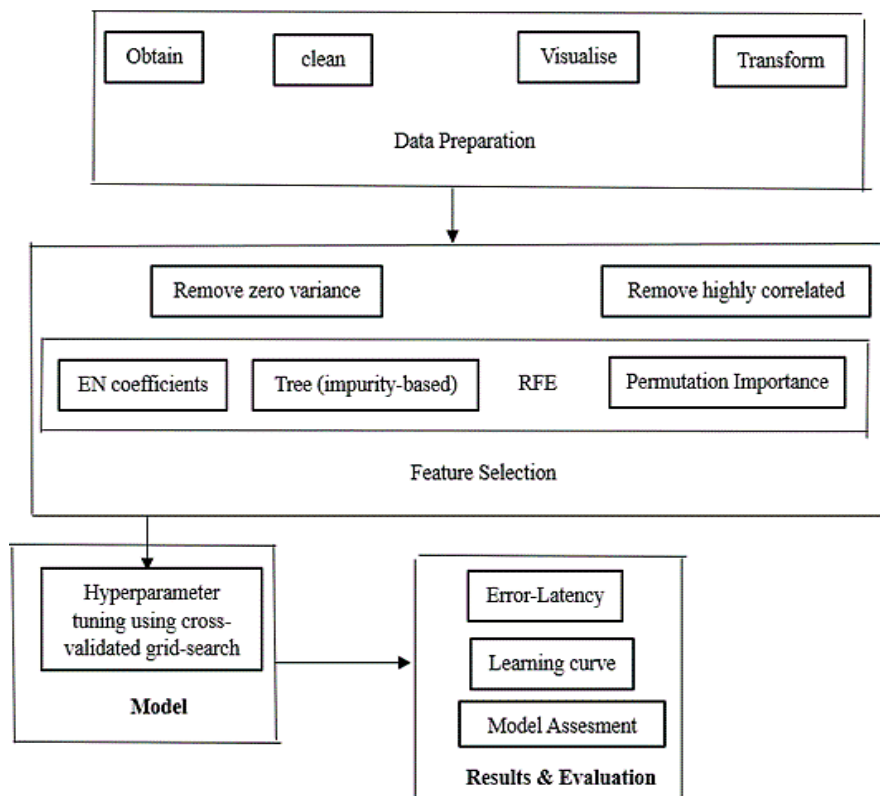


Figure 2: System architecture.

3.1 Modules

We designed the following modules for the project:

- Use the data exploration module to load data into the system.
- The module will be used to read and process data.
- Data would be separated into train and test using this module.
- Constructed models using ElasticNet, Decision Tree, Random Forest, NLP, and Voting Stacking. Algorithm accuracy calculated
- The user signup and login module allows for registration and login.
- User input: This modules provides prediction input.
- Prediction: final prediction showed.

Table 1: Feature set extracted from the publicly available SPEC CPU 2017 database.

Raw Feature	Description
arch	Architecture from lscpu
Nominal_mhz	Processing unit nominal clock frequency
max_mhz	Central processing unit maximum clock frequency
cpus	Number of CPU(s) from lscpu
threads_per_core	Thread(s) per core from lscpu
cores_per_socket	Core(s) per socket from lscpu
sockets	Socket(s) from lscpu
numas	NUMA node(s) from lscpu
l1d_cache_kb	L1 data cache in KB
l1i_cache_kb	L1 instruction cache in KB
l2_cache_kb	L2 cache in KB
l3_cache_kb	L3 cache in KB
mem_kb	Main mem in KB (get it from the Memory field, if there is no /proc/meminfo)
mem_channels	Number of memory channels
channel_kb	Memory channel's capacity
mem_data_rate	Memory transfer rate in MT/s
os	Operating System

Table 1 lists all twenty-one raw features. Data related to the hardware includes the design, processor(s), and memory subsystem. Software

specifics include the operating system, compiler, file system, parallel flag, and thread/copy count. To make them numeric, the operating system, compiler, and file system variables are categorical. Input variables might affect output depending on other input factors. For simplicity, this is the interaction effect. The equation $cpus = sockets \times cores\_per\_sockets \times threads\_per\_core$ illustrates a three-way relationship [9].

4 IMPLEMENTATION

4.1 Algorithms

4.1.1 Multitask ElasticNet

Regularizing regression models with lasso and ridge penalties is elastic net linear regression. By learning from lasso and ridge regression's flaws, the method improves statistical model regularization [10].

4.1.2 Decision Tree

Decision trees are non-parametric supervised learning algorithms for classification and regression. Its tree structure contains a root node, branches, internal nodes, and leaf nodes.

4.1.3 Random Forest

The popular supervised machine learning technique Random Forest technique is used for classification and regression. The more trees a forest has, the stronger it is.

4.1.4 NLP

NLP algorithms are usually machine learning-based. NLP may use machine learning to automatically learn huge sets of rules by studying a corpus (like a book) and generating a statistical conclusion.

4.1.5 Voting Stacking

How votes are aggregated is the main distinction from stacking. Voting aggregates classifiers using user-specified weights, while stacking uses a blender/meta classifier as shown in Figure 3.

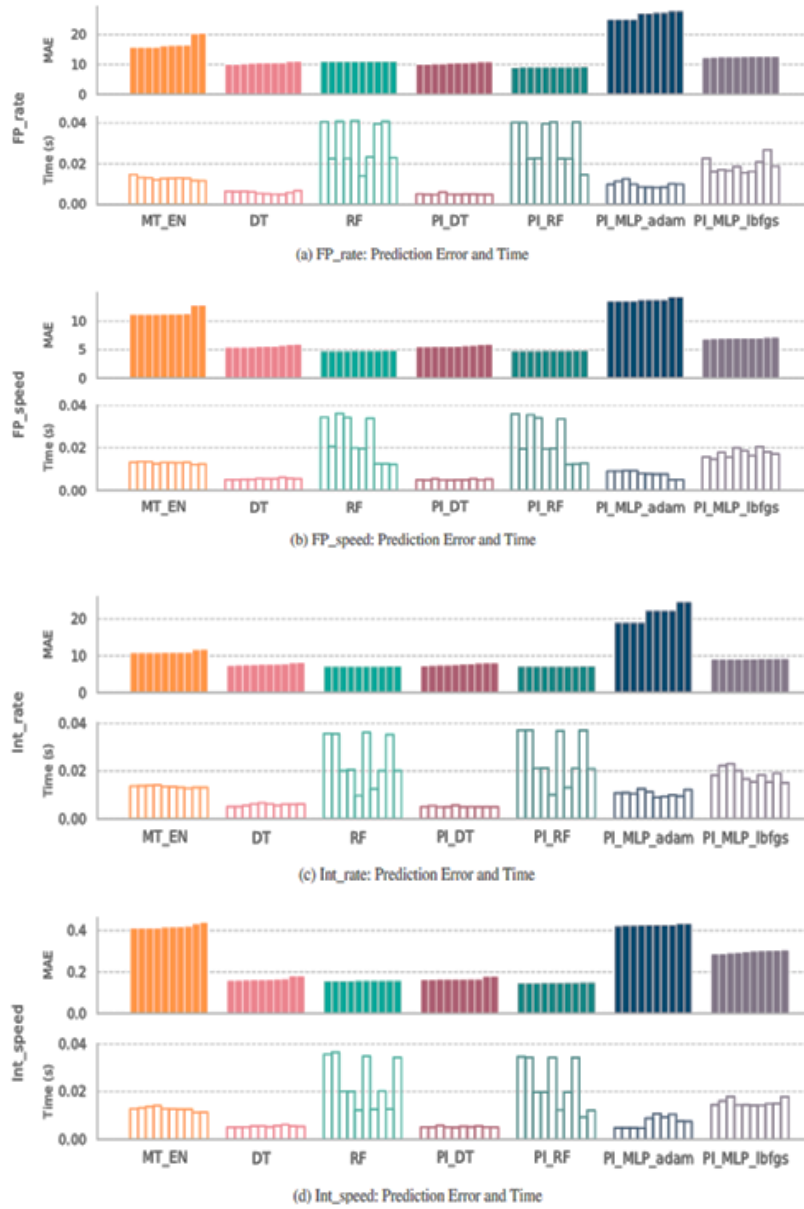


Figure 3. Comparison of forecasting time (in seconds) and mean cross-validation error (MAE) for the evaluated models. Lower values on both metrics indicate better predictive performance and computational efficiency.

5 EXPERIMENTAL RESULTS

To evaluate the developed system, a set of user interface screens and prediction workflows were tested. The system begins with a home screen that introduces the prediction tool and serves as the entry point for the user (Fig. 4). Following this, users are directed to a registration interface that supports the creation of user accounts and secure logins (Fig. 5). Once registered, users access the main interface

where model selection, data upload, and configuration options are provided (Fig. 6). Here, users can upload their input feature values — these include system configuration details such as CPU parameters, memory details, and software-related metadata (Fig. 7). Upon submission, the system processes the input through the pre-trained models and generates performance predictions for the SPEC CPU2017 benchmark, which are then displayed clearly in the prediction output screen (Fig. 8).



Figure 4: Home screen.



Figure 5: User registration.

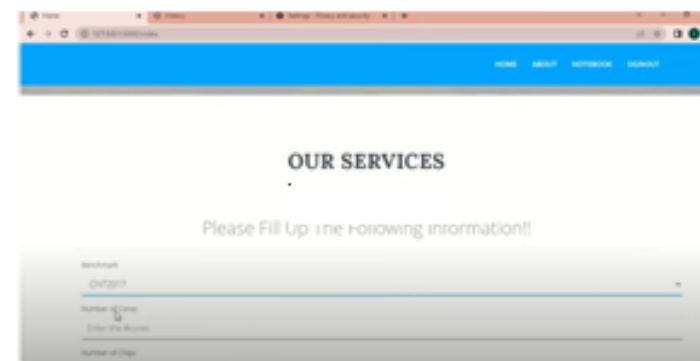


Figure 6: Main page.

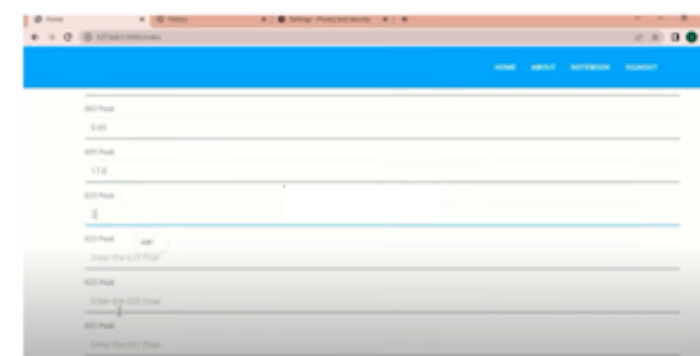


Figure 7: Upload input values.



Figure 8: Prediction result.

The prediction results demonstrate the feasibility and accuracy of our regression models in estimating benchmark performance without executing the full test suite. Tree-based models like Decision Trees and Random Forests consistently outperformed others in terms of both prediction accuracy and execution time. Importantly, the experimental interface enabled seamless interaction with the pipeline and visualization of results, validating both the backend model's performance and the usability of the developed platform.

6 CONCLUSIONS

This research looks at the possibility that supervised learning can forecast how well parallel systems will do on the SPEC benchmarks without actually running the benchmarks themselves. Extensive testing has proven that SPEC CPU2017 parallel and concurrent performance predictions are feasible. We have used grid search to discover the top ten most accurate models after investigating how changing the hyperparameters of the aforementioned four estimators affected their performance. Then, we compared the models' prediction delay and error rates. The most effective models have been those based on trees. They also find that smaller feature sets yield better results when using RFECV for feature selection. Additionally, in order to assess the accuracy of our data-driven performance predictions, we have examined learning curves of top tree-based models. With just 10% of the historical data used as the training set, we were able to demonstrate that this dataset is sufficiently predictive, but with 70% or more data, we were able to reduce 2-3 times the mean error. This dataset is four years old. It would be interesting to see how the dataset and fresh system generations affect the figures. As the final stage, we

compared the final models' average goodness-of-fit (R2) and MAPE on the put-aside test set. Tree-based models (DT and RF) fared better overall and for specific benchmarks in R2 and MAPE. An explanation that stands out from the linear models is that they fail to account for some non-linear correlations (MT_EN). The fact that tree-based models typically perform better when many types of features are present is a probable reason, especially when contrasted with the neural networks MLPs. However, with thousands of samples or more, neural networks may do better. Once again, it will be fascinating to watch as the dataset grows and how these results evolve. With just 29 attributes, Decision trees and random forests may have MAPEs sub 4%. Models with fewer features benefit from random forests, which perform better with 10 features ($1.5\% < \text{MAPE} < 4.5\%$ over four suites). On the other hand, decision trees are the way to go if interpretability is your top priority. Our regression models are more interpretable, have a better track record of properly predicting the SPEC CPU benchmarks, and provide more light on which software and hardware features are most important than in earlier research. By applying the RFE technique, we were able to determine that just a small subset of the available software and hardware attributes (less than five) are crucial to our models, and that a mere ten features are sufficient to produce very accurate predictions on this dataset. Our study provides an effective performance prediction, evaluation, and design space exploration pipeline.

REFERENCES

- [1] N. Wu and Y. Xie, "A survey of machine learning for computer architecture and systems," arXiv preprint arXiv:2102.07952, 2021. [Online]. Available: <https://arxiv.org/abs/2102.07952>.

- [2] D. Das, P. Raghavendra, and A. Ramachandran, "SPECNet: Predicting SPEC scores using deep learning," in *Proc. Companion ACM/SPEC Int. Conf. Performance Engineering*, Apr. 2018, pp. 29–32.
- [3] Y. Wang, V. Lee, G.-Y. Wei, and D. Brooks, "Predicting new workload or CPU performance by analyzing public datasets," *ACM Transactions on Architecture and Code Optimization*, vol. 15, no. 4, pp. 1–21, Jan. 2019.
- [4] B. Ozisikyilmaz, G. Memik, and A. Choudhary, "Machine learning models to predict performance of computer system design alternatives," in *Proc. 37th Int. Conf. Parallel Processing*, Sep. 2008, pp. 495–502.
- [5] A. Tousi and C. Zhu, "Arm research starter kit: System modeling using gem5," *Arm Research*, U.K., Jul. 2017.
- [6] J. Bucek, K.-D. Lange, and J. V. Kistowski, "SPEC CPU2017: Next-generation compute benchmark," in *Proc. ACM/SPEC Int. Conf. Performance Engineering*, 2018, pp. 41–42.
- [7] F. Pedregosa, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Oct. 2011.
- [8] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *Data Mining and Knowledge Discovery*, vol. 5, no. 5, pp. 216–233, 2015.
- [9] D. Kocev, S. Džeroski, M. D. White, G. R. Newell, and P. Griffioen, "Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition," *Ecological Modelling*, vol. 220, no. 8, pp. 1159–1168, Apr. 2009.
- [10] D. Tuia, J. Verrelst, L. Alonso, F. Perez-Cruz, and G. Camps-Valls, "Multioutput support vector regression for remote sensing biophysical parameter estimation," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, no. 4, pp. 804–808, Jul. 2011.

Intelligent Skin Cancer Detection from Dermoscopic Images with Machine and Deep Learning Approaches

Ali Abdul Karim Jawad and Asaad Abbas Hindi

*Department of Geography, University of Basrah, 61004 Basrah, Iraq
pgs.ali.abdulkarim@uobasrah.edu.iq, asaad.hindy @uobasrah.edu.iq*

Keywords: Skin Cancer Detection, Deep Learning, Dermoscopic Images, Machine Learning, Early Diagnosis.

Abstract: Melanoma, in particular, is one of the most common and dangerous cancers in the world, and early diagnosis is critical to improving survival rates. A traditional diagnostic method, such as visual examination or dermoscopy, often requires expert intervention, but it can be challenging to distinguish early-stage melanoma from benign lesions. Artificial intelligence (AI), particularly machine learning and deep learning techniques, are applied to dermoscopic images in this study to detect skin cancer more accurately. Results showed that deep learning models were more accurate, more recallable, and had higher F scores than traditional machine learning algorithms. We compare the performance of Logistic Regression, K-Nearest Neighbors, and advanced deep learning architectures such as Xception, VGG16, and ResNet50 on two public datasets containing dermatoscopic images, HAM10000 and PH2. As a result of the study, deep learning models, especially when fine-tuned, offer significant improvements in detecting skin lesions, including melanoma, allowing for early detection.

1 INTRODUCTION

The early detection of skin cancer is critical to improving the patient's outcome since it is one of the most common types of cancer worldwide. Dermatologists rely heavily on dermoscopic images to diagnose skin cancer, which is diagnosed through examining lesions and moles. In order to detect disease more efficiently, accurately, and scalable, new detection methods are needed to cope with the growing number of patients and the complexity of some cases. AI advances, specifically machine learning (ML) and deep learning (DL), have demonstrated promise in automating and improving the detection of skin cancer from dermoscopic images. The use of machine learning models can help us differentiate benign lesions from malignant lesions [1]. Using multiple layers of neural networks, deep learning is capable of extracting complex features from raw image data automatically.

An individual's skin covers their entire exterior, making it the major organ of the body. Among the human body's first lines of defence is its skin [2]. Protection against external environmental factors, regulation of body temperature, immunity against many diseases, and enhancement of beauty are some

of its functions [3]. Sunlight exposes the skin to harmful ultraviolet rays, but it also produces vitamin D, which is imperative to human health. There are different skin colours (body pigmentation) and moisturizing levels (oily versus dry) depending on where you live in the world [4], [5]. In the presence of sunlight for a long time or when exposure to ultraviolet rays is prolonged, malignant skin diseases are more likely to develop and pigmentation declines. Detecting melanoma early is essential for its survival. As the cancer cells are similar to other skin cells, they are not detected when the disease is in its early stages. Malignant melanomas are caused by abnormal cancerous cells dividing rapidly under the surface of the skin and penetrating deep into the tissue [6].

Treatment of melanoma is challenging when the cancer spreads throughout the body. As a result, early diagnosis is crucial for saving lives. Dermatoscopic images reveal a wide variety of skin conditions. The two most common skin diseases are melanocytosis and nonmelanocytosis [7]. Skin colours vary from person to person, and air bubbles, hair, and artefacts all complicate matters. There are also many similarities between early-stage skin diseases, especially when they are in their early stages. Inadequate medical resources result in 420 million people suffering from skin diseases globally. A lack

of early diagnosis also reduces the quality of life and social advancement in developing countries due to high treatment costs [8].

More than 70,000 cancer cases were reported in the United States in 2017, according to the American Cancer Society. A total of 100,350 cases of melanoma were diagnosed in that country in 2020, with 60,190 men and 40,160 women being diagnosed [9]. Early-stage melanoma and benign lesions are difficult to distinguish due to their similar appearance during the early stages. Computer-assisted diagnosis (AI) is, therefore, essential for early diagnosis. Using computer-aided diagnosis, physicians and experts can diagnose diseases earlier and more efficiently. In this study, we used artificial intelligence techniques to detect skin lesions at an early stage. Using artificial intelligence techniques, specific samples are trained to solve particular problems using many layers and complex neurons. In recent years, deep learning techniques have been used for diagnosing problems that cannot be solved using machine learning [10], [11].

2 LITERATURE REVIEW

The deadly disease of melanomas is responsible for thousands of fatalities worldwide as a result of skin cancer. Throughout history, much has been done to detect and save the victims at the earliest stages of the disease. The first step is to introduce handcrafted features-based techniques for detecting it. In [12], edge-color histograms and local binary patterns (LBP) were used for melanoma lesion detection by using edge and color histograms alongside LBP.

Combining two different techniques to incorporate features of local and global images. The Author [13] proposed two different approaches for localizing melanoma skin cancer. Global image features such as colour, size, shape, texture, and shape of the melanoma lesion are determined using the Laplacian pyramid and a gradient histogram. On the basis of the data obtained, a binary classifier is trained. Bag-of-functions (BoF) classifiers are used to recognize local features in images [14]. Compared to texture features, colour features perform better in this study. Due to the variations in sizes, textures, shapes, and colours of melanoma moles, hand-coded feature-based approaches are often ineffective at detecting skin lesions. According to the Author [15], melanoma lesions can be classified according to the following methods. We calculated the scale-invariant feature transforms (SIFTs) and histograms of oriented gradients (HOGs) of representative features.

Melanoma moles were classified using support vector machines (SVMs) and k-nearest neighbours (KNNs) [16].

Several image-processing techniques were used by the Author [17] to identify lesions on the skin. This approach involved obtaining a medical history before diagnosing. An expert examines the attributes and provides a set of them. Descriptors are derived automatically, and those observable are chosen as a final decision. Segmentation-based techniques are well suited to situations in which the chrominance distribution is uniform and the illumination of an image changes little. Still, in real-time applications, these lighting and illumination variations are usually inevitable. Using skin lesion recognition and segmentation, the Author [18] presented his approach. The input images were then preprocessed, and lesion segmentation was performed using seed region growing and graph-cut approaches. SURF and HOG descriptors were used to compute the features from the segmented lesions. In order to categorize lesions based on the extracted features, an SVM classifier was trained using the extracted features. Although the method is better at segmenting lesions and classifying them, it is computationally expensive.

Due to ML advancements, DL methods are becoming increasingly popular since they provide high accuracy. Radiologists and dermatologists are therefore exploring the use of these techniques for medical imaging as well, allowing them to detect fatal diseases sooner. As a result of deep learning approaches like CNN, it has proven possible to classify skin lesions with good accuracy. The Author [19] suggests that instead of training the CNN from scratch to detect melanoma lesions, pretrained ConvNets are used. The Author has developed a DL-based model for segmenting and categorizing melanomas. Feature calculation from the area of interest was performed using the area-average pooling (RAPooling) method in the first step. A classification framework was proposed using the lesion regions identified in the first step, as well as a segmentation process using RAPooling to guide the classification process. A RankOpt classifier was used for the final step of melanoma classification. Although it performs well for segmenting and categorizing skin moles, it is inefficient economically [20].

Key points from the computed key points were used to train SVMs, KNNs, and Random Forests (RF) to identify melanoma from benign moles. While the method is robust, it needs further improvement to improve its performance in detecting melanoma. In his paper [21], the Author introduced a method of

localizing melanoma lesions automatically. For the first step, feature correlations were calculated using a linearly independent and linear prediction method (LP). The second step involved computing a representative set of keypoint vectors using RGB colour histograms and SIFT. The input samples were classified into melanoma and nonmelanoma categories based on the computed key points by SVM classifiers. While computationally efficient, this method may not be appropriate for samples whose chrominance changes are intense.

A variety of factors have hindered previously conducted studies, including hair, air bubbles, artefacts, and light reflections. Furthermore, similar characteristics between diseases pose a major challenge when diagnosing and separating them. There have been no differences between this study and previous studies in terms of issues addressed. The images were cleaned with average and Laplacian filters in order to remove artefacts, air bubbles, skin lines, and reflections. A high degree of accuracy can also be achieved when removing hair from images using dull razor technology. Several diseases share similar features, so these features were extracted from each image and combined into one vector using three hybrid algorithms. Consequently, each disease has its characteristics. A deep and representative characteristic of each disease was extracted using ResNet-50 and AlexNet.

3 PROPOSED METHODOLOGY

3.1 Methods and Materials

In this study, not only melanoma and nevi are distinguished from pigmented lesions that commonly appear in clinics, but also no melanocytic pigments. With the help of two publicly available datasets, pre-trained CNNs are used to classify skin cancer images. First, we have the HAM1000 dataset [22]. This dataset is a benchmark for academic machine learning.

In addition, there is the PH2 dataset [23]. In this database, 200 dermatoscopic images of lesions with melanocytic cells are included. This database contains 80 samples of atypical nevi (Atyp NV) and 40 samples of melanoma (Mel), which we used to expand our diagnostic categories. A 768 x 560-pixel resolution RGB image in 8-bit colour. This paper contains 8 diagnostic classes in total.

Researchers have developed CNN in recent years to solve computer vision problems with greater accuracy. Our paper uses 4 deep CNN trained on

ImageNet [24] using Tensor Flow [24], a framework developed by Google for deep learning [25].

3.2 Artificial Intelligence

In addition to planning, problem-solving, understanding natural language, and learning, machines can also perform some other tasks characteristic of human intelligence thanks to artificial intelligence. Our discussion of artificial intelligence will be divided into two main categories. There is also a discussion of pre-trained networks and transfer learning (TL).

3.2.1 Machine Learning

In machine learning, useful information is extracted from data by separate algorithms. After training, machine learning extracts salient features and natural patterns that are used to assign attributes to samples or cluster them based on cluster identifications. Models can be used to predict data that was previously unknown based on this information. We will outline the main ML models in the following sections:

- Decision Trees. DTs can be used for categorical as well as numerical variables in ML since they do not require assumptions regarding data distribution or classifier structure. Consequently, they are suitable for classification, regression, and multi-output analysis. Datasets of any size and complexity can be classified effectively and efficiently using them. Decision trees are assembled into random forests (RFs). When DTs are multiplied, the variance of each DT is reduced, resulting in RFs that are more robust and generalizable.
- Support Vector Machines. An SVM can be used to perform linear and nonlinear classifications (kernel methods) and regressions, as well as to detect outliers. In spite of this, it is important to keep in mind that SVMs perform best when the dataset is complex. As a result of using SVMs for classification, hyperplanes are constructed. As a result of every hyperplane, two distinct classes of features can be separated and differentiated from one another.
- K-Nearest Neighbors. Using KNN, new data is classified based on how similar it is to the nearest labelled data. As soon as a KNN classifier's parameters, namely the number of

nearest neighbours and the distance metric, are chosen (Euclidean and Manhattan distances are most relevant), a majority vote is used to assign a label to the new data set. Typically, K is set between 3 and 10 to prevent overfitting and underfitting.

- Artificial Neural Networks. ANNs were introduced as an alternative method of solving complex problems based on neuronal connections studies. Despite early attempts to mimic the human brain and synaptic connections between neurons, only a small amount of its mechanisms could be understood. The ANNs are implemented using simpler and more ordered architectures consisting of neurons (or nodes), which correspond to synaptic connections and layering. Using neural networks, data is analyzed based on the relevance of representations to the problem at hand. The process of learning occurs when neurons in a neural network change their connections between each other as they learn (e.g., as they gain experience). Methods for training neural networks are called learning paradigms.

3.2.2 Deep Learning

Due to the fact that it mimics the human brain in its architecture, deep learning has proven to be the most successful solution in DL. Computer vision algorithms based on CNNs gained popularity in the early 2000s. Developed on the basis of neural networks in the visual cortex, they have excellent results. Each CNN layer has a distinct function, and there are several types. Some of them can indeed be trained, while others only need to implement established functions. CNN architectures commonly use the following layers:

- Convolutional layers. In addition to learning local patterns, convolutional layers can learn spatial hierarchies of patterns as well as translation-invariant patterns. With the increasing depth of the network, CNN is able to learn increasingly complex visual concepts efficiently. In addition to being translation-invariant, convolutional layers can learn spatial hierarchies of patterns, and these two properties allow them to learn local patterns efficiently. Layers that perform convolution over input images produce feature maps to be sent to the following layers as a result of each layer's convolution operation.
- Normalization layers. There is no trainable parameter in the normalization function, and forward propagation is the only available option. In recent years, these layers have become less popular.
- Regularization layer. During training, these layers ignore a portion of neurons randomly in order to reduce overfitting. Dropout is the most common regularization technique.
- Pooling layers. By pooling layers, feature maps can be subsampled while maintaining the main information presented therein, thus reducing model parameters and computation costs. Feature maps are convolved using average pooling and max pooling, like convolutional layers, but no trainable parameters are included.
- Fully connected layers. There is a connection between a neuron and an activation function in another layer. Using a CNN classifier, the classification results are determined after a fully connected layer (FC).

3.2.3 Transfer Learning and Pre-Trained Modelsgoogle's Inception V3

Based on Google's Inception v3 architecture, we were able to categorize 8 diagnostic categories by connecting two fully connected layers, one average pooling layer, and a softmax layer [26]. To make this model compatible with input images, they were all resized to (299, 299). Based on a decay and momentum of 0.9 and a learning rate of 0.0007, the stochastic gradient descent (SGD) algorithm was used to optimize:

- InceptionResNet v2. Using our dataset, the Inception ResNet v2 architecture [27] was retrained by replacing the top layer with a global-average pooling layer, followed by a fully connected layer, followed by a softmax layer, allowing the classification of eight diagnostic categories.
- ResNet 152. Using our dataset, ResNet 152 architecture [28] was retrained by replacing top layers with one average-pooling, one fully connected, and finally, one softmax layer, which allowed the organization of 8 diagnostic categories.
- DenseNet 201. An eight-category classification system was built using DenseNet 201 architecture [29] and a global average pooling layer plus an SL.

3.3 Model Performance Evaluation

The effectiveness of the proposed model was assessed through a comprehensive performance evaluation using a confusion matrix framework. Prior to model development, the dataset was partitioned into distinct training and testing sets to ensure unbiased evaluation. After model training, performance assessment was conducted exclusively on the test set to measure generalization capability.

Following the methodology presented in [30], the skin cancer detection effectiveness of the proposed model was evaluated using four key performance metrics derived from the confusion matrix: accuracy, precision, recall, and F1-score. These metrics collectively provide a comprehensive assessment of the model's classification performance across all test instances.

4 RESULT ANALYSIS AND DISCUSSION

The five machine learning algorithms presented in Figure 1 – LR, LDA, KNN, Classification, and RT, and GNB – are compared based on four metrics: accuracy, precision, recall, and F-score. True positives are more likely to be detected using KNN and CART, particularly in terms of recall and F-score. There is, however, a lower recall for LR and GNB, which indicates that they miss more positive results. Each algorithm for detecting skin cancer is ranked according to its strengths and weaknesses in the chart below.

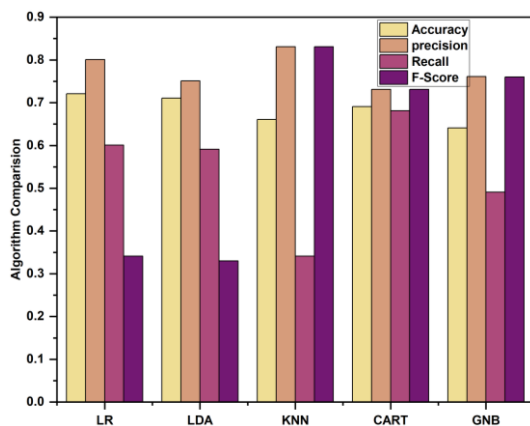


Figure 1: Kaggle-based machine learning models for skin cancer prediction.

In Figure 2, three digital learning models are compared with the frozen base approach to detect

skin cancers (Xception, VGG16, and ResNet50). Xception, ResNet50, and VGG16 achieve the best accuracy and F-score. There is good true positive detection in every model, but Xception performs the best overall in terms of recall, indicating it is the most reliable.

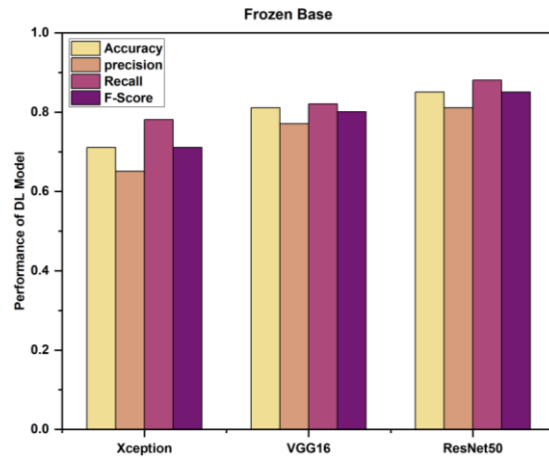


Figure 2: Performance of DL Model versus Frozen Base.

Figure 3 compares the performance of three deep learning models, VGG16 and ResNet50-on skin cancer detection, using a "Fine-Tuning" approach, in which the pre-trained models are more extensively modified to improve results. The Xception model performs better in this scenario compared to the "Frozen Base" scenario, followed closely by ResNet50 and VGG16. There is strong precision and recall across all three models, indicating their ability to detect true positives effectively. As a result of this fine-tuning configuration, Xception remains the most effective model, demonstrating increased performance in all metrics.

In Figure 4, accuracy values for a deep learning model in both frozen and tuned configurations are shown over 25 epochs. It is evident that Tuned Accuracy and Tuned Validation Accuracy are improving steadily, especially after the 10th epoch, surpassing the Frozen Accuracy and Frozen Validation Accuracy, which remain constant. Comparing fine-tuning to using frozen layers highlights the benefits of fine-tuning.

For a frozen deep learning model and a tuned deep learning model, Figure 5 shows loss values over 25 epochs. Losses for Tuned and Tuned Validation are steadily decreasing, reaching lower values than for Frozen and Frozen Validation Losses, which fluctuate more. Using fine-tuned layers is more efficient than using frozen layers for convergence and performance.

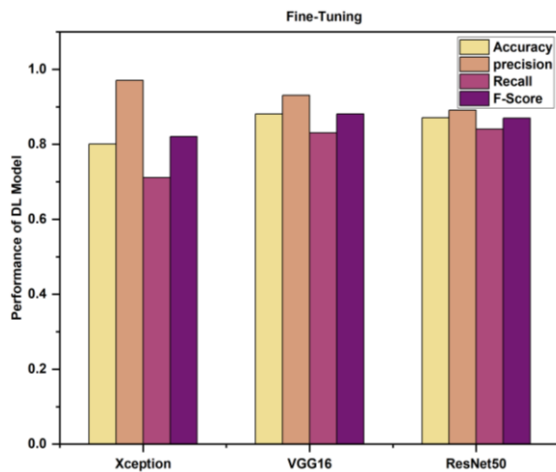


Figure 3: Performance of DL Model versus Fine-Tuning.

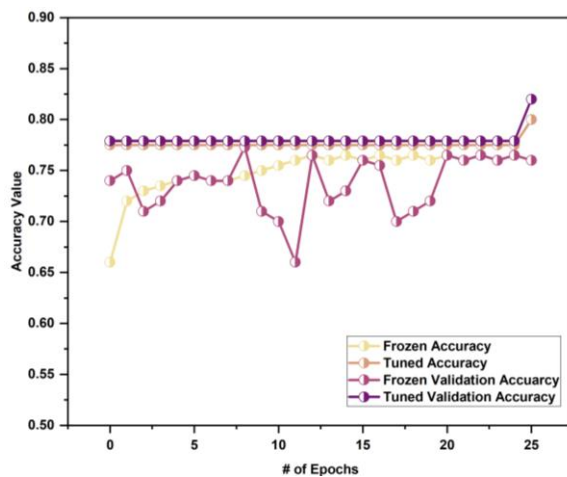


Figure 4: Model accuracy comparison across epochs for frozen and tuned models.

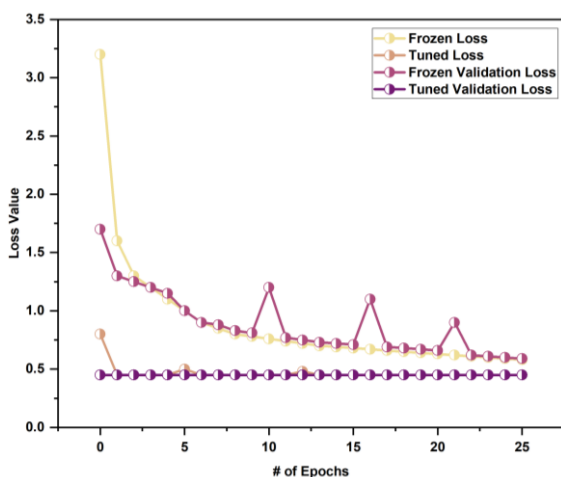


Figure 5: Epoch-by-epoch comparison of frozen and tuned models.

5 CONCLUSIONS

Using machine learning and deep learning, this study demonstrated that skin cancer can be effectively detected in the early stages from dermoscopic images. The comparative analysis clearly showed that pre-trained deep learning models significantly outperformed traditional machine learning approaches, particularly when enhanced through advanced fine-tuning techniques. Among the evaluated models, Xception consistently exhibited superior performance over ResNet50 and VGG16, achieving higher accuracy, precision, and sensitivity across all tested metrics. This highlights that fine-tuning deep learning models results in more precise and dependable skin cancer detection, positioning these tools as highly valuable resources for dermatologists, clinicians, and patients, enabling faster and more accurate diagnoses. Future research could further explore how variations in skin types, pigmentation, demographic factors, and environmental influences impact the reliability and efficacy of these AI models when integrated into real-time clinical diagnostic systems.

6 REFERENCES

- [1] P. Rani, S. Verma, S. P. Yadav, B. K. Rai, M. S. Naruka, and D. Kumar, "Simulation of the lightweight blockchain technique based on privacy and security for healthcare data for the cloud system," *Int. J. E-Health Med. Commun. (IJEHMC)*, vol. 13, no. 4, pp. 1–15, 2022.
- [2] S. Rosenbaum, "Data Governance and Stewardship: Designing Data Stewardship Entities and Advancing Data Access," *Health Serv. Res.*, vol. 45, no. 5p2, pp. 1442–1455, Oct. 2010, doi: 10.1111/j.1475-6773.2010.01140.x.
- [3] E. Proksch, J. M. Brandner, and J. Jensen, "The skin: an indispensable barrier," *Exp. Dermatol.*, vol. 17, no. 12, pp. 1063–1072, Dec. 2008, doi: 10.1111/j.1600-0625.2008.00786.x.
- [4] T. H. Sakuma and H. I. Maibach, "Oily Skin: An Overview," *Skin Pharmacol. Physiol.*, vol. 25, no. 5, pp. 227–235, 2012, doi: 10.1159/000338978.
- [5] P. Rani, S. P. Yadav, P. N. Singh, and M. Almusawi, "Real-World Case Studies: Transforming Mental Healthcare With Natural Language Processing," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Biju, and G. C. Deka, Eds., IGI Global, 2025, pp. 303–324, doi: 10.4018/979-8-3693-4203-9.ch016.
- [6] M. Nasir et al., "Melanoma Detection and Classification using Computerized Analysis of Dermoscopic Systems: A Review," *Curr. Med. Imaging Former. Curr. Med. Imaging Rev.*, vol. 16,

- no. 7, pp. 794–822, Sep. 2020, doi: 10.2174/1573405615666191223122401.
- [7] P. Rani, K. Ur Rehman, S. P. Yadav, and L. Hussein, "Deep Learning and AI in Behavioral Analysis for Revolutionizing Mental Healthcare," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Biju, and G. C. Deka, Eds., IGI Global, 2025, pp. 263–282, doi: 10.4018/979-8-3693-4203-9.ch014.
- [8] B. Zhang et al., "Opportunities and Challenges: Classification of Skin Disease Based on Deep Learning," *Chin. J. Mech. Eng.*, vol. 34, no. 1, p. 112, Dec. 2021, doi: 10.1186/s10033-021-00629-5.
- [9] C. Herman, "Emerging technologies for the detection of melanoma: achieving better outcomes," *Clin. Cosmet. Investig. Dermatol.*, p. 195, Nov. 2012, doi: 10.2147/CCID.S27902.
- [10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [11] G. Ansari, P. Rani, and V. Kumar, "A novel technique of mixed gas identification based on the group method of data handling (GMDH) on time-dependent MOX gas sensor data," in *Proc. Int. Conf. Recent Trends Comput. (ICRTC 2022)*, Springer, 2023, pp. 641–654.
- [12] N. C. F. Codella et al., "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM J. Res. Dev.*, vol. 61, no. 4/5, pp. 5:1–5:15, Jul. 2017, doi: 10.1147/JRD.2017.2708299.
- [13] C. Barata, M. Ruela, M. Francisco, T. Mendonca, and J. S. Marques, "Two Systems for the Detection of Melanomas in Dermoscopy Images Using Texture and Color Features," *IEEE Syst. J.*, vol. 8, no. 3, pp. 965–979, Sep. 2014, doi: 10.1109/JSYST.2013.2271540.
- [14] P. Rani, D. S. Mohan, S. P. Yadav, G. K. Rajput, and M. A. Farouni, "Sentiment Analysis and Emotional Recognition: Enhancing Therapeutic Interventions," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Biju, and G. C. Deka, Eds., IGI Global, 2025, pp. 283–302, doi: 10.4018/979-8-3693-4203-9.ch015.
- [15] J. Daghrir, L. Tlig, M. Bouchouicha, and M. Sayadi, "Melanoma skin cancer detection using deep learning and classical machine learning techniques: A hybrid approach," in *Proc. 5th Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, Sousse, Tunisia: IEEE, Sep. 2020, pp. 1–5, doi: 10.1109/ATSIP49331.2020.9231544.
- [16] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," *Comput. Electr. Eng.*, vol. 105, p. 108543, 2023.
- [17] J. Fernandez Alcon et al., "Automatic Imaging System With Decision Support for Inspection of Pigmented Skin Lesions and Melanoma Diagnosis," *IEEE J. Sel. Top. Signal Process.*, vol. 3, no. 1, pp. 14–25, Feb. 2009, doi: 10.1109/JSTSP.2008.2011156.
- [18] A. Rehman, M. A. Khan, Z. Mehmood, T. Saba, M. Sardaraz, and M. Rashid, "Microscopic melanoma detection and classification: A framework of pixel-based fusion and multilevel features reduction," *Microsc. Res. Tech.*, vol. 83, no. 4, pp. 410–423, Apr. 2020, doi: 10.1002/jemt.23429.
- [19] J. Kawahara, A. BenTaieb, and G. Hamarneh, "Deep features to classify skin lesions," in *Proc. IEEE 13th Int. Symp. Biomed. Imaging (ISBI)*, Prague, Czech Republic: IEEE, Apr. 2016, pp. 1397–1400, doi: 10.1109/ISBI.2016.7493528.
- [20] S. Verma et al., "An automated face mask detection system using transfer learning based neural network to preventing viral infection," *Expert Syst.*, p. e13507, 2024.
- [21] K. Hu et al., "Classification of melanoma based on feature similarity measurement for codebook learning in the bag-of-features model," *Biomed. Signal Process. Control*, vol. 51, pp. 200–209, May 2019, doi: 10.1016/j.bspc.2019.02.018.
- [22] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, 2018.
- [23] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "PH² - A dermoscopic image database for research and benchmarking," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2013, pp. 5437–5440. [Online]. Available: <https://ieeexplore.ieee.org/document/6610779/>.
- [24] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.
- [25] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Systems Design and Implementation (OSDI 16)*, 2016, pp. 265–283. [Online]. Available: <https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi>.
- [26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html.
- [27] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017. [Online]. Available: <https://ojs.aaai.org/index.php/aaai/article/view/11231>.
- [28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.
- [29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.
- [30] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, pp. 37–63, 2011, doi: 10.9735/2229-3981.

Intelligent IoT-Based Data Analytics System for Precision Farming Using Regression Techniques

Anidi Anidi¹, Arna Juwariyah², Ahmed Aljuboori³ and Shahad Lateef Abdulwahid⁴

¹Department of Early Childhood Education, Universitas Sulawesi Tenggara, 93121 Kendari, Indonesia

²Department of Premary Teacher Education, Universitas Sulawesi Tenggara, 93121 Kendari, Indonesia

³Department of Computer Science, Dijlah University College, 10021 Baghdad, Iraq

⁴Faculty of Computer Engineering, Imam Ja'afar Al-Sadiq University, 62001, Maysan, Iraq

anidi@un-sultra.ac.id, arnajuwariyah@un-sultra.ac.id, a.s.aljuboori@duc.edu.iq, Shahadlateef17@gmail.com

Keywords: Precision Farming, Internet of Things (IoT), Regression Techniques, Machine Learning, Smart Agriculture

Abstract: Increasing global demands for sustainable agricultural practices, triggered by food security and climate change, have evolved precision farming. By combining machine learning with the Internet of Things (IoT), precision agriculture can optimize resource use and crop yields. Data analytics based on IoT are used here to enhance agricultural decision-making by using regression techniques like Support Vector Machines (SVMs) and Multilayer Perceptrons (MLPs). Data from environmental and soil sensors are collected in agricultural fields, including temperature, humidity, nitrogen, phosphorus, potassium, pH, and rainfall. Utilizing machine learning algorithms, this data is processed to predict which crops will yield the highest yields and utilize the most resources. Based on a comparative analysis, MLP models exhibit superior performance to SVM models with respect to training time, testing time, and regression error (lower RMSE). It achieves the highest classification accuracy (92%) among existing models such as SPS, CSMS, and SRIM. When farmers have access to real-time, data-driven insights, they can make better decisions, increase productivity, and adopt more sustainable farming practices.

1 INTRODUCTION

The Internet of Things (IoT) has revolutionized many industries, including agriculture, thanks to its rapid growth. The concept of precision farming, which maximizes agricultural practices based on real-time data, has emerged as a powerful way to enhance productivity, resource management, and sustainability. IoT-based data analytics systems play an important role in this sector due to the collection and analysis of sensor and device data on the ground [1]. By making informed decisions, farmers are able to increase yields and reduce costs by irrigating, fertilizing, controlling pests, and harvesting effectively. In agricultural data analysis, regression techniques are particularly valuable. Based on historical and real-time inputs, they help predict key variables such as soil moisture levels and crop growth rates. The integration of IoT devices with intelligent regression models enables highly accurate forecasting systems to be developed, which can adjust to dynamic environmental conditions and optimize farm management. It is designed to provide farmers with a

robust tool to enhance decision-making processes by bridging the gap between data collection and actionable insights. This study aims to harness the application of advanced analytics to agricultural IoT data in order to promote environmental conservation, ensure food security, and develop sustainable agricultural practices as a result of global challenges such as climate change and population growth.

With precision farming, modern agriculture is becoming more productive, efficient, and sustainable. With the growing demand for food and the challenges posed by climate change, agriculture practices must be optimized. Many benefits can be achieved by analyzing data using IoT and machine learning. Real-time monitoring of environmental conditions, soil quality, crop health, and other critical parameters is made possible by the Internet of Things. To provide actionable insights, these devices require massive amounts of data to be processed intelligently. A regression algorithm is a powerful tool for analyzing and predicting agricultural outcomes, including yield estimates, irrigation requirements, and pest management measures. Agrarian practices have

become outdated due to a shortage of supplies, incompetence with technology, and a lack of knowledge among farmers [2], [3]. It is also common for pests and insects to reduce the yield of certain crops. Pests and insects have attacked several crops. Animals and birds can be poisoned by some insects and pesticides, making them ineffective. Additionally, it damages animal and food chains' natural food webs [4]. Due to crop diseases, there is a significant reduction in throughput. Approximately 40% of agricultural yield loss occurs due to insects, pests, viruses, animals, and weeds, according to [5]. Pests, diseases, animals, and weeds are responsible for 40% of agricultural yield losses, according to [5]. Moreover, they have both short- and long-term effects, some of which are temporary and others are permanent [6]. A lot depends on the weather when it comes to agriculture. The weather has a significant impact on agriculture.

We live in a world where technology is everywhere. Currently, remote monitoring techniques are used to provide relevant information to farmers [7]. Several factors contribute to this, including WiFi sensor networks and the Internet of Things [1]. Miniaturization led to the development of the Internet of Things [8]. During his 1999 lecture on Supply Chain Management, Kevin Ashton mentioned the Internet of Things for the first time [9]. In the Internet of Things, a sensor can uniquely identify a smart object. Any object, sensor, person, or smart device that connects and shares information is considered a "thing" [10]. The traditional method of detecting diseases and pests by hand and calculating quantity and production based on statistics resulted in human error in the past [11], [12]. The technology learns from its experiences through machine learning. By analyzing and modelling large datasets collected from crop fields, valuable insights can be gained. This technique identifies hidden patterns in horticulture, including salt content, temperature, and humidity. Several machine-learning approaches can be used for crop disease prediction, such as artificial neural networks, SVM regressions, logistic regressions, fuzzy technologies, etc [13]. [14]. The use of machine learning to predict apple diseases has been developed by scientists. Additional information about coronaries can be derived from apple leaf images in addition to apple scab images. Four algorithms were used to classify the same dataset: Support Vector Machine, K Nearest Neighbor, Decision Tree, and Naive Bayes [13]. Matlab 2016 was used for simulations. According to this study, KNN categorizes diseases with 99.4% accuracy. Himachal Pradesh and Uttarakhand developed it as an alternative to existing

systems that were unreliable and expensive. Predicting crop diseases using IoT and machine learning was discussed during the presentation [15]. The system's model was developed by combining IoT and machine learning. Several environmental sensors collected data, including a temperature sensor and a humidity sensor.

2 LITERATURE REVIEW

In the soil, plants can find the nutrients they need to thrive. Minerals in the soil are necessary for growth, but if any are missing, the plant will have trouble growing. Soil composition must be regularly tested to ensure that plants receive enough nutrients. An application of fertilizer high in nutrients may correct nutrient deficiency problems in the soil. Fertilizers have positively influenced agricultural output, but their extensive use has caused ecological damage. The importance of soil nutrient testing in agriculture can be attributed to this fact. Even though conventional soil testing provides accurate data, it is unsuitable for precision agriculture because of the cost and time involved in obtaining the results. Tests that test a larger number of samples are prohibitively expensive, so they cannot measure a field's geographic heterogeneity. Therefore, fast, portable, economical, and highly precise methods are essential for achieving the best results [16].

An expanding population and careful management are essential for agriculture. According to the authors [17], the modern agricultural system relies heavily on human labour but is highly mechanized. The return on a thirty per cent investment from 1920 to 1970 was one hundred eighty per cent. A significant increase in productivity was not the result of a rise in data sources but rather of advances in farming techniques. It has been found that sifting machines, mechanical innovations, and synthetic manures contribute to agricultural profitability. Technology has become more important to farmers over the last decade for communicating and storing information. They can, therefore, better monitor their financial data and interactions with third parties. We live in a world where information is readily available. As a result, farmers can easily collect data and conduct statistical analyses using field observations in horticulture.

According to the authors [18], several proprietary techniques can be used to improve agricultural monitoring. When it comes to keeping track of geographical regions and climatic punctuations, researchers have found more complex frameworks. Over time, Farm Management Information Systems

(FMIS) have evolved to address the specific needs and activities of farms. In the present day, these frameworks are integrating into the Internet age by utilizing established systems management and responses to strengthen agricultural structures. As time has progressed, Farm Management Information Systems, or FMIS, have evolved to meet the specific needs of farms. However, many people think the Internet is not perfect, particularly when it comes to managing a large number of connected devices, such as IoT devices or stakeholder devices. There is, however, no standardized solution that can ensure reliable interoperability between relevant authorities. FI has been providing frameworks to help close these gaps since that time.

In recent years, farmers have been urged to adopt thorough management practices, and a sensor-based approach could finally allow them to do so. It has been demonstrated that sensors play an important role in agriculture, and their fundamental role has been defined. To measure farming efficiency, the Precision Agriculture Monitor System (PAMS) uses Shining sensors. A monitoring and regulation system such as the IFarm Framework is recommended to increase agricultural production by reinforcing socioeconomic elements. Monitoring and regulating water usage may be easier with this system. Using a variety of characteristics, [19], [20] classified sensor technologies according to their performance.

The media, government agencies, and farm-based equipment are the main sources of precision agricultural information in Texas and New Mexico, two states that grow cotton [21]. Precision agriculture technology can be influenced in various ways by data from a variety of sources. GPS-enabled yield monitors and soil survey maps are popular examples. Dealers contribute to the implementation of zone soil sampling and soil survey maps in varying degrees. Smart farming relies heavily on ML and the IoT. When implementing these practices, farmers face several challenges, including predicting crop diseases. Several diseases affect apple crops, but apple scab is the most common. Using WSNs [22], [23] in apple orchards allows real-time data collection and early disease prediction. Additionally, he discussed the challenges farmers face when handling hardware units and sensors due to environmental factors [24]. As part of precision farming, automated devices, Internet of Things sensors, real-time data collection, and cloud storage are used, as well as data analysis. Irrigation systems and greenhouses can be controlled smartly using a framework proposed by one author. Nutritional, climate, and irrigation data can be stored, managed, and analyzed with it [25]. Because the soil

nutrition level decreases from year to year due to cultivation, this method maximizes soil fertility [26]. Combined with IoT sensors and smart tools, big data plays an increasingly important role. An article addressed the volume of sensor data generated, cloud storage availability, and challenges associated with real-time data analysis and visualization [27].

It was proposed that WSNs and IoT be used for precision agriculture and irrigation monitoring. The implementation and maintenance costs of fully sensor-based agriculture, as well as farmers' lack of knowledge, have been noted as limitations. A farmer could, therefore, make an appropriate decision based on updated information. According to one author, soil samples could be used to predict the type of crop that would be suitable for a particular field. ESP8266 WiFi module, Arduino board, and other sensors were used to collect soil temperatures, moisture, and mineral values. The algorithms that performed most accurately on the rainfall dataset were naive Bayes, logistic, and C 4.5 [28].

3 METHODOLOGY

Agricultural producers have benefited from the Internet of Things by improving irrigation efficiency, increasing yields, and reducing costs. Through the integration of agriculture and information technology, an intelligent agricultural solution can be created. With the advent of IoT technology, three aspects have emerged.

- For years, WSNs have been driving the growth of precision agriculture by enhancing crop yields through the use of advanced technology.
- Crops are currently produced using farming facilities that yield high-quality yields. Inputs are high, outputs are high, capital is high, and labour is high in the production process.
- Producing and managing crops on a contract basis is a relatively new concept in agriculture. Globally, urbanization has outpaced rural development, creating a growing gap between the two. According to statistics, 80% of extremely poor people live in rural areas, and 75% of moderately poor people live in rural areas.

3.1 IoT Framework for Agriculture

Cloud database management systems will access real-world data to train machine learning models using real-world data, which is accessed through storage

media. Figure 2 illustrates one of 22 crops that could be implemented as a result of the module.

3.1.1 Data Mining and Network Implementation

A first-level architecture enables utilities to capture and communicate data. Gateways and base stations are connected to the sensor network. The second level incorporates classification algorithms and specifications. Next, the machine learning algorithms must be implemented to acquire the server's results. Irrigation crops are obtained from the server using a trained module. Specific sensors measure temperature, humidity, and rainfall, whereas nitrogen, phosphorus, potassium, and pH parameters are measured by analytical sensors. A spreadsheet was used to compile the data, with 22 different crops as the ground truth. Using machine learning algorithms, the module is trained to recommend crops for irrigation based on the model attained. In the article, it demonstrates how IoT-based smart agriculture can be integrated into the field by combining physical structure, data collection, data processing, and data analytics. From Kaggle, we acquired the datasheet [29]. The crop is irrigated as recommended to predict crop yields and obtain maximum yields.

A variety of environmental factors determine how well crops will be irrigated, such as soil fertility, i.e., nitrogen, phosphorous, and moisture. By considering the 7 attributes, a crop can be irrigated to maximize yield. Based on this article, it can be used to make better decisions about planting different crops. With WEKA or Waikato Environment for Knowledge Analysis, a free, open-source software program licensed under the GNU public license, you can

perform a wide range of fact-mining tasks as shown in Figure 1.

3.2 Agriculture Sensor Layer

Mobile devices (smartphones, sensor nodes, etc.) use GPS to create various IoT devices for smart agriculture, such as field sensors, greenhouse sensors, photovoltaic farms, solar insecticide lamps, etc. Consequently, IoT devices can be integrated and adapted to serve two purposes in agriculture. Our first commitment is to provide nutrient solutions in a trustworthy manner, as well as distributing them in a timely manner. As well as improving consumption management, lowering costs, and reducing losses, we strive for a better solution. In addition to being beneficial for the economy, it will also be beneficial for the environment. SCADA systems (supervisory control and data acquisition) are used in agriculture to control operations. IoT sensors and meters are proposed for greenhouses:

IoT sensors and meters are proposed for greenhouses:

- Calculate the water pumping system pressure, drip rate, and surface to be irrigated using IoT devices.
- Water meters that display real-time information display the status of water storage.
- Systems that use IoT to filter water based on physical properties (e.g., sand filters).
- Fertiliser meters should be able to provide real-time updates for injections and storage tanks of NPK fertilisers.
- It is possible to control the pH and electrical conductivity of nutrient solutions through the Internet of Things.

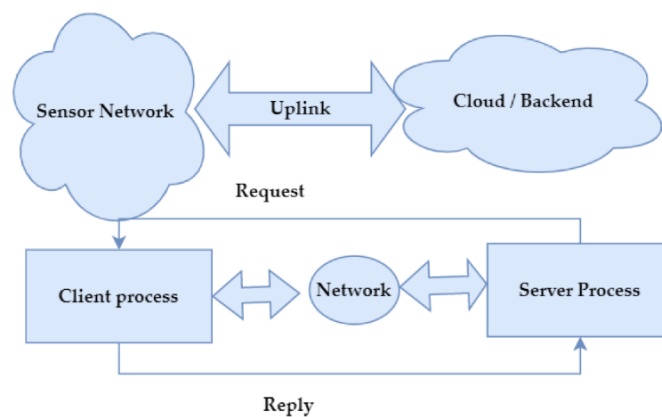


Figure 1: A proposed IoT client-server architecture.

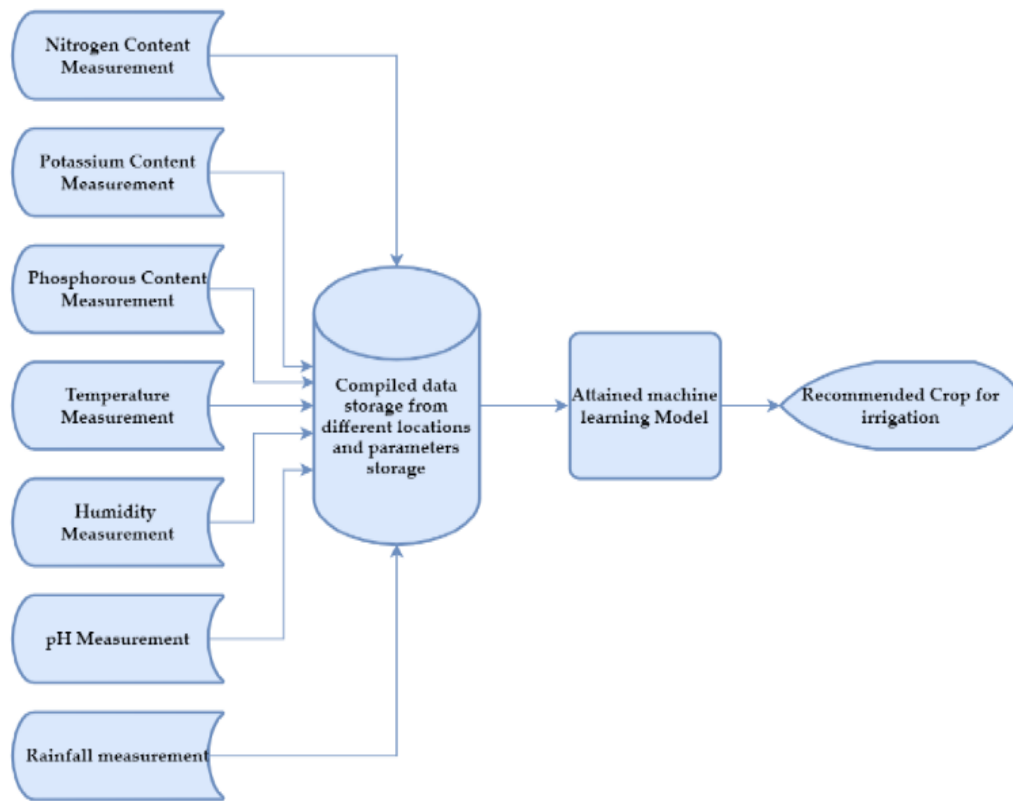


Figure 2: Using machine learning modules to construct a block diagram.

3.3 Fog Computing Layer

Analyzing and making informed decisions is enabled by the analytics and decision-making model, which provides farmers with reports that accompany their decisions. Several states, as well as machine intelligence, are predicted to be created as a result of the in-charge order. On the fog layer of data analytics systems, machine learning algorithms and classification models are typically used.

Sensors initially generate data, which is then acquired by edge devices. A preprocessing step cleans and corrects stored data. Data designed for a particular purpose is used to train machine learning algorithms, which are then initialized to reflect that purpose. A sensor collects data in the Internet of Things, which may be collected in real-time or in batches (for instance, when to pump water) (for example, temperature and humidity). A multilayer perceptron neural network (MLP-ANN) makes predictions and displays data in graphs using support vector machines. In addition to its ability to learn nonlinear models in high-dimensional spaces, this algorithm offers several advantages.

3.3.1 Multilayer Perceptron

This type of NN can also supplement feedforward neural networks. There are three layers in the MLP: input, hidden, and output. As opposed to linear functions, MLP approximates continuous functions. Perceptions, or neurons, are part of the MLP. In this case, $(x = x_1 + x_2 + x_4 + x_5 \dots \dots x_n)$ receives n features [30] as input $(x = x_1 + x_2 + x_4 + x_5 \dots \dots x_n)$.

Weighted sums are calculated for input layers based on the n features sent to u .

$$u_x = \sum_{i=1}^n w_i x_i. \quad (1)$$

A result of this type should be passed on to the activation function [f]. In this article, sigmoid nodes 0 to 35 are supposed to be passed on. A hidden layer or layers can be found in MLP. There is, however, an external layer between the input and output layers.

As far as support vector machines (SVMs) are concerned, they are divided into two categories: classification and regression. The goal of SVM is to

shift nonlinear data into a linear space where it can be separated. Two conditions must be met for the hyperplane to effectively separate data: the distance between vectors and the hyperplane must be adapted to differing aspects of the vectors. Here is what the assumption function looks like:

$$f(x_i) = \begin{cases} +1 & \text{if } w \cdot x + b \geq 0 \\ -1 & \text{if } w \cdot x + b < 0 \end{cases}. \quad (2)$$

Class +1 place points cannot be found above, below, or on the hyperplane, while class-1 place points can be found above, below, and on the hyperplane.

Perceptrons with multiple layers in an artificial neural network (ANN) are called multilayer perceptrons. The animate nervous system is controlled by perceptrons, which are interconnected systems. A deep learning neural network simulates nonlinear functions of high order as its foundation. Here are the steps for calculating the degree of accuracy in output node j based on the above example:

$$e_j(n) = d_j(n) - y_j(n). \quad (3)$$

Perceptrons produce output values based on goal values. Weight adjustments can reduce the output error by adjusting the nodes' weights.

$$e(n) = \frac{1}{2} \sum_j e_j^2(n). \quad (4)$$

3.4 Edge Network Layer

Sensors are detected and connected to low-power microcontrollers at remote locations using IoT microcontrollers designed for IoT. Sensor data can be collected, analyzed, and transmitted to the edge layer's base station using the ESP 32 Node MCU. The calibration and comparison of sensors are required for analogue and digital data collection. A healthy and unhealthy climate is collected in order to ensure crop survival.

A climate-based model tracks Gerbera and Broccoli. The parameters of greenhouses are monitored using Node MCU ESP 32 microcontrollers. The data collected by personal computers is serialized with timestamps. A DHT11 sensor measures temperature and humidity, an LDR sensor measures light intensity, an MQ2 sensor measures CO₂, and a Cu lead measures soil moisture. Ten days are monitored continuously using the MQTT protocol by Adafruit IO over the cloud platform using a specific time interval. Computers monitor sensor data in real time using cloud-based controllers.

3.5 Cloud Computing Layer

Through an application that uses data from edge layer nodes to process and control at the base station, farmers can monitor crop cultivation progress. The Adafruit IO platform displays, acts on, and interacts with sensor data. Additionally, MQTT ensures that the dossier remains private and secure. The MQTT protocol is an inconsequential problem-solving protocol that is located on TCP/IP. A message broker routes messages through MQTT, connecting senders and receivers that send and receive messages. Publishing and subscribing to messages can be done with the same client. Data is sent from temperature sensors based on a specific subject, such as in a greenhouse system.

4 RESULTS AND DISCUSSION

In Figure 3, the training and testing times for SVM and MLP models are compared. The MLP takes about 450 seconds to train (about 320 seconds), while the SVM takes about 70 to 80 seconds to test. During training, MLP uses more computation than SVM, but during testing, it is similar to SVM.

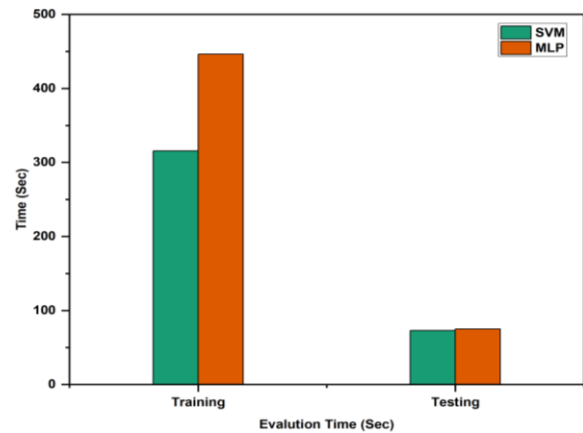


Figure 3: Representation of evaluation time.

A comparison of the SVM and MLP models is presented in Figure 4. Pump and Fan data show that SVM is more accurate than MLP in the Pump and Fan category. Compared to SVM, MLP performs significantly better in the Lift category. Pump and fan performance are better with SVM, whereas lift performance is better with MLP, as shown in Figure 4.

This bar chart compares pumps, fans, and lifts using SVM and MLP models. Pump and Fan performance is slightly better with SVM, while Lift performance is significantly better with MLP. Pump and fan performance are better with SVM, whereas lift performance is better with MLP.

A bar chart representing the average RMSE (Root Mean Square Error) for regression tasks for SVMs and MLPs is shown in Figure 6. MLP is represented by the blue bar, while SVM is represented by the yellow bar. Based on the chart, MLP is better at regression with lower prediction error than SVM (around 0.08), which indicates better regression performance with MLP.

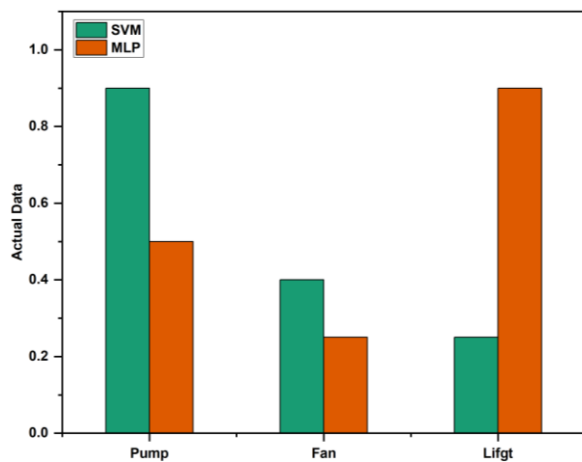


Figure 4: Comparison of SVM and MLP performance across equipment categories.

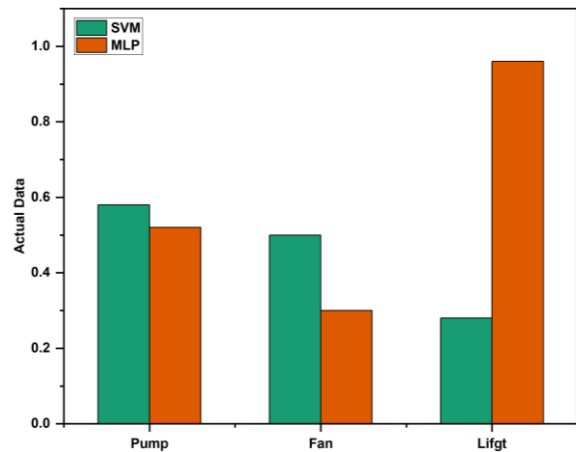


Figure 5: Performance Comparison of SVM and MLP across different equipment types.

In Figure 7, four models are compared, each of which has an accuracy of 92%: the Proposed Model (92%), the SPS (90%), the CSMS (66%), and the SRIM (89%). According to the chart, SRIM and SPS

achieve a higher degree of accuracy than the Proposed Model, while CSMS achieves the lowest degree of accuracy. Classification accuracy is better with the Proposed Model than with the other models.

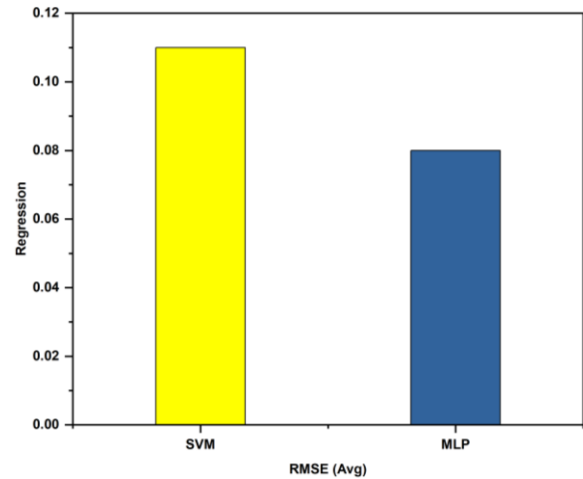


Figure 6: Comparison of average RMSE between SVM and MLP models.

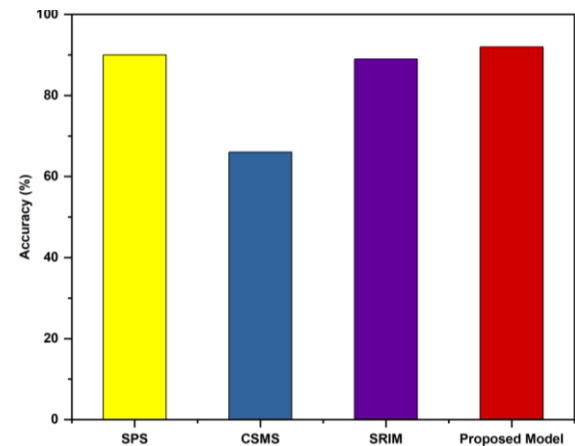


Figure 7: Accuracy comparison of SPS, CSMS, SRIM, and proposed model.

5 CONCLUSIONS

Precision agriculture was the focus of this paper, which presented a data analytics system based on IoT that is intelligent and IoT-based. Real-time data collection, analysis, and actionable insights are delivered through the use of IoT sensors and machine-learning regression models. Training and testing experiments have consistently shown that Multi-Layer Perceptrons (MLPs) outperform Support Vector Machines (SVMs) in terms of Root Mean Square

Error (RMSE) and accuracy metrics. A comprehensive dataset obtained from Kaggle, comprising detailed soil fertility indicators and varied climatic conditions, was employed to rigorously test and validate the effectiveness of the proposed system. The model accurately recommended suitable crops tailored to specific environmental conditions, demonstrating practical applicability. Furthermore, the proposed approach achieved the highest accuracy of 92% when compared with other advanced frameworks such as SPS, CSMS, and SRIM. These results validate that intelligent IoT-based agricultural systems are not only feasible but essential for achieving long-term sustainability and food security. By providing farmers with actionable predictive insights, this technology reduces reliance on traditional, less-efficient agricultural practices, optimizes resource allocation, minimizes environmental impact, and significantly enhances the resilience, productivity, and overall efficiency of modern farming systems.

REFERENCES

- [1] B. Bhola et al., "Quality-enabled decentralized dynamic IoT platform with scalable resources integration," *IET Communications*, p. cmu2.12514, Oct. 2022, doi: 10.1049/cmu2.12514.
- [2] K. Bhargava, A. Kashyap, and T. A. Gonsalves, "Wireless sensor network based advisory system for Apple Scab prevention," in *Proc. 2014 Twentieth National Conference on Communications (NCC)*, IEEE, 2014, pp. 1–6. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/6811263/>.
- [3] A. Singh et al., "Smart Traffic Monitoring Through Real-Time Moving Vehicle Detection Using Deep Learning via Aerial Images for Consumer Application," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 4, pp. 7302–7309, Nov. 2024, doi: 10.1109/TCE.2024.3445728.
- [4] S. S. Shinde and M. Kulkarni, "Review Paper on Prediction of Crop Disease Using IoT and Machine Learning," in *Proc. 2017 Int. Conf. Transforming Engineering Education (ICTEE)*, Pune: IEEE, Dec. 2017, pp. 1–4, doi: 10.1109/ICTEED.2017.8586207.
- [5] E.-C. Oerke, "Crop losses to pests," *Journal of Agricultural Science*, vol. 144, no. 1, pp. 31–43, 2006.
- [6] S. Savary, A. Ficke, J.-N. Aubertot, and C. Hollier, "Crop losses due to diseases and their implications for global food production losses and food security," *Food Security*, vol. 4, no. 4, pp. 519–537, Dec. 2012, doi: 10.1007/s12571-012-0200-5.
- [7] P. Rani, S. P. Yadav, P. N. Singh, and M. Almusawi, "Real-World Case Studies: Transforming Mental Healthcare With Natural Language Processing," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Biju, and G. C. Deka, Eds., IGI Global, 2025, pp. 303–324, doi: 10.4018/979-8-3693-4203-9.ch016.
- [8] R. Buyya and A. V. Dastjerdi, *Internet of Things: Principles and Paradigms*, Elsevier, 2016. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=_k11CwAAQBAJ.
- [9] K. Ashton, "That 'internet of things' thing," *RFID Journal*, vol. 22, no. 7, pp. 97–114, 2009.
- [10] P. Rani et al., "An Implementation of Modified Blowfish Technique with Honey Bee Behavior Optimization for Load Balancing in Cloud System Environment," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–14, May 2022, doi: 10.1155/2022/3365392.
- [11] T. Rumpf et al., "Early detection and classification of plant diseases with support vector machines based on hyperspectral reflectance," *Computers and Electronics in Agriculture*, vol. 74, no. 1, pp. 91–99, 2010.
- [12] P. Rani et al., "Simulation of the lightweight blockchain technique based on privacy and security for healthcare data for the cloud system," *International Journal of E-Health and Medical Communications (IJEHMC)*, vol. 13, no. 4, pp. 1–15, 2022.
- [13] P. Rani and R. Sharma, "Intelligent transportation system for internet of vehicles based vehicular networks for smart cities," *Computers & Electrical Engineering*, vol. 105, p. 108543, 2023.
- [14] S. Singh and S. Gupta, "Digital image processing techniques for early detection and classification of different diseased plants," *International Journal of Bio-Science and Bio-Technology*, vol. 8, no. 4, pp. 61–66, 2016.
- [15] G. Ansari, P. Rani, and V. Kumar, "A Novel Technique of Mixed Gas Identification Based on the Group Method of Data Handling (GMDH) on Time-Dependent MOX Gas Sensor Data," in *Proc. Int. Conf. Recent Trends in Computing*, vol. 600, Springer Nature Singapore, 2023, pp. 641–654, doi: 10.1007/978-981-19-8825-7_55.
- [16] P. Rani et al., "Deep Learning and AI in Behavioral Analysis for Revolutionizing Mental Healthcare," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra et al., Eds., IGI Global, 2025, pp. 263–282, doi: 10.4018/979-8-3693-4203-9.ch014.
- [17] A. L. Olmstead and P. W. Rhode, "Conceptual issues for the comparative study of agricultural development," in *Agriculture and Economic Development in Europe since 1870*, Routledge, 2008, pp. 47–71. [Online]. Available: <https://www.taylorfrancis.com/chapters/edit/10.4324/9780203928776-9>.

- [18] M. A. Alqahtani, "Machine learning techniques for malware detection with challenges and future directions," *International Journal of Communication Networks and Information Security*, vol. 13, no. 2, pp. 258–270, 2021.
- [19] M. H. Anisi, G. Abdul-Salaam, and A. H. Abdullah, "A survey of wireless sensor network approaches and their energy consumption for monitoring farm fields in precision agriculture," *Precision Agriculture*, vol. 16, no. 2, pp. 216–238, Apr. 2015, doi: 10.1007/s11119-014-9371-8.
- [20] P. Rani and M. H. Falaah, "Real-Time Congestion Control and Load Optimization in Cloud-MANETs Using Predictive Algorithms," *NJF Intelligent Engineering Journal*, vol. 1, no. 1, pp. 66–76, 2024.
- [21] M. Csótó, "Information flow in agriculture—through new channels for improved effectiveness," *Journal of Agricultural Informatics*, vol. 1, no. 2, 2010. [Online]. Available: <http://magisz.org/journal/index.php/jai/article/view/17>
- [22] N. Kumar et al., "Teeech: Three-tier extended energy efficient clustering hierarchy protocol for heterogeneous wireless sensor network," *Expert Systems with Applications*, vol. 216, p. 119448, 2023.
- [23] N. Kumar et al., "THWSN: Enhanced energy-efficient clustering approach for three-tier heterogeneous wireless sensor networks," *IEEE Sensors Journal*, vol. 22, no. 20, pp. 20053–20062, 2022.
- [24] R. Akhter and S. A. Sofi, "Precision agriculture using IoT data analytics and machine learning," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 5602–5618, Sep. 2022, doi: 10.1016/j.jksuci.2021.05.013.
- [25] M. A. Zamora-Izquierdo et al., "Smart farming IoT platform based on edge and cloud computing," *Biosystems Engineering*, vol. 177, pp. 4–17, Jan. 2019, doi: 10.1016/j.biosystemseng.2018.10.014.
- [26] U. Ahmed et al., "A nutrient recommendation system for soil fertilization based on evolutionary computation," *Computers and Electronics in Agriculture*, vol. 189, p. 106407, Oct. 2021, doi: 10.1016/j.compag.2021.106407.
- [27] E. Ahmed et al., "The role of big data analytics in Internet of Things," *Computer Networks*, vol. 129, pp. 459–471, Dec. 2017, doi: 10.1016/j.comnet.2017.06.013.
- [28] S. Prabhu et al., "Soil analysis and crop prediction," *International Journal of Scientific Research in Science and Technology*, vol. 7, no. 4, pp. 117–123, 2020.
- [29] Crop Recommendation Dataset. [Online]. Available: <https://www.kaggle.com/atharvaingle/crop-recommendation-dataset>.
- [30] T. Vijayan, M. Sangeetha, A. Kumaravel, and B. Karthik, "Feature Selection for Simple Color Histogram Filter based on Retinal Fundus Images for Diabetic Retinopathy Recognition," *IETE Journal of Research*, vol. 69, no. 2, pp. 987–994, Feb. 2023, doi: 10.1080/03772063.2020.1844082.

AI-Enhanced Intelligent Healthcare: Advancements in Remote Monitoring, Predictive Analytics and Disease Diagnosis

Hussein Qahtan Al Gburi^{1,2}, Abeer Tariq Maolood¹ and Akbas Ezaldeen Ali¹

¹Collage of Computer Science, University of Technology, 10066 Baghdad, Iraq

²Administrative Affairs, Ministry of Education, 100066 Baghdad, Iraq

Hussein.kahtan@gmail.com, abeer.t.maolood@uotechnology.edu.iq, akbas.e.ali@uotechnology.edu.iq

Keywords: Artificial Intelligence, Deep Learning, CNN, Machine Learning, Internet of Things, Healthcare.

Abstract: Artificial Intelligence (AI) and Deep Learning (DL) in healthcare have brought advanced health monitoring and predictive analytics. By amalgamating AI with the IoT, Remote Healthcare Monitoring (RHM) systems have been generated, allowing the constant monitoring of patients' health statuses along with decreasing healthcare expenditures. In this paper, we analyze AI-based healthcare innovations exerting their impact in modern medical practices and the need for more advancement in the healthcare predictive solutions from the perspective of AI and the role AI plays in healthcare, that is, the applications of AI in healthcare like disease diagnosis, biomedical research, and predictive analytics. Weather Clinical Decision Support Systems (CDSS) use AI to make specific therapy suggestions based on patient's characteristics and facilitate the best outcomes. Additionally, AI smart city healthcare systems provide effective solutions for remote applications and in improving smart city urban healthcare services. There have been recent advancements in deep learning being used to accurately identify deep learning through Convolutional Neural Networks (CNNs) and predictive modeling of chronic diseases. While AI increases healthcare efficiency and precision to a great extent, existing issues such as data privacy, model interpretability, and system integration still exist. Further studies should focus on further improving natural language processing (NLP) and the adaptability of the AI models along with looking into ethical issues before the widespread usage of AI can be allowed.

1 INTRODUCTION

AI stands for the machine-based cognitive capability different from mental functions of both humans and other biological systems. AI investigates "intelligent agents" which refer to any entity or device that renders perception and understanding of its surroundings and then takes appropriate actions to boost its goal achievement possibilities [1]. The term AI encompasses instances of machine operations that replicate cognitive functions including learning analysis to solve problems. The same intelligence functions under the name machine learning (ML) [2]. The system basis of AI combines hardware elements together with software components into one integrated whole. AI technology uses algorithms as its primary element from a software viewpoint. The theoretical model of artificial neural networks operates as a method for deploying AI algorithms. The system performs as an artificial brain simulation which uses connecting neurons across multiple networks with adjustable message frequencies. A

single neuron detects several stimuli from its neighboring neurons as the entire network changes states according to environmental input changes. The neural network produces outputs when environmental stimuli interact due to its structural similarity with human brain responses. Functional neural networks operate with multiple interconnected layers using different network configurations [3].

Researchers have developed neural networks capable of supervised learning, which involves inferring a function that correlates inputs to outputs based on example pairs; unsupervised learning, which entails extracting insights from unlabeled, unclassified, or uncategorized test data to discern common features, responding to the presence or absence of these features in new data rather than relying on system feedback [4] and reinforcement learning, which focuses on operating within a given environment to optimize rewards and minimize penalties, adhering to an accumulative framework [5]. The advancement of computational power makes neural networks deeper which enables

them to mimic human brain operations to learn better, Figure 1 shows ML Types.

To enable AI deployment on edge devices such as wireless network phones and IoT sensor nodes the technology requires power consumption optimization together with physical simplification [6]. Scientists have tested the deployment of AI algorithms through memristors [7], analog united circuits [8] and spintronics [9]. Memristor crossbar circuits represent one innovative platform that eliminates the requirement to access the "memory wall" because they combine memory operations directly on the computation device [10]. To change these settings, one needs direct access to internet networks. The process of maximizing AI implementation efficiency became the focus of contemporary academic research through work on minimizing data representation bit count. The accuracy of calculations remains stable when data storage bits decrease from 32 or 16 to just 8. The execution speed with lower power requirements along with reduced size benefits these systems [11].

The performance of artificial neural networks remains satisfactory when suitable training methodologies are combined with balanced datasets and adequate data volume and continuous dataset accessibility following application of pre-training techniques or deep training instead of superficial training [12]. The AI software and hardware advancement speed allows implementation across IoT, autopilot driving, Robotic Technical Fields, Image recognition and NLP [13],[14],[15],[16]. The medical field benefits from enhanced analysis tools and treatment results that result from AI research conducted by biomedical researchers who work to improve healthcare industry effectiveness [16].

2 RELATED WORKS

Deep learning and the IoT have combined to transform several industries in recent years by facilitating automated decision-making, real-time monitoring, and intelligent data processing. Applications of this convergence include industrial automation, environmental sensing, infrastructure management, and health monitoring. IoT systems are better equipped to decipher intricate data patterns, increase forecast accuracy, and respond adaptively thanks to deep learning algorithms. To comprehend the promise and difficulties in this quickly developing

field, this review examines recent developments in deep learning-based IoT systems, highlighting significant contributions, approaches, and limits of recent research as shown in Table 1.

3 AI FOR LIVING ASSISTANCE

Artificial intelligence through advanced robotic systems develops better quality life options within supported living spaces for elderly and disabled populations. The ambient intelligent system called RUDO helps people with visual impairments to live with sighted people in both informatics and electronics domains [23]. Through one interface people who have visual impairments can access various features of this intelligent system. During key maternity periods a smart artificial intelligence assistant will offer dietary suggestions together with necessary guidance to expectant mothers. Such personal assistance provides advanced recommendations through its cognitive capabilities which combine "cloud-based communication media among all relevant parties" [24]. A fall-detection system utilizes the integration of a Bayesian classifier and radar Doppler time-frequency signatures that helps protect elderly people from accidents and their associated challenges [25].

The ambient assisted living (AAL) was developed to let AI analyze data from multiple communication systems and recognize network-related incidents while determining assistance needs of senior citizens [26]. Smart houses under "ambient intelligence" perform activity surveillance to provide support services for elderly people which allows AAL settings to enable "aging in place" at home. The analysis of motions using ML techniques monitors risky actions so that preventive actions can be started [27]. The model for AAL is depicted in Figure 2.

Sensors in this setup collect information about both environment conditions and human activity before cloud computing or edge intelligence performs evaluations. The computer system determines necessary steps that activate warning systems together with preventative procedures. AI experts unite with mobile technology and personal digital assistants (PDAs) to provide memory support for people with permanent memory problems allowing them to maintain independent personal tasks [28].

Table 1: Reviewing the latest developments in deep learning-based IoT systems.

Authors	Key Findings	Description	Limitations
Naseri et al. 2023 [17]	Improved classification accuracy using adaptive sailfish moth flame optimization.	Created an IoT-based face mask detection model using hybrid DL.	May struggle with varying lighting conditions and occlusions.
Hamza et al. 2023 [18]	Achieved superior predictive performance in sleep quality monitoring.	Employs a Deep Belief Network for Predicting Sleep Quality.	Limited to sleep monitoring; requires further validation on other health conditions.
Zhu et al. 2023 [19]	Improved early fault detection and predictive maintenance.	Developed a hybrid CNN-LSTM model with KPCA for wind turbine fault prediction.	Limited to wind turbine applications; generalization to other domains needed.
Moradi et al. 2023 [20]	Achieved 81.77% accuracy on prognostic criteria.	Developed a semi-supervised DNN for Health Indicator construction using SHM data fusion.	Requires more real-world testing; focused on composite panel fatigue stress only.
Helmi et al. 2023 [21]	High classification performance using Marine Predator Algorithm for feature selection.	Designed an HAR system with residual convolutional networks and BiGRU for wearable sensors.	Computational cost may be high for real-time applications.
Kondaka et al. 2023 [22]	Enhanced patient monitoring and predictive analytics with cloud-based AI.	Proposed iCloud-Assisted Intensive Deep Learning for real-time smart medical monitoring.	Cloud dependency may introduce latency and security concerns.

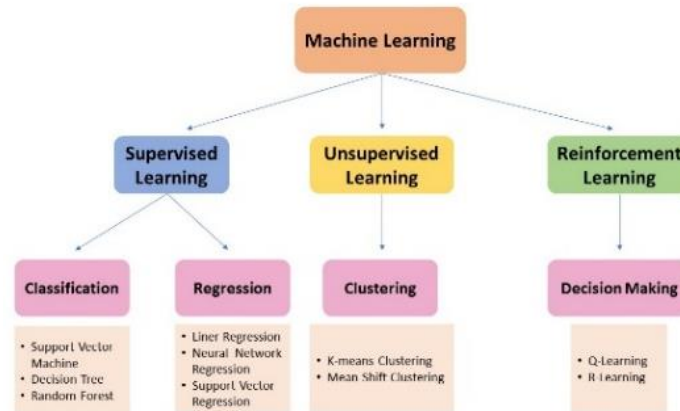


Figure 1: Types of AI learning.

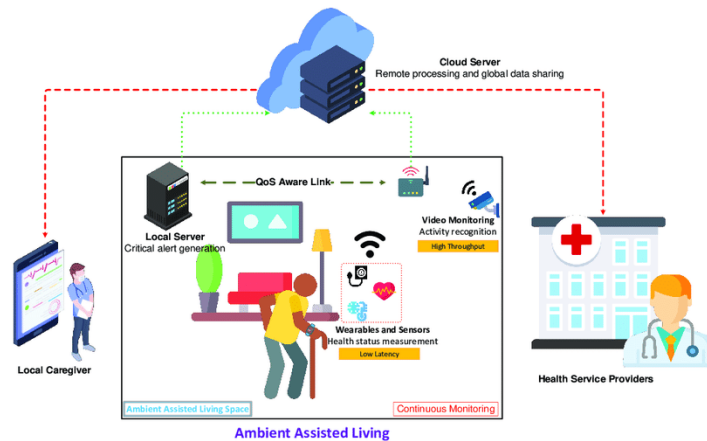


Figure 2: The model for AAL.

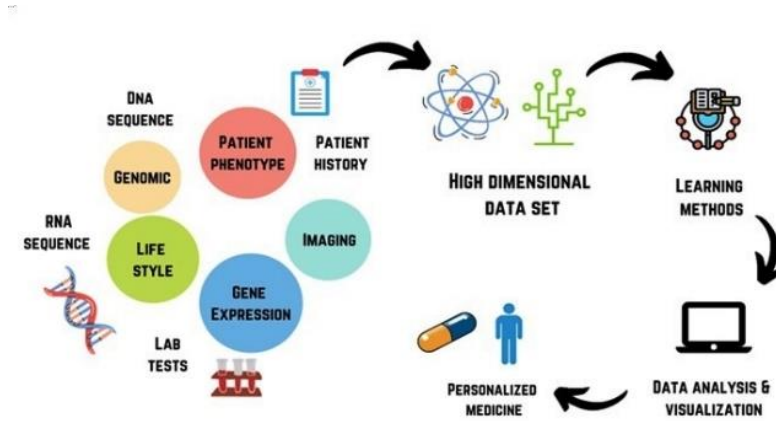


Figure 3: AI in healthcare and data processing.

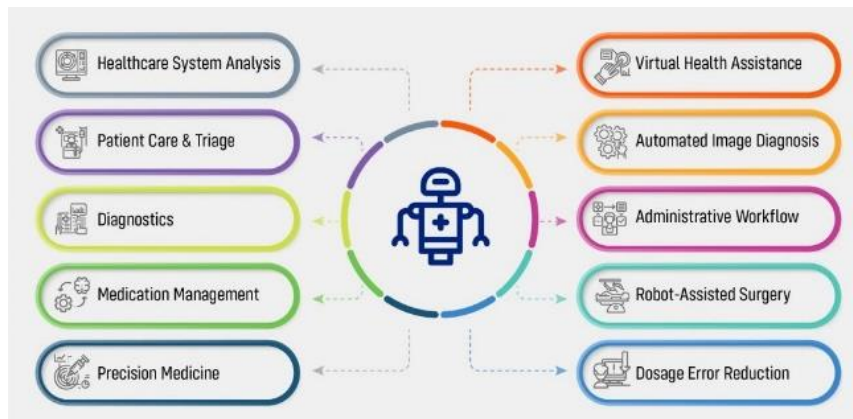


Figure 4: AI in healthcare applications.

3.1 Artificial Intelligence in Healthcare Data Processing

A major development has taken place in NLP applications for the biomedical field. Biomedical Question Answering (BioQA) describes the task of efficient document-database query that retrieves precise responses to user inquiries from established collections. Algorithms based on NLP will actively pursue informative responses according to [29]. The initially classification of biomedical inquiries enables researchers to obtain useful information from gathered answers. According to [30] machine learning performs four-category biomedical inquiry categorization with over 90% accuracy.

A long-term acquisition of biological data across multiple sources requires performing critical operations which involve medical data integration and analysis in addition to resolving conflicting information [31]. Endless human labor along with

intense manual work and inadequate results have streaked through these kinds of tasks throughout history. The accuracy of AI in performing these tasks proves equal to professional evaluators thus leading to more efficient and precise operations [32]. The acquired dataset knowledge resulting from machine learning algorithms enables researchers to apply the integrated logical reasoning to elaborate clinical information including diverse biomedical data types [33]. Figure 3 shows AI in healthcare.

3.2 AI in Healthcare Research

AI demonstrates potential to become powerful tool for biomedical research in addition to its current role as an electronic diagnostic and management system [34]. The current study domain investigates tumor-suppressor processes, the extraction of protein-protein interactions, and the establishment of genome-based genetic connections to enhance healthcare practice implementations [35,36].

Advanced medical technology shows increasing signs of gaining awareness which biomedical research can analyze. Using the computational modeling assistant (CMA) sophisticated agent biomedical researcher can build executable simulation models from their conceptual frameworks according to [37]. The CMA system employs its intelligence to process all gathered knowledge alongside simulation models which converts hypothesis-based concepts into actual simulation codes. The researcher conducts an evaluation to select the best models for the CMA before the CMA creates simulation codes from these selected models. Research time reduction along with enhanced production stands as a key benefit of working with the CMA system. Figure 4 shows AI in Healthcare applications.

4 PROGNOSIS AND DISEASE DETECTION

Disease detection represents the most essential reason why AI is crucial in biomedicine. The field has generated several important advancements during its recent development period. Artificial intelligence technology allows medical providers to make accurate diagnoses earlier in time for various diseases [38] - [40]. Artificial intelligence uses machine learning to analyze microarray data through which it establishes patient-specific patterns for diagnostic purposes. The classification of cancer microarray data using this approach serves as a new application for diagnostics in oncology [41]. The combination of integrated AI helps biosensors function with point-of-care testing devices to detect cardiovascular diseases during early diagnosis stages [42]. AI demonstrates its use in predicting survival rates of cancer patients with colon cancer in addition to diagnosis tasks [43].

AI holds strong potential for clinical usage in both the diagnostic and prognostic domains. The field of illness diagnostics utilizes two fundamental processing approaches which include two-dimensional medical imaging and one-dimensional signal analysis and computation. The approaches serve medical practitioners for diagnosing patients and handling their illness and predicting health outcomes [44]. The processing of one-dimensional biomedical signals through AI features the extraction of vital information through EEG and EMG and ECG

signals. EEG technology represents an essential method to forecast epileptic seizures. Proper prediction of seizures helps medical professionals reduce the impacts on patients [45] - [47].

AI applications in medical imaging targets three areas for both quality enhancement and workflow optimization it includes thermal imaging, image segmentation and multidimensional imaging [48]. The implementation of AI systems improves standard decision support systems (DSSs) by enhancing diagnostic accuracy and automation of illness management which reduces the workload for medical staff. Figure 5 shows AI diagnosis in heart disease.

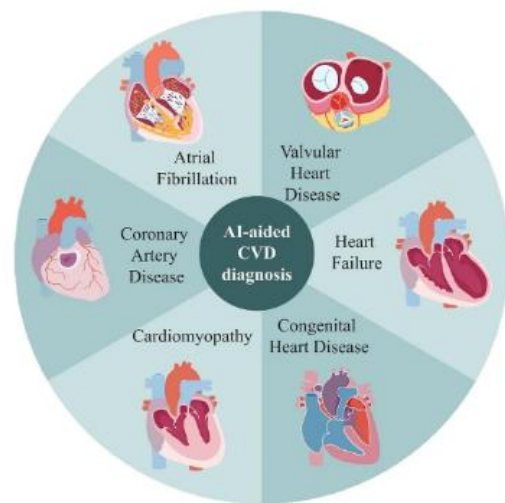


Figure 5: AI diagnosis in heart disease.

5 SMART CITY AND HEALTHCARE FACILITIES

Numerous cities that do not have smart city status seek to build standard healthcare facilities in order to achieve the smart city healthcare approach. Smart cities need to achieve these objectives to provide effective healthcare services to people by ensuring high-quality residential areas and healthcare services within an environment supporting healthy living [49], [50]. The establishment of a mechanism is needed to develop innovative and effective healthcare solutions. The necessary elements of IoT healthcare systems emerge from a combination of different systems and architectures and frameworks as shown in Figure 6.

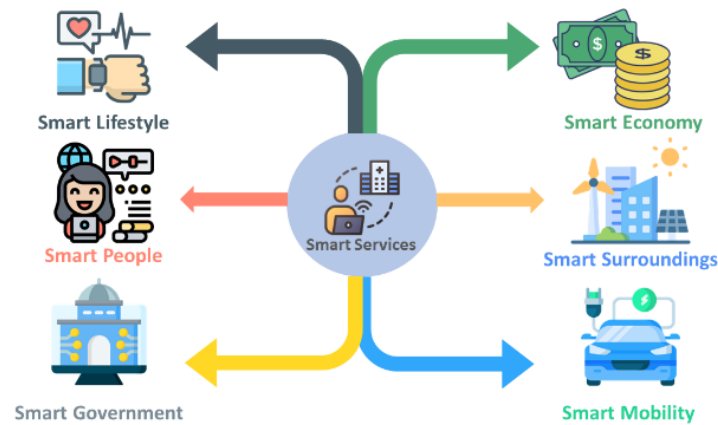


Figure 6: IoT healthcare systems architectures and frameworks.

Table 2: Machine learning applications in smart cities.

Algorithm	Application	Function
Clinical Correlation Analysis (CCA)	Used to create medical answers by integrating devices, instruments, and cases.	Extracts feature from data; Highlights connections between data sets through visualization and comparison.
Neural Network (NN)	Enhances medical instruments, software and hardware; Applied in rehabilitation administration.	Provides a learning approach for addressing approximate function evaluation problems.
Support Vector Machines (SVM)	Forecasting healthcare solutions; Real-time detection of blood sugar levels.	Controls and categorizes large volumes of data; Handles diverse data types; High-speed data processing and transmission.
K-means	Healthcare delivered in the convenience of your residence and within the vibrancy of an urban environment.	for handling massive datasets, it uses clustering to detect patterns in data not identified yet.
DB-SCAN	M-IoT network processes data from human sensors to gauge communication efforts.	Facilitates the visualization of the organization of unstructured data; Ensures compatibility with multiple file formats.

The six core elements which constitute the concept of "smart services" are namely Smart economy; smart surroundings; smart mobility, smart governments, smart lifestyle and smart people. The various manifestations of services have been comprehensively classified. These services enable citizens to access and utilize a number of intelligent devices for their benefits [51]. Internet users who give extensive personal data access result in highly sophisticated services built upon complex network structures. Telemedicine centers help reduce disease spread because they provide medical care to patients who remain in isolation [52]. Machine learning applications in smart cities appear as described in Table 2.

6 MOBILE-BASED HEALTH MONITORING

Mobiles stand as the most beneficial devices present in modern society. The speech monitoring function in smartphones works as a powerful evaluation tool. Multiple hardware systems have been created to exploit this flexibility. The modern smartphone design relies on NFC (near field communication) sensors and GPS trackers, fingerprint scanners, Bluetooth modules, accelerometers, barometers, magnetometers and gyroscopes and proximity sensors and wireless sensors and cameras for health

monitoring system development [53]. Cellphones provide various benefits to users with their significant storage capability that allows them to keep data securely. Modern smartphones possess enough primary storage to save patient information. Android smartphones enable instant access to information dissemination at the same time as device management and real-time communication capabilities [54].

7 CONCLUSIONS

Medical care improves through AI and Smart Health Monitoring systems by solving problems of delayed medicine delivery while delivering preventive care measures and providing prompt real-time assistance. Remote Health Monitoring (RHM) receives enhanced security and personalized healthcare solutions through cloud computing and blockchain technologies which also establish better patient trust. Biomedicine experiences continuous expansion of AI functions which develops new possibilities for diagnostic assessment and disease prognostication as well as biomedical scientific study. The parallel progress of biomedical sciences and artificial intelligence (AI) is giving rise to increasingly intelligent and efficient healthcare solutions. As these two fields continue to evolve together, they pave the way for systems that not only enhance diagnostic accuracy and treatment planning but also streamline healthcare delivery across various levels. In this context, further research in the area of Health Management is essential to fully explore its broader applications in continuous health monitoring, early disease detection, personalized medicine, and predictive analytics. Leveraging the rapid advancements in AI – such as machine learning, deep learning, and data-driven decision-making – can lead to significant improvements in healthcare outcomes. This synergy has the potential to revolutionize how patient care is administered, leading to more proactive and preventive approaches. Ultimately, integrating AI into Health Management frameworks will contribute to enhancing the overall quality of life, promoting well-being, and ensuring more equitable and accessible healthcare systems worldwide.

8 FUTURE WORK

Historically deep neural networks together with deep learning techniques operated as black boxes which made their core processing mechanisms poorly understandable. Understanding deep learning models

becomes possible through visual insights into learned representations together with implemented models which enables simpler development and optimization of deep learning applications for complex machine health monitoring tasks. Transfer learning describes the method of applying deep learning techniques between different information domains. Machine health monitoring studies benefit from this investigation since problems have different levels of training data availability. Machine learning models as well as deep learning models have the capability to move from one domain to another.

REFERENCES

- [1] S. Amil et al., "Interactive Conversational Agents for Perinatal Health: A Mixed Methods Systematic Review," *Healthcare*, vol. 13, no. 4, p. 363, 2025, [Online]. Available: <https://doi.org/10.3390/healthcare13040363>.
- [2] K. Krishna and K. Uma, "A Review on Current Applications of Artificial Intelligence in Medical Field," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 626-633, 2023, [Online]. Available: <https://doi.org/10.48175/IJARSC-11497>.
- [3] S. Hossain, S. Umer, R. K. Rout, and H. A. Marzouqi, "A Deep Quantum Convolutional Neural Network Based Facial Expression Recognition For Mental Health Analysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 1556-1565, 2024, doi: 10.1109/TNSRE.2024.3385336.
- [4] M. S. A. Nadeem et al., "Hybridizing Artificial Neural Networks Through Feature Selection Based Supervised Weight Initialization and Traditional Machine Learning Algorithms for Improved Colon Cancer Prediction," *IEEE Access*, vol. 12, pp. 97099-97114, 2024, doi: 10.1109/ACCESS.2024.3422317.
- [5] H. Saadi and A. M. Abdulshahed, "Reinforcement Learning Applications in Medical Imaging," in *2024 8th International Conference on Image and Signal Processing and their Applications (ISPA)*, Biskra, Algeria, 2024, pp. 1-6, doi: 10.1109/ISPA59904.2024.10536861.
- [6] K. Yu et al., "Hybrid Quantum Classical Optimization for Low-Carbon Sustainable Edge Architecture in RIS-Assisted AIoT Healthcare Systems," *IEEE Internet of Things Journal*, vol. 11, no. 24, pp. 38987-38998, 2024, doi: 10.1109/JIOT.2024.3399234.
- [7] S. Bak, J. Park, J. Lee, and J. Jeong, "Memristor-Based CNNs for Detecting Stress Using Brain Imaging Signals," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 1, pp. 140-149, 2024, doi: 10.1109/TETCI.2023.3297841.
- [8] M. Manokhin, P. Chollet, and P. Desgreys, "Flexible Analog-to-Feature Converter for Wireless Smart Healthcare Sensors," in *2023 21st IEEE Interregional NEWCAS Conference (NEWCAS)*, Edinburgh, United Kingdom, 2023, pp. 1-5, doi: 10.1109/NEWCAS57931.2023.10198150.

- [9] A. Lone et al., "Multilayer Ferromagnetic Spintronic Devices for Neuromorphic Computing Applications," Research Square, 2023, [Online]. Available: <https://doi.org/10.21203/rs.3.rs-3839002/v1>.
- [10] X. Zhang, A. Huang, Q. Hu, Z. Xiao, and P. K. Chu, "Neuromorphic computing with memristor crossbar," *Physica Status Solidi (A)*, vol. 215, no. 13, p. 1700875, 2018.
- [11] S. Babu et al., "AI-Driven Healthcare: Predictive Analytics for Disease Diagnosis and Treatment," *International Journal for Modern Trends in Science and Technology*, vol. 10, pp. 5-9, 2024, [Online]. Available: <https://doi.org/10.46501/IJMTST1006002>.
- [12] H. Al Gburi, M. Algburi, and H. Al-Saedi, "Compact Antenna Design for RFID and IoT Applications," in 2022 International Conference on Microelectronics (ICMI), 2022, pp. 1-4, doi: 10.1109/ICMI55296.2022.9873727.
- [13] D. Jaafar and M. H. Algburi, "Path Planning in 3D space using Butterfly Optimization Algorithm," *Periodicals of Engineering and Natural Sciences*, vol. 9, no. 2, pp. 976-985, 2021, [Online]. Available: <https://doi.org/10.21533/pen.v9i2.1812>.
- [14] H. Al Gburi and A. Al-baiati, "Path-Planning Dynamic 3D Space Using Modified A* Algorithm," in IOP Conference Series: Materials Science and Engineering, vol. 928, p. 032016, 2020, [Online]. Available: <https://doi.org/10.1088/1757-899X/928/3/032016>.
- [15] A. M. A. Al-Sabaawi et al., "Exploiting the Crow Search Algorithm to Overcome the Bandpass Problem," in *Forthcoming Networks and Sustainability in the AIoT Era*, vol. 1036, pp. 184-198, Springer, Cham, 2024, [Online]. Available: https://doi.org/10.1007/978-3-031-62881-8_14.
- [16] K. Huang, H. Choudhary, and P. Kuo, "Artificial Intelligent Agent Architecture and Clinical Decision-Making in the Healthcare Sector," *Cureus*, vol. 16, p. e64115, 2024, [Online]. Available: <https://doi.org/10.7759/cureus.64115>.
- [17] R. A. S. Naseri, A. Kurnaz, and H. M. Farhan, "Optimized face detector-based intelligent face mask detection model in IoT using deep learning approach," *Applied Soft Computing*, vol. 134, p. 109933, 2023, [Online]. Available: <https://doi.org/10.1016/j.asoc.2022.109933>.
- [18] M. A. Hamza et al., "Wearables-assisted smart health monitoring for sleep quality prediction using optimal deep learning," *Sustainability*, vol. 15, no. 2, p. 1084, 2023, [Online]. Available: <https://doi.org/10.3390/su15021084>.
- [19] A. Zhu et al., "Condition monitoring of wind turbine based on deep learning networks and kernel principal component analysis," *Computers and Electrical Engineering*, vol. 105, p. 108538, 2023, [Online]. Available: <https://doi.org/10.1016/j.compeleceng.2022.108538>.
- [20] M. Moradi et al., "Intelligent health indicator construction for prognostics of composite structures utilizing a semi-supervised deep neural network and SHM data," *Engineering Applications of Artificial Intelligence*, vol. 117, p. 105502, 2023, [Online]. Available: <https://doi.org/10.1016/j.engappai.2022.105502>.
- [21] A. M. Helmi, M. A. Al-qaness, A. Dahou, and M. Abd Elaziz, "Human activity recognition using marine predators algorithm with deep learning," *Future Generation Computer Systems*, vol. 142, pp. 340-350, 2023, [Online]. Available: <https://doi.org/10.1016/j.future.2023.01.006>.
- [22] L. S. Kondaka, M. Thenmozhi, K. Vijayakumar, and R. Kohli, "An intensive healthcare monitoring paradigm by using IoT based machine learning strategies," *Multimedia Tools and Applications*, vol. 81, no. 26, pp. 36891-36905, 2022.
- [23] M. Hudec and Z. Smutny, "RUDO: A Home Ambient Intelligence System for Blind People," *Sensors*, vol. 17, no. 8, p. 1926, 2017, [Online]. Available: <https://doi.org/10.3390/s17081926>.
- [24] I. Machorro-Cano et al., "Cloud-Based Platforms for Health Monitoring: A Review," *Informatics*, vol. 11, no. 1, p. 2, 2024, [Online]. Available: <https://doi.org/10.3390/informatics11010002>.
- [25] H. Cho et al., "Fall Detection Based on Continuous Wave Radar Sensor Using Binarized Neural Networks," *Applied Sciences*, vol. 15, no. 2, p. 546, 2025, [Online]. Available: <https://doi.org/10.3390/app15020546>.
- [26] M. A. Qureshi et al., "Deep learning-based ambient assisted living for self-management of cardiovascular conditions," *Neural Computing and Applications*, vol. 34, pp. 10449-10467, 2022, [Online]. Available: <https://doi.org/10.1007/s00521-020-05678-w>.
- [27] M. H. Abidi, "Multimodal data-based human motion intention prediction using adaptive hybrid deep learning network for movement challenged person," *Scientific Reports*, vol. 14, p. 30633, 2024, [Online]. Available: <https://doi.org/10.1038/s41598-024-82624-z>.
- [28] F. H. Saif, M. N. Al-Andoli, and W. M. Y. W. Bejuri, "Explainable AI for Alzheimer Detection: A Review of Current Methods and Applications," *Applied Sciences*, vol. 14, no. 22, p. 10121, 2024, [Online]. Available: <https://doi.org/10.3390/app142210121>.
- [29] N. Jofche et al., "PharmKE: Knowledge Extraction Platform for Pharmaceutical Texts Using Transfer Learning," *Computers*, vol. 12, no. 1, p. 17, 2023, [Online]. Available: <https://doi.org/10.3390/computers12010017>.
- [30] H. Bolat and B. Şen, "Document Retrieval System for Biomedical Question Answering," *Applied Sciences*, vol. 14, no. 6, p. 2613, 2024, [Online]. Available: <https://doi.org/10.3390/app14062613>.
- [31] M. Voskamp et al., "Integrating Text Mining into the Curation of Disease Maps," *Biomolecules*, vol. 12, no. 9, p. 1278, 2022, [Online]. Available: <https://doi.org/10.3390/biom12091278>.
- [32] E. Victor and T. Ogbonna, "Developing and Analyzing Deep Learning and Natural Language Processing Systems in the Context of Medical Information Processing," *International Journal of Research and Innovation in Applied Science*, vol. IX, pp. 733-739, 2024, [Online]. Available: <https://doi.org/10.51584/IJRIAS.2024.908065>.
- [33] H. J. Duggirala et al., "Artificial intelligence and machine learning in veterinary medicine: a regulatory perspective on current initiatives and future prospects," *American Journal of Veterinary Research*, 2025, [Online]. Available: <https://doi.org/10.2460/ajvr.24.09.0285>.

- [34] H. Nishiki et al., "Integrated Analysis of Somatic DNA Variants and DNA Methylation of Tumor Suppressor Genes in Colorectal Cancer," *International Journal of Molecular Sciences*, vol. 26, no. 4, p. 1642, 2025, [Online]. Available: <https://doi.org/10.3390/ijms26041642>.
- [35] S. Tuladhar et al., "A novel solution of deep learning for endoscopic ultrasound image segmentation: enhanced computer aided diagnosis of gastrointestinal stromal tumor," *Multimedia Tools and Applications*, vol. 81, pp. 1-21, 2022, [Online]. Available: <https://doi.org/10.1007/s11042-022-11936-x>.
- [36] C. Qin et al., "A Novel Rational Medicine Use System Based on Domain Knowledge Graph," *Electronics*, vol. 13, no. 16, p. 3156, 2024, [Online]. Available: <https://doi.org/10.3390/electronics13163156>.
- [37] A. Sanobar et al., "Plastic Surgery Awareness and Perceptions Among Healthcare Professionals: A Single Centre Study at Liaquat University Hospital," *Pakistan Journal of Health Sciences*, vol. 6, no. 1, pp. 185-190, 2025, [Online]. Available: <https://doi.org/10.54393/pjhs.v6i1.2610>.
- [38] S. Zabyan and A. Ali, "Detection Covid-19 Based on Chest X-ray Images Using Convolution Neural Networks," *International Journal of Computers, Communications & Control*, vol. 22, pp. 34-42, 2022, [Online]. Available: <https://doi.org/10.33103/uot.ijccce.22.1.4>.
- [39] S. Zabyan and A. Ali, "Identification of Pneumonia Based on Chest X-Ray Images Using Wavelet Scattering Network," *Periodicals of Engineering and Natural Sciences*, vol. 9, no. 2, pp. 976-985, 2021, [Online]. Available: <https://doi.org/10.21533/pen.v9i2.1953>.
- [40] D. Tripathi et al., "Artificial Intelligence in Biomedical Engineering and Its Influence on Healthcare Structure: Current and Future Prospects," *Bioengineering*, vol. 12, no. 2, p. 163, 2025, [Online]. Available: <https://doi.org/10.3390/bioengineering12020163>.
- [41] J.-H. Wang et al., "Cancer Diagnosis by Gene-Environment Interactions via Combination of SMOTE-Tomek and Overlapped Group Screening Approaches with Application to Imbalanced TCGA Clinical and Genomic Data," *Mathematics*, vol. 12, no. 14, p. 2209, 2024, [Online]. Available: <https://doi.org/10.3390/math12142209>.
- [42] A. M. Alyami et al., "Implantable Biosensors for Vascular Diseases: Directions for the Next Generation of Active Diagnostic and Therapeutic Medical Device Technologies," *Biosensors*, vol. 15, no. 3, p. 147, 2025, [Online]. Available: <https://doi.org/10.3390/bios15030147>.
- [43] O. Attallah, "Lung and Colon Cancer Classification Using Multiscale Deep Features Integration of Compact Convolutional Neural Networks and Feature Selection," *Technologies*, vol. 13, no. 2, p. 54, 2025, [Online]. Available: <https://doi.org/10.3390/technologies13020054>.
- [44] M. Almutairi et al., "Detecting Elderly Behaviors Based on Deep Learning for Healthcare: Recent Advances, Methods, Real-World Applications and Challenges," *IEEE Access*, vol. 10, pp. 69802-69821, 2022, doi: 10.1109/ACCESS.2022.3186701.
- [45] I. T. Abdullah et al., "Joining of Aa1050 sheets via two stages of friction spot technique," *Journal of Mechanical Engineering Research and Developments*, vol. 44, no. 4, pp. 305-317, 2021.
- [46] M. Q. Algburi, *Wearable Antenna Design for IoT and Biomedical Applications*, Saarbrücken, Germany: LAP Lambert Academic Publishing, pp. 1-132, 2022.
- [47] M. Algburi and M. Ilyas, "A Novel Design Reconfigurable Antenna Based on the Metamaterial for Wearable Applications," in *Journal of Physics: Conference Series*, vol. 1973, p. 012042, 2021.
- [48] R. M. H. et al., "Computer Vision and Image Processing the Challenges and Opportunities for New Technologies Approach: A paper review," *Journal of Physics: Conference Series*, vol. 1973, p. 012002, 2021, [Online]. Available: <https://doi.org/10.1088/1742-6596/1973/1/012002>.
- [49] Q. H. Mohammed and A. Namburu, "Hybrid Model and Framework for Predicting Air Pollutants in Smart Cities," *Journal of Engineering and Sustainable Development*, vol. 28, no. 3, pp. 392-406, 2024, [Online]. Available: <https://doi.org/10.31272/jeasd.28.3.9>.
- [50] G. A. Aramice, A. H. Miry, and T. M. Salman, "Optimal Long-Range-Wide-Area-Network Parameters Configuration for Internet of Vehicles Applications in Suburban Environments," *Journal of Engineering and Sustainable Development*, vol. 27, no. 6, pp. 754-770, 2023, [Online]. Available: <https://doi.org/10.31272/jeasd.27.6.7>.
- [51] W. A. C. Castañeda and P. Bertemes Filho, "Improvement of an Edge-IoT Architecture Driven by Artificial Intelligence for Smart-Health Chronic Disease Management," *Sensors*, vol. 24, no. 24, p. 7965, 2024, [Online]. Available: <https://doi.org/10.3390/s24247965>.
- [52] K. Mani, K. K. Singh, and R. Litoriya, "IoT-Driven Revolution: Transforming Elderly Healthcare Through Dynamic Heart Health Monitoring and Analysis," in *2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET)*, Indore, India, 2024, pp. 1-8, doi: 10.1109/ACROSET62108.2024.10743704.
- [53] M. R. Ananda et al., "Design of Application Framework for Vital Monitoring Mobile-Based System," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 10, no. 2, pp. 252-264, 2024, [Online]. Available: <https://doi.org/10.26555/jiteki.v10i2.28416>.
- [54] A. N. Muhi et al., "Innovative Designs for Enhanced Connectivity for 3G and 4G Cellular Networks," in *2024 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, Bandung, Indonesia, 2024, pp. 56-61, doi: 10.1109/ICRAMET62801.2024.10809318.

SECTION 5

STATISTICAL MODELLING AND ANALYSIS AND THEIR APPLICATIONS IN LIFE SCIENCES

Proceedings of the 13th International Conference on Applied Innovations in IT (ICAIIIT)

JULY 2025

Integrated Features Based on Graph Clustering and Gene Expression

Sura Ibrahim Mohammed Ali^{1,2} and Sura Zaki Al Rashid¹

¹Department of Software, College of Information Technology, University of Babylon, 51001 Babylon, Iraq

²Department of Mathematics and Computer Applications, College of Science, Al-Muthanna University, 66001 Sumaway, Iraq
suraibraheem@mu.edu.iq, suraibrahimm.sw@student.uobabylon.edu.iq, sura.alrashid@uobabylon.edu.iq

Keywords: Topological Analysis, Multi-Feature Framework, Gene Expression, Graph Clustering, Gene Regulatory NetWork.

Abstract: Integrating different biological features – for example, the informativeness of topological features and gene expression – is challenging because each feature must be accounted for individually if the features are used to help forecast models. In this process, ensuring that the outcomes reflect the underlying biological structure of the network information while minimizing noise and irrelevant data is crucial. This study identifies the importance of rigorous pre-analyses in determining statistically significant correlations and joint effects among preprocess features before applying machine-learning techniques. Thus, when deploying multidimensional datasets, a systematic multi-feature methodology is presented in this paper to unify optimized graph clustering, weighted Jaccard similarity, and dimension reduction based on principal component analysis (PCA). Specifically, the objective was to identify novel uncharacterized gene associations in complex biological networks. Moreover, this study offers more refined insights into gene interactions within their networks, revealing patterns and relationships that might be hidden by broad data analysis. The method's performance was validated according to the benchmarks for a Dialogue on Reverse Engineering Assessment of Methods, fifth edition (DREAM5) challenge project, to determine its ability to analyze complex biological networks.

1 INTRODUCTION

Biological systems often involve complex interactions among numerous genes and chemicals within cellular networks, especially in multifactorial diseases [1], [2]. Advances in experimental and computational methods now enable precise mapping of physical (e.g., protein-protein, signaling, and regulatory) and functional (Yet, the scale of these networks poses challenges for extracting biological insights, making community detection a vital graph-clustering task for identifying functional modules [3].

While many algorithms have been tested on benchmark in silico networks, their utility in real molecular contexts remains uncertain. Nevertheless, identifying modules is critical for downstream analyses like link prediction and disease mechanism interpretation. Features from community structures and gene expression data can be combined to detect regulatory patterns and deeper functional associations. Dimensionality reduction techniques such as PCA help extract biologically relevant features from high-dimensional gene expression data. As early as 2010, PCA was shown to improve cancer

subtype classification by incorporating biological structure [4]. Moreover, integrating regulatory networks into statistical models allows for a richer interpretation of genetic architecture and biological function [5]. By combining community data with PCA loadings, researchers gain deeper insights into genetic network organization, surpassing traditional analytical methods [6]. This paper presents a novel multi-feature framework that integrates gene expression profiles, community structures, and principal components to identify highly variable genes within selected modules – prioritizing targets for experimental validation and enhancing biological understanding.

2 RELATED WORK

The graph structures are non-Euclidean, meaning they are characterized by irregular arrangements, make it hard to identify the nearby genes of a given point in the data, and vary in the number of surrounding nodes of various genes. Several recent studies have highlighted the importance of clustering

as a critical step in single-cell data integration and analysis. The method's primary innovation in this study [7] is its simultaneous consideration of the condition manifold's and gene manifold's internal geometric structures, this novel approach used DGPCA, its Laplacian embedding to approximate the cluster membership indicators and obtain principal components (PCs) to characterize the data. The condition manifold and gene manifold, two internal geometric structures that are appropriate for bi-clustering, are features of DGPCAA new interpretable framework for solving the single-cell RNA-Seq clustering challenge was proposed in [8]. The system could generate many clustering's of the same dataset at a low cost, in addition to retrieving results from a wide range of single-cell research. The algorithm was able to provide interpretable justifications for each of its choices and be used as a backend algorithm for interactive interpretation and analysis. In [9] to extract attribute information, the scDFN algorithm uses a dual mechanism that includes an autoencoder. This is the first study to use an enhanced graph network for single-cell topological data representation. Using the triple self-supervision technique and the cross-network information fusion mechanism, the information fusion module merges the retrieved attribute and topology data. To maximize the cell clustering representation throughout the entire model, quadruple joint losses are used. scGMAI, a novel Gaussian mixture clustering technique based on autoencoder networks and FastICA, was proposed in [10]. It is a powerful tool for precisely classifying and distinguishing cell types from scRNA-Seq data and demonstrates the wide range of applications it can have in scRNA-Seq data analysis. The defining genes of cell clusters are more precisely identified by this methodology. In scGMAI, FastICA selects the most important independent features to create a low-dimensional space that retains all of the data's fundamental properties. In order to assess how effectively biological variation is maintained during integration, the study in [11] has extensively benchmarked data integration techniques utilizing Louvain clustering. Researchers were able to statistically evaluate the quality of integration by comparing cluster assignments with reference cell type labels using measures such as NMI and ARI. This method has made clustering, when combined with strong validation metrics, a common technique for single-cell genomics batch correction evaluation and biological discovery.

3 METHODOLOGY

The nature of molecular networks is modular, and the subsets of nodes are more interconnected than pure chance would suggest. These subsets often consist of genes or proteins that share biological functions. Therefore, community detection is an intrinsic step towards deriving findings from network data [12]. However, biological networks' sheer scale and complexity present an obstacle to their analysis. Consequently, the present researcher has focused on community structures to identify functionally related groups of genes or proteins involved in biological processes. Here, together with module identification and graph clustering, the traditional network science task of community detection was the objective of applying a number of proposed methods [13].

In particular, this paper [14] reports on detecting protein complexes based on gene expressions. However, the resulting framework addresses broader challenges, such as integrating multiple data sources and analyzing complex networks. Nevertheless, despite these advances, there is still very little understanding of how the different approaches identify biologically meaningful communities in molecular networks. For instance, although the Dialogue on Reverse Engineering Assessment of Methods (DREAM) project demonstrated robust methods of network inference [15], very little downstream analysis has been conducted of the reconstructed networks, for example, in predicting regulatory pairs or discovering new links.

Hence, a broad three-component framework is proposed in this study, deploying robust methods of community detection to identify groups of functionally related genes accurately. Consequently, additional biological insights were anticipated, together with a PCA-based feature selection strategy for reducing the dimensionality of gene expression data by selecting various pertinent principal components (PCs), according to cumulative variance. This framework is schematically represented in Figure 1, which illustrates the workflow and further clarifies the functions of the regulatory gene network. Here, highly diverse features are typically gathered from a variety of databases (see top panel).

A feature table or matrix is then produced by merging the feature data with the already identified gene regulatory network, sourced from open databases like DREAM5. From this gene regulatory network, appropriate algorithms are applied to the relevant genes, selected by graph clustering from, for example, an *E. coli* dataset.

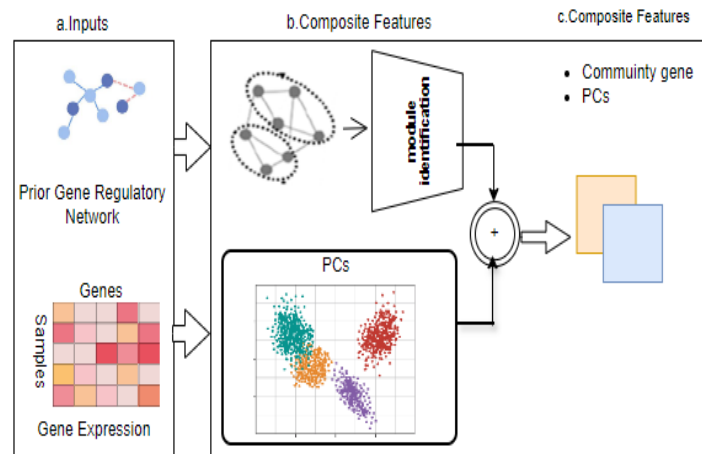


Figure 1: Overview of the proposed scheme for integrating biological data.

3.1 Feature Selection and Extraction

Feature selection enhances predictive accuracy and deepens understanding of biological networks, especially in the context of genomics and personalized medicine where accurate prediction is crucial [16]. To achieve this, the paper employs two main criteria: the "Predicting Communities" approach, which leverages prior knowledge of gene regulatory networks, and Principal Component Analysis (PCA), used to select the top 10 features (principal components) based on variance. Detailed descriptions of these criteria follow in the methodology section.

Features were extracted from both regulatory and inferred gene expression networks, focusing on dynamic elements such as transcription factors and co-expressed genes. Communities of highly interconnected genes and transcription factors were also incorporated. These were merged into a unified feature matrix, offering an integrated view of diverse biological data. Subsequent unsupervised clustering, based purely on network topology, identified non-overlapping communities ranging from 3 to 226 elements. This integrative approach supports robust exploration of molecular network structure.

3.2 Predicting Communities

In order to predict communities of related genes referred to as 'regulatory communities' or 'gene regulatory networks', it is crucial to cluster features using robust methods of community detection.

A framework of module predictions is proposed, consisting of four phases and represented in Figure 2. The architecture of the proposed framework consists of four key phases:

- 1) community detection to identify functional groups;
- 2) parameter optimization;
- 3) integration of transcription factors and target genes;
- 4) module validation to link biological functions with the detected communities.

In brief, Phase 1 consists of communities being used to seed or initialize groups of genes, in order to preprocess communities. The groups are refined by ranking and selecting those features that will best preserve the local structure, using mutual information.

In Phase 2, weighted Jaccard similarity takes place. The key difference between normal and weighted Jaccard is that the latter does not operate with binary features alone; instead, the numerical values of each feature used to calculate similarity. Specifically, in weighted Jaccard, the numerical values of each feature considered to compute similarity, as in the proposal, using the numerical values of the genes, for example, PageRank, in_degree, etc. Genes that show very similar numerical values have high similarity, thereby increasing the accuracy of the gene grouping in populations.

In Phase 3, information about aggregation attributes, shared neighbours, and connectivity is used to build similarity matrices, thereby enhancing the performance of the existing community detection algorithms.

Finally, in Phase 4, communities are clustered on the similarity matrix to obtain the final clustering (module). All four of these phases are described in detail in the following sub-sections.

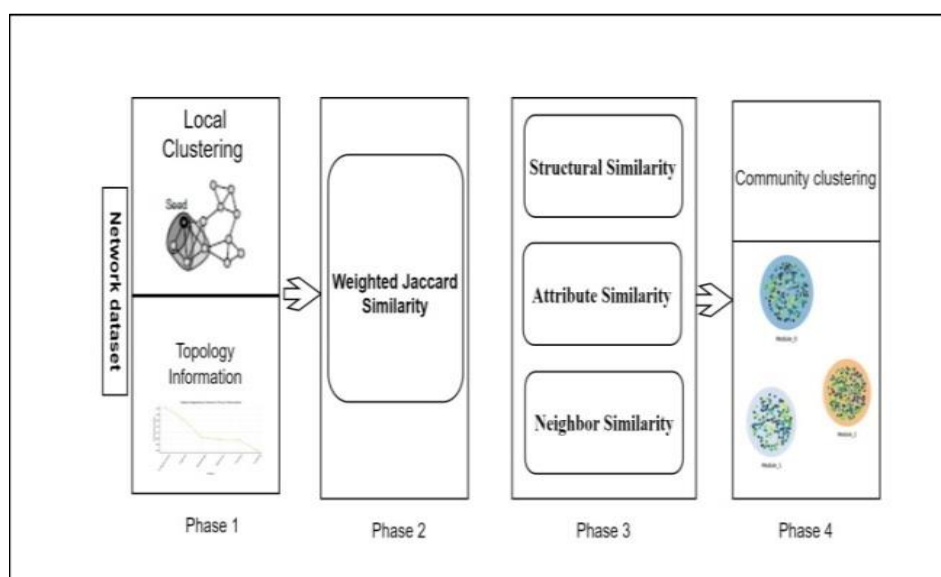


Figure 2: Framework of module predictions.

The communities illustrated in Figure 2 are made up of groups of genes and transcription factors that collectively perform specific biological functions, such as stress response or cell differentiation. Regulatory communities not only help shed light on the regulation of gene expression but also reveal the functional relationships between genes. This information helps to identify biological pathways and understand cellular processes. For instance, in stress response, a regulatory module could include stress resistance genes and their transcriptional regulators. Meanwhile, a cell differentiation network may contain differentiation genes and the transcription factors that control their expression. Therefore, this paper continues and extends the work reported in [17], introducing community detection algorithms that are enhanced by parameter learning. The latter has previously been applied in the literature to social networks but is now extended to biological networks. This parameter controls the tendency of nodes to form communities, consequently influencing community size by optimizing the modularity function.

The parameter would inversely influence the size of the detected communities, thereby increasing the value results in smaller, more tightly knit communities. This adjustment offers a scalable and flexible approach to analyzing large and complex biological networks. Unsupervised clustering algorithms were therefore employed, relying solely on network structure (link and attribute information) without additional biological annotations, consequently ensuring unbiased module identification.

3.2.1 Phase 1: Local Clustering and Topological Features

Phase 1 was subdivided into two stages: local clustering and topological feature extraction:

- 1) Local Clustering Phase. Initial clusters were formed using cluster density from the gene network, followed by attribute weighting and refinement via the Louvain modularity optimization algorithm [18]. The DICCA algorithm was employed to group genes based on proximity and attributes, targeting densely connected subgroups. Jaccard similarity and modularity optimization further enhanced cluster identification and weighting.
- 2) Topological Feature Stage. Topological features – such as eigenvalue, closeness, and edge – were extracted to assess their influence on the target variable. Mutual information was used to rank these features, with the most informative ones selected for prediction tasks.

3.2.2 Phase 2: Weighted Jaccard Similarity Algorithm

Once the initial clusters are formed, the weighted Jaccard similarity algorithm can be applied as the key approach to detecting communities within biological networks. This algorithm is specifically designed to render similarity measures more relevant by accounting for the strength and importance of interactions. Other than the original Jaccard index,

which assigns equal weights to all attributes of interest, the weighted Jaccard similarity algorithm weights attributes according to their importance, minimizing the effect of weak/nonspecific interaction. In this paper, the weighted Jaccard similarity of two genes was calculated based on (1):

$$\text{Similarity}(i, j) = \frac{\sum \min(\text{atti}, \text{attj})}{\sum \max(\text{atti}, \text{attj})} \quad (1)$$

In (1), *atti* and *attj* are the attribute vectors for the clustered *i*-th and *j*-th genes. This ensures that more importance is given to the features with higher values, while calculating the similarity score between the genes within a single cluster. It helps to identify biologically significant gene communities.

After constructing a Jaccard similarity matrix (see Definition 2) and calculating attribute weights (see Definition 3), the next logical steps consist of leveraging these definitions, in order to gain further insights into gene clustering and network analysis. On the basis of the attribute weights calculated therefrom, the algorithm emphasizes the importance of significant interactions and minimizes the influence of nonspecific interactions.

Definition 2: Jaccard similarity matrix construction: $J = [\text{similarity}(i, j)]$, where J is the Jaccard similarity matrix for a cluster, and i, j represent the index genes within that cluster. This matrix is populated by computing the Jaccard similarity for each pair of genes in the cluster.

Definition 3: Attribute weight calculation:

$$w_k = \frac{1}{n} \sum_{i=1}^n J_{ik} \quad (2)$$

where w_k is the weight for the k -th attribute, calculated as the average Jaccard similarity involving that attribute across all pairs of genes within the cluster. Meanwhile, n denotes the number of genes in the cluster, and J_{ik} signifies the element of the Jaccard matrix for the i -th gene and k -th attributes.

Definition 4: Weighted gene similarity calculation: Using the attribute weight w_k derived in Definition 3, a weighted similarity score can be computed as (2) for any pair of i and j genes within a cluster. The weighted similarity measure adjusts the raw Jaccard similarity by amplifying the contribution of attributes with higher weights:

$$\text{Weighted Similarity}(i, j) = \sum_{k=1}^m w_k \cdot J_{ik} \quad (3)$$

Where: w_k is the weight of the k -th attribute, J_{ik} is the Jaccard similarity matrix element for gene i and attribute k ; m is the total number of attributes.

This step ensures that attributes with greater biological significance exert a stronger influence on the clustering process.

3.2.3 Phase 3: Aggregation of Information

Building on this, the algorithm incorporated a hybrid similarity measure (3) to integrate multiple aspects of similarity, including structural, attribute, and neighbour-based metrics:

$$\begin{aligned} \text{Hybrid Similarity} = & \alpha \times \text{Structural Similarity} + \\ & (1 - \alpha) \times (\beta \times \text{Attribute Similarity} + \\ & (1 - \beta) \times \text{Neighbour Similarity}) \end{aligned} \quad (3)$$

Structural similarity is captured by the adjacency matrix A while attribute similarity is represented by matrix W , and shared neighbour similarity by SNsim . The balance among these is controlled by parameters α and β , with $\alpha \geq 0.5$ ensuring structural dominance. Jaccard similarity was integrated with these metrics to form biologically meaningful gene clusters. This method enhances network analysis accuracy and interpretability. Modularity, as defined in (4), was used to assess community quality by measuring intra-community density versus inter-community separation, thus validating the cohesiveness of detected structures. Using the following (4), modularity can be defined in formal terms, placing this evaluation in a mathematical framework:

$$Q = \frac{1}{2m} \sum \left[A_{ij} - \frac{K_i K_j}{2m} \right] \delta(C_i, C_j) \quad (4)$$

Where:

- A_{ij} is entry into an adjacency matrix if the edge between the genes is i and j , otherwise valued as 0.
- K_i and K_j represent the value of genes i and j , respectively. This refers to the number of edges that connect each gene.
- m is the total number of edges in the network.
- $\delta(C_i, C_j)$ represents the delta, equal to 1 when genes i and j are in the same module but otherwise equal to 0.

The Modularity formula evaluates the strength of communities by comparing actual and expected edge distributions within them – higher values indicate well-defined structures. This paper enhances modularity by optimizing weights α and β using network feature analysis rather than fixed values. Key features include the In-Degree Z-Score (which identifies influential nodes), PageRank (which assesses node importance via connectivity), and Betweenness Centrality (which highlights bridging nodes). This tuning aligns modularity with structural

and functional network properties, thereby enabling more precise detection of biologically and structurally significant communities.

3.2.4 Phase 4: Predicted Communities in Priors Network Analysis

Phase 4 consisted of constructing the hybrid similarity matrix, integrated with the Louvain algorithm to extract optimum community clusters. The predicted communities were compared with known biological databases to validate their biological relevance and interpret their genetic roles. In this study, the communities were grouped into three broad categories: (1) High Diversity, (2) High Potential, and (3) Moderate/Low (see Table 1). Meanwhile, some of the communities based on graph clusters and predicted in this study retained a considerable level of complexity and the potential for important biological activities, whereas others were expected to possess a wide range of biological roles and adaptabilities. Therefore, most of the communities were expected to have a less flexible or more specialized genetic configuration, which could be concentrated in a small number of vital biological processes.

3.3 Principal Component Analysis Feature

Principal Component Analysis (PCA) was employed to reduce dimensionality while preserving key transcriptional signals [19]. In this study, PCA loadings were integrated with gene community information within a single interaction network to explore intra-community gene relationships and their biological implications. The normalized, log-transformed expression matrix was analyzed across all samples, retaining the top 10 principal components to capture maximum variance. To correct for batch effects, the Harmony algorithm was applied to the PCs [20], ensuring the preservation of true biological signals. Focusing on the components with the highest contributions enhanced interpretability and supported more accurate modeling of gene relationships, ultimately improving feature matrices and biological prediction outcomes.

3.4 Biological Information Aggregation Using Multi-feature Reconstruction

Multi-feature reconstruction integrates outputs from PCA, community predictions, and gene expression

into a unified feature matrix. This matrix captures key biological patterns by combining variance in gene expression, structural relationships, and community affiliation. Each sample is represented by a row, with columns encoding the top 10 principal components, expression values, and binary indicators of community membership. This integrative approach enhances the representation of latent factors influencing gene expression and improves predictive accuracy. By unifying multidimensional data, the matrix serves as a robust framework for identifying novel functional activities, structural links, and regulatory interactions – insights that would be missed using isolated data sources.

4 RESULT DISCUSSION

This study utilized gene expression microarray data from *E. coli*, sourced from the DREAM5 Challenge [21], comprising 4511 genes, 805 samples, and 2066 experimentally validated regulatory interactions. The network was constructed based on the RegulonDB database, which includes only interactions classified as strong evidence [22], [23]. Given *E. coli*'s status as a well-studied model organism, it serves as a reliable foundation for regulatory network analysis. In this context, validated interactions were treated as true positives, while others were considered negatives. The datasets used included *E.coli_chip_features*, *E.coli_expression_data*, and the gold standard network.

For a clearer understanding of the data patterns, a smooth density curve is included with the histogram (see Figure 3), showing the distribution of average hybrid similarity among the clusters. Eight clusters are centred in the first noticeable peak, which is located at an average hybrid similarity of around 0.1. Another cluster grouping with eight clusters is indicated by the second peak, amounting to around 0.4. Based on the hybrid similarity metric, the distribution shows diverse levels of similarity in the nodes within clusters, ranging from 0.05 to 0.6. However, clusters close to 0.1 may have more varied features or weaker relationships.

The Louvain algorithm provided stability in the partitions. Section 3.3 lists a set of criteria besides community size that can be used to determine the likelihood of undiscovered genes and transcription factors. These criteria would help gain a better understanding of the regulatory network. Overall, the interactions between the gene communities were likely to be complex, involving shared regulators,

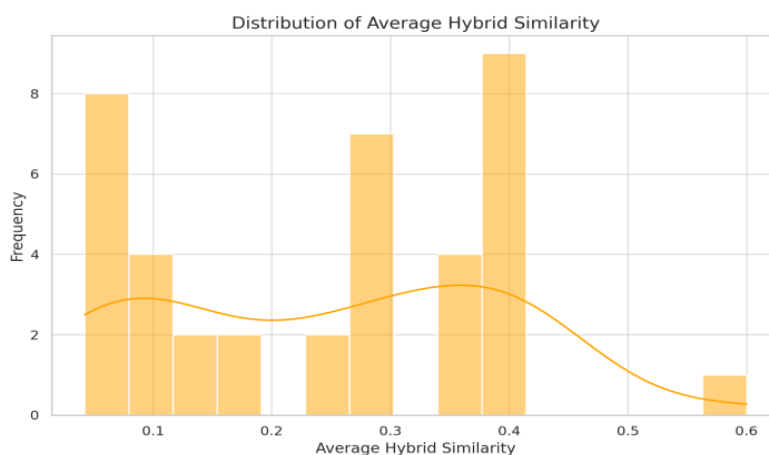


Figure 3: Distribution of hybrid similarity.

Table 1: Performance comparison of the proposed method with existing approaches.

Studies	Algorithm Used	Dataset	Modularity Score	No. of Predicted Communities
Choobdar et al. (2019)	Random Walk Clustering	DREAM (multiple species)	0.45~	60
Proposed method	Weighted Jaccard Similarity Louvain Clustering	DREAM (E.coli)	0.6488	38

metabolic pathway connections, stress response mechanisms, regulatory cascades, and transport/signalling processes. However, the evaluation of predicted modules is challenging because there is no ground truth of 'correct' modules in molecular networks. Therefore, in this study, a framework was introduced to empirically assess modules based on their association with complex traits using STRING¹. The updated STRING version was evaluated and the prediction of *E. coli* genes in the communities of 1081 was assessed, in order to further ascertain the accuracy of community predictions using the current researcher's methods. All of the communities added to the most recent version of STRING had a true percentage of > 90%, with the exception of two communities that had < 90%, namely, Communities 1 and 9, which had respective Jaccard index values of 0.52 and 0.75. Average precision for the total number of given communities was approximately 0.97.

The modularity score, calculated as 0.6488 using (4), was used to assess community detection quality. Compared to the 0.45 score reported by Choobdar et al. [20], this reflects a 44% improvement, indicating enhanced clustering and network organization. The results demonstrate the method's effectiveness in identifying biologically meaningful structures and improving the interpretation of gene interactions.

Table 1 demonstrates the performance of this method and the method proposed in [24] to predict communities across DREAM datasets. As this method has significantly outperformed other competing methods, it indicates the formativeness of topological properties and content in predicting communities.

A set of genes was analyzed within specific communities. These genes were selected based on their highest PC variance of average gene expression across samples. Thus, the goal of the current approach was achieved by integrating biological information from different networks to enable more accurate prediction of new gene connections and to support a deeper understanding of complex biological systems. These communities prioritize novel candidate genes; reveal pathway-level rate was then applied to the genes with the highest variance for each component, to identify important genes. In Figure 4, horizontal axis reflects the amount of variance of each gene, showing how gene expression varied across the PCs. The vertical axis represents the mean expression of each gene for all samples. Each dot refers to a gene with the highest PC variance indicated. The orange dots denote the position of the genes, whereas the text accompanying each dot (in blue) presents the names of the corresponding genes.

¹ <https://string-db.org/>

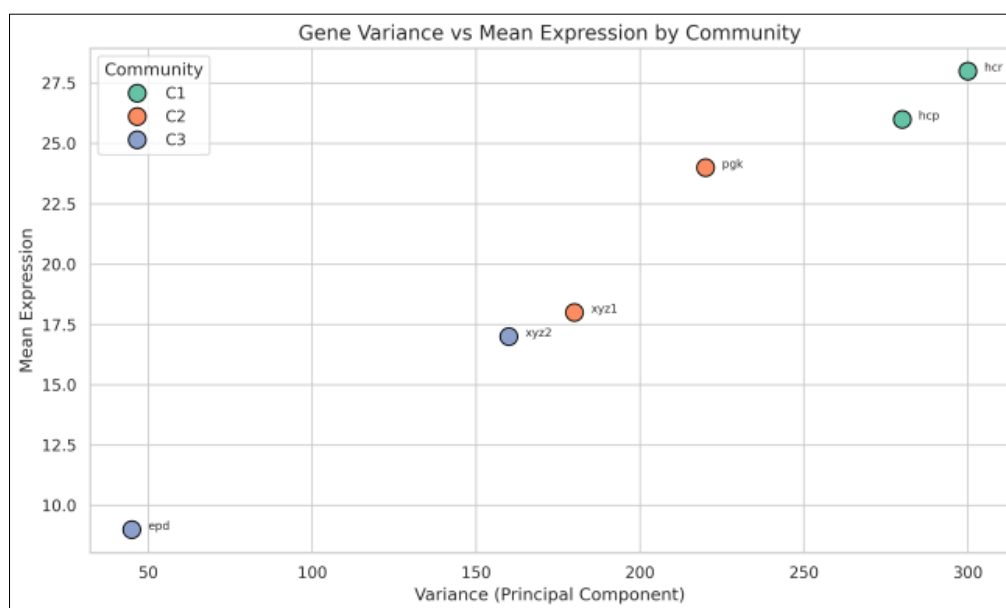


Figure 4: Relationships between variance and mean.

Table 2: Top genes selected by graph clustering dataset.

Genes		Biological Process	P-value
Community_11	ada, hns, lrp, tdh, ...	Metabolic & Regulatory Networks	0.0283
Community_12	epd, aer, crr, crp,...	Cellular Regulation & Adaptation	0.00554
Community_25	fnr, hmp, hcp,...	Stress Response & Detoxification	3.47×10^{-7}
Community_2	hpt, acs, gcd, ...	Energy Homeostasis & Metabolism	0.157
Community_16	lpd, icd, ssb	DNA Interaction & Metabolism	0.21
Community_23	uof, rob, slp, ...	Environmental & Metabolic Adaptation	0.132

The genes to the upper right of the graphic (for example, hcr and hcp) exhibited high variance and high mean expression, indicating their potentially important role in a specific biological process. In contrast, the genes illustrated lower left (for example, epd) displayed low variance and low mean expression, and may therefore be less important in this context. The relationship between high variance and high mean expression can be interpreted as representing genes that are dynamically active and associated with a response or function.

Focusing on the results in Table 2, which correspond to the integration of expression data with topological properties, it may be seen that the method of integrating the expression data played an essential role in performance. Meanwhile, the p-values in the various groups denote the levels of statistical significance and stand to offer new insights into

functional integration among the genes in each group. By way of illustration, Community_25 was topped with genes such as fnr, hmp, hcr, and hcp and had a p-value of 3.47×10^{-7} , exhibiting highly significant interactions and integrations as physiological responses to oxidative conditions and nitrosative stress. Here, some communities such as 29 and 0 had p-values at 1, reflecting comparatively lower significance in their interactions. In contrast, high p-values in groups such as Community_29 and Community_0 could be the starting point for further, more detailed investigation into new relations and interactions, which could be relevant to deeper insights into biological networks.

Moreover, advanced bioinformatics tools and techniques may give way to new horizons opening up for possible discoveries in molecular biology and genomics. Conversely, Community_11 is composed

of *ada*, *hns*, *lrp*, *tdh*, and *kbl* genes with a p-value of 0.0283, thereby evidencing significant functional cooperation in terms of the impact on cellular regulatory networks and metabolic pathways. This difference in p-values enabled highly interactive groups to be identified, namely, high functional integration, which suggests the important biological functions of these genes in the face of environmental and physio-logical change. The identified genes were associated with critical biological functions such as gene regulation, cell signalling, and basic biological processes, consequently enhancing their biological significance.

5 CONCLUSIONS

This paper presents a comprehensive framework that leverages multiple types of features and advanced graph clustering techniques to predict gene interactions with notable precision. By integrating information from diverse biological data sources, the approach not only increased the accuracy of link prediction but also allowed for a nuanced exploration of both dynamic and static aspects within gene networks. The two-step clustering strategy proved effective in disentangling these facets, ultimately shedding light on key gene relationships that might otherwise remain obscured by traditional analysis methods. Importantly, the findings highlight the value of combining different data perspectives – such as gene expression profiles, topological network properties, and community structures – to gain a more holistic understanding of genetic connectivity. The resulting framework has enabled a more detailed mapping of gene regulatory mechanisms and offered new insights into the fundamental organization of cellular processes. This work not only enriches our understanding of gene networks but also establishes a methodological foundation for future studies aiming to interpret complex molecular interactions.

Moreover, the methodology has practical implications beyond basic research. Its ability to reliably predict novel gene interactions positions it as a valuable tool for applications in fields such as precision medicine and drug discovery, where understanding the intricacies of gene regulation is critical. The structured, data-driven approach laid out in this research provides a reproducible pathway for future exploration and validation in both experimental and clinical settings. Ultimately, this study contributes meaningfully to the ongoing effort to decode the complexity of biological systems through computational innovation.

REFERENCES

- [1] D. Marbach et al., "Perturbations Across Complex Diseases," *Genome Biology*, vol. 18, no. 1, p. 236, 2016, [Online]. Available: <https://doi.org/10.1038/nmeth.3799>.
- [2] M. J. Bonder, R. Luijk, D. V. Zhernakova, and M. Moed, "Disease variants alter transcription factor levels and methylation of their binding sites," 2015.
- [3] S. S. Ahmed, S. Roy, and J. Kalita, "Assessing the Effectiveness of Causality Inference Methods for Gene Regulatory Networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 17, no. 1, pp. 56–70, 2020, [Online]. Available: <https://doi.org/10.1109/TCBB.2018.2853728>.
- [4] B. A. Logsdon and J. Mezey, "Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations," *PLoS Computational Biology*, vol. 6, no. 12, 2010, [Online]. Available: <https://doi.org/10.1371/journal.pcbi.1001014>.
- [5] S. Z. AlRashid, M. H. Dosh, and A. J. Obaid, "Classification of the Senescence-Accelerated Mouse (SAM) Strains With Its Behaviour Using Deep Learning," *International Journal of e-Collaboration*, vol. 18, no. 2, pp. 1–13, 2022, [Online]. Available: <https://doi.org/10.4018/IJeC.304035>.
- [6] F. Wagner, "GO-PCA: An unsupervised method to explore gene expression data using prior knowledge," *PLoS One*, vol. 10, no. 11, pp. 1–26, 2015, [Online]. Available: <https://doi.org/10.1371/journal.pone.0143196>.
- [7] J. X. Liu, C. M. Feng, X. Z. Kong, and Y. Xu, "Dual Graph-Laplacian PCA: A Closed-Form Solution for Bi-Clustering to Find 'Checkerboard' Structures on Gene Expression Data," *IEEE Access*, vol. 7, pp. 151329–151338, 2019, [Online]. Available: <https://doi.org/10.1109/ACCESS.2019.2941227>.
- [8] S. R. Datasets, J. M. Zhang, J. Fan, H. C. Fan, D. Rosenfeld, and D. N. Tse, "An Interpretable Framework for Clustering," *bioRxiv*, pp. 1–15, 2017.
- [9] T. Liu and C. Jia, "scDFN: enhancing single-cell RNA-seq clustering with," vol. 25, no. 6, 2024.
- [10] B. Yu et al., "ScGMAI: A Gaussian mixture model for clustering single-cell RNA-Seq data based on deep autoencoder," *Briefings in Bioinformatics*, vol. 22, no. 4, pp. 1–10, 2021, [Online]. Available: <https://doi.org/10.1093/bib/bbaa316>.
- [11] M. D. Luecken et al., "Benchmarking atlas-level data integration in single-cell genomics," *Nature Methods*, vol. 19, no. 1, pp. 41–50, 2022, [Online]. Available: <https://doi.org/10.1038/s41592-021-01336-8>.
- [12] N. A. A. Shanan, H. A. Lafta, and S. Z. Al Rashid, "Using alignment-free methods as preprocessing stage to classification whole genomes," *International Journal of Nonlinear Analysis and Applications*, vol. 12, no. 2, pp. 1531–1539, 2021, [Online]. Available: <https://doi.org/10.22075/ijnaa.2021.5281>.
- [13] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1–44, 2016, [Online]. Available: <https://doi.org/10.1016/j.physrep.2016.09.002>.

- [14] S. Noori, N. Al-A'araji, and E. Al-Shamery, "Construction of dynamic protein interaction network based on gene expression data and quartile one principle," *Proteins: Structure, Function, and Bioinformatics*, vol. 90, no. 5, pp. 1219–1228, 2022, [Online]. Available: <https://doi.org/10.1002/prot.26304>.
- [15] S. M. Hill et al., "Inferring causal molecular networks: Empirical assessment through a community-based effort," *Nature Methods*, vol. 13, no. 4, pp. 310–322, 2016, [Online]. Available: <https://doi.org/10.1038/nmeth.3773>.
- [16] C. G. Urzúa-Traslaviña et al., "Improving gene function predictions using independent transcriptional components," *Nature Communications*, vol. 12, no. 1, 2021, [Online]. Available: <https://doi.org/10.1038/s41467-021-21671-w>.
- [17] A. Bhih, P. Johnson, and M. Randles, "An optimisation tool for robust community detection algorithms using content and topology information," *The Journal of Supercomputing*, vol. 76, no. 1, pp. 226–254, 2020, [Online]. Available: <https://doi.org/10.1007/s11227-019-03018-x>.
- [18] A. Bhih, P. Johnson, T. Nguyen, and M. Randles, "Decentralized iterative community clustering approach (DICCAs)," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, vol. 2017-Octob, pp. 1–7, 2017, [Online]. Available: <https://doi.org/10.1109/PIMRC.2017.8292677>.
- [19] N. Alrefaai and S. Z. Alrashid, "Classification of gene expression dataset for type 1 diabetes using machine learning methods," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 5, pp. 2986–2992, 2023, [Online]. Available: <https://doi.org/10.11591/eei.v12i5.4322>.
- [20] I. Korsunsky et al., "Harmony 2," *Nature Methods*, vol. 16, no. 12, pp. 1289–1296, 2019, [Online]. Available: <https://doi.org/10.1038/s41592-019-0619-0>.
- [21] D. Marbach et al., "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, no. 8, pp. 796–804, 2012, [Online]. Available: <https://doi.org/10.1038/nmeth.2016>.
- [22] H. Salgado, A. Santos, U. Garza-Ramos, J. Van Helden, E. Díaz, and J. Collado-Vides, "RegulonDB (version 2.0): A database on transcriptional regulation in *Escherichia coli*," *Nucleic Acids Research*, vol. 27, no. 1, pp. 59–60, 1999, [Online]. Available: <https://doi.org/10.1093/nar/27.1.59>.
- [23] G. J. Kang, S. R. Ewing-Nelson, L. Mackey, J. T. Schlitt, A. Marathe, and K. M. Abbas, "Neonatal Rat Ventricular Myocyte Isolation: HHS Public Access," *Physiology & Behavior*, vol. 176, no. 1, pp. 139–148, 2018. <https://doi.org/10.1002/cpbi.43>.
- [24] S. Choobdar et al., "Assessment of network module identification across complex diseases," *Nature Methods*, vol. 16, no. 9, pp. 843–852, 2019, [Online]. Available: <https://doi.org/10.1038/s41592-019-0509-5>.

Using the Rooted Tree to Find the Optimal Solution for the Transportation Problems

Athraa Abdul Ghani and Mohammed Shakir Mahdi

*College of Education for Women, Department of Mathematics, University of Kufa, 54003 Najaf, Iraq
athraaa.alsani@student.uokufa.edu.iq, mohammedsh.mahdi@uokufa.edu.iq*

Keywords: Transportation Problems, Optimal Solution, Graph Theory.

Abstract: Considering economic aspects of the transportation problem and the fact that it is a particular instance of linear programming problems that rely on figuring out the best way to distribute goods from supply centers to customers in a shorter amount of time or at a lower cost, as well as the growing importance of the globalization and rapid development era, resolving the transportation problem has become crucial in the field of operations research to enhance ways of bringing the product in the quickest amount of time or at the lowest cost from the original source to the customer. By developing a new algorithm to address transportation issues, including logistical issues for the supply of goods and their arrival at their destination, we offer a simplified, guaranteed, and less expensive approach than earlier approaches for businesses to deal with transportation problems, thereby improving efficiency, reducing operational risks, and significantly increasing overall competitive advantage for companies.

1 INTRODUCTION

Transportation problems (TP) of the repeatedly used applications of programming with linear techniques is for mutating and solvin transportation problems as a Lpp. in this problem, the main target is obtain an optimal schedule for shipping the commodity while satisfy in the demands in every destination, French mathematician Gaspard Monge in 1781, proposed a mathematical model that in volves sealing and transporting soil between different locations at the lowest possible cost. This achieved through aheavy cooperation with Napoleon Bonaparte's army that worked to build forts and military ways. The American mathematician, Frank L. Hitchcock, proposed the TP in 1941 [1], and A.N. Tolstoi [2] published his Monge's problem solution which represents atheoretical foundation for solving TP. in1947, Tjalling C. Koopmans, offered his paper to solve TP [3]. the tow mentioned studies are the main achievements in the advancement of the different techniques for solving the transportation model. also, in 1949 T.C. Koopmans developed the transportation model and G. B. Dantzing towards the for mulation and solution of linear programming problem in 1951 [4] with using simplex method for getting optimal solution. Bazarra, Jarvis and Sherali defined

the linear programming problems in 1990 [5]. However, it includes aconsiderable volume of variables and restrictions, and solve them by simplex technique requires of effort and a protracted time. Abdul Staar Soomro et.al. suggested an adjustment vogel's approximate method for solving transportation problem in 2015 [6]. Many researchers have presented other methods to get (IBFS) to the transportation problems. Lakh veer Kaur and others discussed now to improve the method of maximum difference for finding (IBFS) for the TP in 2018 [7]. In arelated context, S.C.Zelibe and C.P.Ugwuanyi developed novel remedies for the problem of transportation, 2019 [8]. In 2020 Al-Saeedi 1 proposed a technique for finding objective solutions for transportation types. Subsequently, Noor proposed a technique for finding the optimal or near-optimal solution [9]. In this work we present a new approach to find the optimal solution for the transportation problems which minimize is cost by using an ordered tree, where a tree is an directed graph in which any two vertices are connected by exactly one path, or equivalently a connected acyclic directed graph [10], so we sought to provide the best results that showed the efficiency of the new method by comparing their solution to the old ones (NWCM, LCM and VAM) and the method Saeedi 1. We used

the Matlab 2013 program to calculate the optimal solution; it is given to the optimal solution.

2 SOME BASIC CONCEPTS OF TRANSPORTATION PROBLEMS TP AND GRAPH THEORY (GT)

2.1 Transportation Model

In TP, there exist m origins of supply $S_1, S_2, \dots, S_m$ and n destinations of demand $D_1, D_2, \dots, D_n$, each one is specially represented. The brackets symbolize the tracks connecting origins and destinations. Bracket (i, j) connection from origin i ($i = 1, 2, \dots, m$) into destination j ($j = 1, 2, 3, \dots, n$), where C_{ij} represents the transportation cost for each unit of product, while X_{ij} is the charged quantity. The available units at supply in origin i are represented by a_i , while the available units at request in destination j are represented by b_j .

$$\text{Minimize } Z = \sum_{i=1}^m \sum_{j=1}^n C_{ij} X_{ij} \text{ (gross cost)} \quad (1)$$

Subject to

$$\sum_{j=1}^n X_{ij} = a_i, \quad i = 1, 2, \dots, m \text{ (supply restrictions)} \quad (2)$$

$$\sum_{i=1}^m X_{ij} = b_j, \quad j = 1, 2, \dots, n \text{ (demand restrictions)} \quad (3)$$

2.1.1 Representation of Transportation Problems

TP represent a particular model of linear programming problems (LPP). The transportation problem (Fig. 1) is shipping different quantities of homogeneous goods from various sources (e.g., factories) to various destinations (e.g., warehouses) so that the total transportation cost (or time) is minimized.

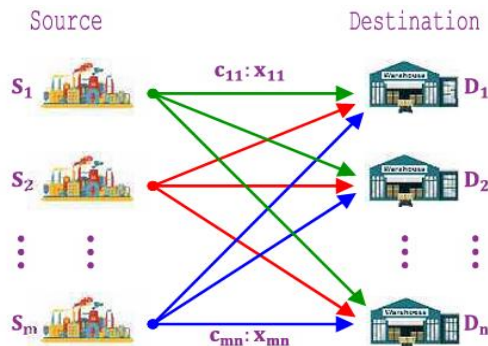


Figure 1: The graph structure of TP.

2.1.2 Schedule of Transportation

The TP is represented in a distinctive table (Fig. 2) that employs the mathematical model for the TP linear programming. A tabular format can be used to represent the transportation problem model with all related parameters. The supply (a_i Availability), visualized in the transportation schedule, is shown in every source in the right column, while the destination requirement (b_j) is found in every cell in the bottom row. The cost of oneness transfer (c_{ij}) is located in the top corner of the table, right Side, while the shipped matter quantity (x_{ij}) is represented in the cell center [11].

Sources (i)	Destinations (j)				Supply (a_i)
	D_1	D_2	$\dots$	D_n	
S_1	c_{11} x_{11}	c_{12} x_{12}	$\dots$	c_{1n} x_{1n}	a_1
S_2	c_{21} x_{21}	c_{22} x_{22}	$\dots$	c_{2n} x_{2n}	a_2
$\vdots$	$\vdots$	$\vdots$	c_{ij} x_{ij}	$\vdots$	$\vdots$
S_m	c_{m1} x_{m1}	c_{m2} x_{m2}	$\dots$	c_{mn} x_{mn}	a_m
Demand (b_j)	b_1	b_2	$\dots$	b_n	$\sum_{i=1}^m a_i = \sum_{j=1}^n b_j$

Figure 2: Schedule of TP.

2.1.3 Definition

2.1.3.1 Feasible Solution (FS)

A feasible solution to transportation problems is a collection of non-negative assigned quantities $X_{ij} \geq 0$ that cater for restrictions (supply constraints and demand restrictions in transportation problems) [12].

2.1.3.2 Basic Feasible Solution (BFS)

We say that a feasible solution of transportation problems is a basic feasible solution (BFS) when it includes no more than $(m \cdot n - 1)$ positive assigned quantities, where m stands for the number of rows and n stands for the number of columns of TP [12], [13].

3 GENERAL PROCEDURE TO SOLVE A TRANSPORTATION PROBLEM

The solution of a TP typically follows a systematic procedure designed to ensure both feasibility and optimality. The general approach consists of the following main steps:

- 1) Mathematical formulation of the table of TP.
- 2) Find an initial Basic Feasible Solution (BFS).
- 3) Modified the initial basic feasible solution (IBFS) obtained in Phase 2 to find the optimal solution.

3.1 Types of Transportation Problems

3.1.1 Balanced TP

The TP is called balanced when the quantum of products in the apportionment pivot is equivalent to the quantum of products desired by the request pivot i.e., $\sum_{i=1}^m a_i = \sum_{j=1}^n b_j$.

3.1.2 Unbalanced TP

The TP is called unbalanced when the quantum of products ready in the apportionment pivot is not equivalent to the quantum of products desired by the request pivot i.e. $\sum_{i=1}^m a_i \neq \sum_{j=1}^n b_j$.

3.2 Methods for Finding IBFS of TP

The initial solution that we obtain from the traditional solution approaches or the new solution approaches must contain the following properties:

- It should be feasible.
- It must satisfy the non-negativity.
- It should be basic.
- Some of the classic methods used to get the first solution for TP are: North-West Corner Method (NWCM).
- Least Cost Method (LCM).
- Vogel's Approximation Method (VAM).

3.3 Methods of Finding Optimal Solution

The moment of determining the initial basic feasible solution will lead to performing the optimality test in an effort to see whether the obtained feasible solution is optimal or not. This test can be conducted through applying the following methods:

- 1) Stepping Stone Method (SSM).
- 2) Modified Distribution (MODI) Method.

3.4 Graph Theory

Graph theory is a branch of discrete mathematics that studies mathematical structures known as graphs. A graph is formally defined as an ordered pair

$G = (V, E)$ where V is the set of vertices, and E is the set of edges connecting pairs of vertices. Graph theory models binary relationships and solves optimization problems involving networks and distributional problems [10].

3.4.1 Tree

A tree is a connected, cycle-free undirected graph, meaning that there is a single path between any pair of nodes. Thus, if a tree $G = (V, E)$ contains n nodes, then the number of edges $|E|$ is $n - 1$. Trees represent hierarchical structures, making them an essential tool in many optimization and algorithmic analysis applications [14].

3.4.2 Rooted Tree

A rooted tree is one in which one node is identified as the root, forcing the other nodes into a hierarchical order. Every node, with the exception of the root, has a distinct path to the root. A node's "children" are the nodes that come right after it, while its "parent" is the node that comes before it. Numerous methods and mathematical models used in optimization are based on this hierarchical structure [15].

3.4.3 Tree in Graph

In graph theory, a "tree in a graph" is defined as a subgraph consisting of a set of vertices and edges of the original graph, which is connected and acyclic. In other words, if the graph $G = (V, E)$ and a tree $T = (V', E')$ such that $V' \subseteq V$ and $E' \subseteq E$, then T is a tree in G if the following conditions are met [10], [16]:

- 1) Connected. A path exists between any pair of vertices in T .
- 2) Cycle-free. No route can begin and finish at the same vertex without repeating edges.

When T includes all vertices V in the original graph, this subgraph is called a universal tree.

3.4.4 Definitions

Definition 1. An m -ary tree ($m \geq 2$) is a rooted tree in which every vertex has m or fewer children [15] as in Figure 3.

Definition 2. A complete m -ary tree is an m -ary tree in which every internal vertex has exactly m children and all leaves have the same depth, as shown in Figure 3 [15].

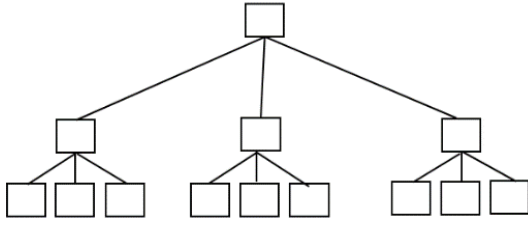


Figure 3: Complete 3-ary tree.

4 THE PROPOSED ALGORITHM

We will discuss the algorithm we proposed to find the optimal solution (OS) directly. This algorithm replaces the existing IBFS and OB in sub-sections 3.2 and 3.3 by relying on the rooted tree.

4.1 Algorithm Overview

This subsection describes the overall concept and purpose of the proposed method. The algorithm aims to optimize transportation costs by constructing a rooted tree instead of relying on traditional methods such as IBFS (Initial Basic Feasible Solution) and OB (Optimality by MODI or SSM). The rooted tree mechanism allows systematic allocation of supply and demand across the cost matrix, forming a basis for cost minimization.

4.2 Algorithm Procedure

The proposed algorithm follows a systematic approach to solve the transportation problem through the following steps:

Step 1. We start with the cell located at the North West corner, which takes its share from the allocation between supply and demand.

Step 2. From this cell, we begin forming a tree extension, taking the cells in the second row, with each cell taking its share between supply and demand, starting from the first cell.

Step 3. We take the branches of the cells that branched from the cell in the second row and continue the tree extension for the remaining cells until the last cell in the next row.

Step 4. We repeat the allocation for the cells so that each cell takes its share between supply and demand until the tree branching is complete.

Step 5. Repeat the above steps for each cell of the cost matrix and calculate the cost of the transportation problem.

Step 6. We choose the lowest cost for the transportation problem.

Note. We get the lowest cost from step 6 of the proposed algorithm, which matches the optimal solution (OS) after comparing it with SSM and MODI.

5 NUMERICAL EXAMPLES

We will provide illustrative examples of the proposed algorithm. The input data for the transportation problem are summarized in Figure 4.

Source	Destination			Supply
	D_1	D_2	D_3	
S_1	1	2	2	15
S_2	6	4	5	35
S_3	3	2	9	25
Demand	23	33	19	75

Figure 4: The transportation problems.

The following numerical example demonstrates the application of our algorithm to this dataset:

- 1) In Figure 5, we start from the first row and allocate the cell (northwest) until we complete allocating the elements of the first row.

Based on the first cell, we calculate the cost as

$$Z = \sum_{i=1}^3 \sum_{j=1}^3 x_{ij} c_{ij}.$$

$$Z = (15*1) + (8*6) + (27*4) + (6*2) + (19*9) = 354 \text{ units.}$$

- 2) We repeat the algorithm for the first row by allocating the second cell, as in the following Figure 6.

Based on the second cell, we calculate the cost as

$$Z = \sum_{i=1}^3 \sum_{j=1}^3 x_{ij} c_{ij}.$$

$$Z = (15*2) + (23*6) + (12*4) + (6*2) + (19*9) = 399 \text{ units.}$$

- 3) We repeat the algorithm for the first row by allocating the third cell, as shown in Figure 7.

Based on the third cell, we calculate the cost as

$$Z = \sum_{i=1}^3 \sum_{j=1}^3 x_{ij} c_{ij}.$$

$$Z = (15*2) + (23*6) + (12*4) + (21*2) + (4*9) = 294 \text{ units.}$$

Now, we replaced the third row with the second row while maintaining the allocation for each column and row. By repeating the proposed algorithm on each cell, we will be able to calculate the lowest cost of the cost matrix that matches the optimal solution based on OS, as in the following Figures 8, 9, and 10.

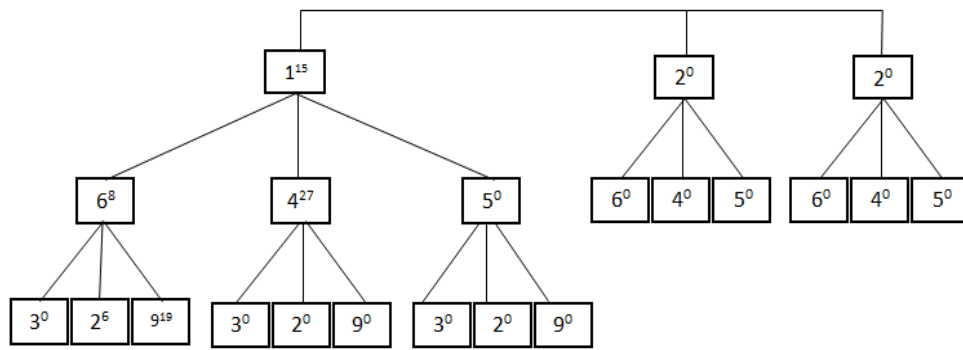


Figure 5: The tree of the first row rooted tree by c_{11} .

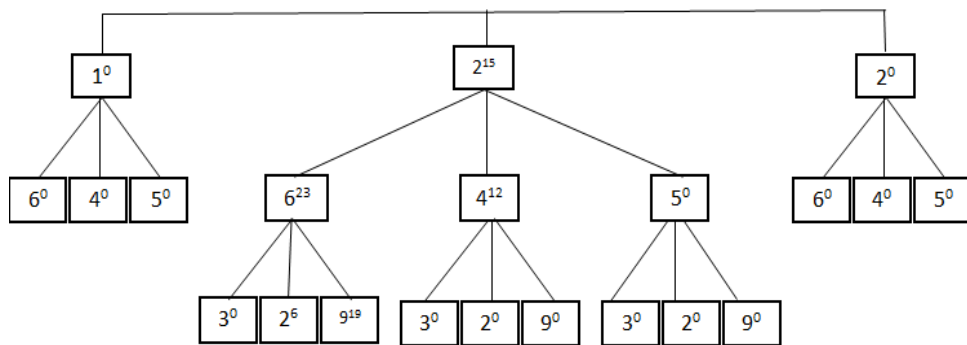


Figure 6: The tree of the first row rooted tree by c_{12} .

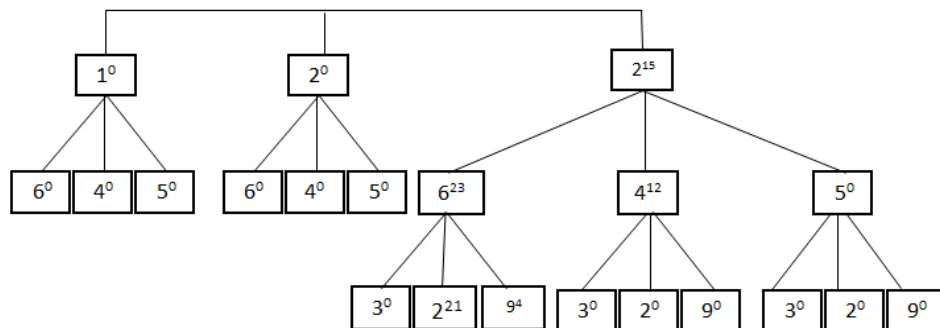


Figure 7: The tree of the first row rooted tree by c_{13} .

Source	Destination			Supply
	D_1	D_2	D_3	
s_1	1	2	2	15
s_3	3	2	9	25
s_2	6	4	5	35
Demand	23	33	19	75

Figure 8: The optimal solution based on OS.

We repeat the application of the algorithm on each element of the cost matrix, and using the MATLAB 2013 program, we obtain the optimal solution as in Figure 11.

Then, $Z = (15*1) + (8*3) + (17*2) + (16*4) + (19*5) = 232$ units, which is the optimal solution.

Source	Destination				supply
	D_1	D_2	D_3	D_4	
S_1	1	2	3	4	30
S_2	7	6	2	5	50
S_3	4	3	2	7	35
Demand	15	30	25	45	115

Figure 9: The transportation problems.

Using the proposed algorithm, programmed in MATLAB 2013, we obtained a Rooted Tree, which gives us the optimal solution. We moved the fourth column to become the second while maintaining the allocation for each column and row. As a result, we obtained the following Figure 10.

Source	Destination				Supply
	D_1	D_2	D_3	D_4	
S_1	1	4	3	2	30
S_2	7	5	2	6	50
S_3	4	7	2	3	35
Demand	15	45	25	30	115

Figure 10: The lowest possible cost.

We repeat the application of the algorithm to each element of the cost matrix to obtain the lowest possible cost by using MATLAB 2013, we get the lowest possible cost from the given Figure 12.

Then, $Z = (15 * 1) + (15 * 2) + (45 * 5) + (5 * 2) + (20 * 2) + (15 * 3) = 365$ units.

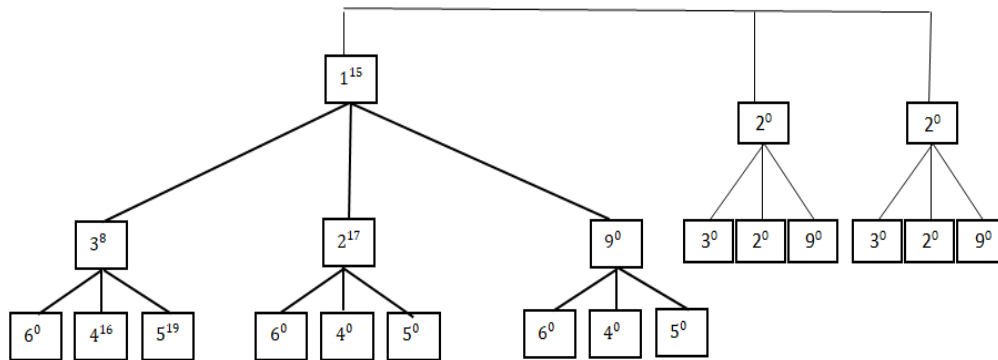
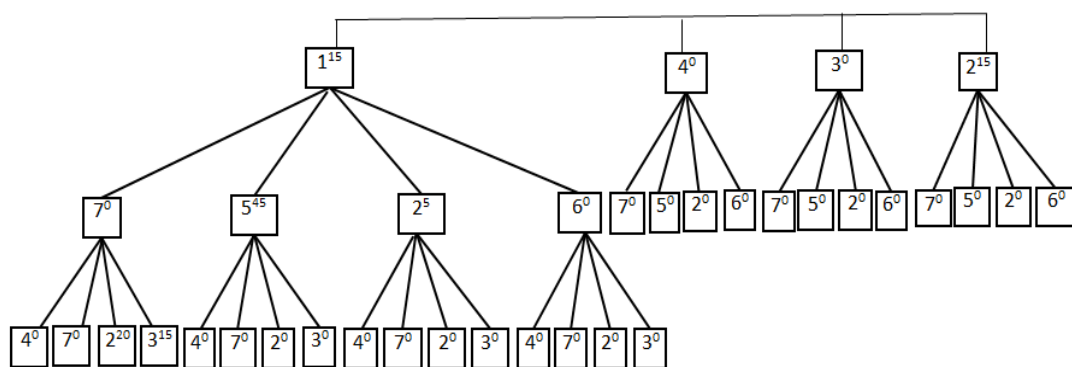

 Figure 11: The tree of the first row rooted tree by c_{11} .


Figure 12: The tree of the first row rooted tree.

5 COMPARISON

The outcomes of the instance solving conducted in the current study were compared to prove the workability of the new technique. The comparison has been conducted between the classic methods of solution (NWCM, LCM, VAM and method Al-Saeedi1). A superior IBFS is produced by the proposed technique, rather than the traditional algorithms. This algorithm has close or ideal positions to optimize the IBFS that is obtained by the current suggested algorithm. This algorithm is identified as superior to those acquired by employing Vogel's Approximation approach. The following Figure 13 shows the difference between methods to solve TP in the cost function.

Name of method	NWCM	LCM	VAM	Saeedi1	The proposed algorithm	OS
Example1	354	240	247	247	232	232
Example2	480	405	375	405	365	365

Figure 13: The difference between the methods for solving TP in the cost function.

6 CONCLUSIONS

The results from using the tree approach in this work are found to be better than the results obtained with the three traditional methods and the Saeedi method 1. Given what was shown earlier, it is clear that the new algorithm provides satisfactory and preferred solutions. As a result, the TP was solved using an objective function that was designed to be minimized. The understanding and use of the new algorithm are both considered to be simple in many cases. It saved me both time and energy in discovering the optimal solution (OS). On the basis of the best answer, we achieve a wise selection. Consequently, our proposed approach gives the method's algorithm a potential improvement by enhancing computational efficiency, reducing overall complexity, and facilitating easier implementation. These advantages make it particularly suitable for practical applications in industries where rapid and

cost-effective solutions are essential for maintaining competitiveness.

REFERENCES

- [1] F. L. Hitchcock, "The distribution of a product from several sources to numerous localities," *Journal of Mathematics and Physics*, vol. 20, pp. 224–230, 1941.
- [2] H. A. H. J. Al-Saeedi and M. A. S. Al-Jenabi, "Solving transportation problems by using a new modification of Vogel's approximation method," pp. 30–40, 2021.
- [3] T. C. Koopmans, "Optimum utilization of the transportation system," *Econometrica*, vol. 17, pp. 136–146, 1949.
- [4] L. Almamalik, "A new method for finding an optimal solution transportation problems," in *1st Piksi International Conference on Knowledge and Sciences*, 2020, pp. 1–2.
- [5] R. G. Patel, B. S. Patel, and P. H. Bhathawala, "On optimal solution of a transportation problem," *Global Journal of Pure and Applied Mathematics*, vol. 13, no. 9, pp. 6201–6202, 2017.
- [6] M. S. M. Zabiba, H. A. H. Al-Dallal, K. H. Hashim, M. M. Mahdi, and M. A. K. Shiker, "A new technique to solve the maximization of the transportation problems," *Journal of Physics*, pp. 1–2, 2020.
- [7] L. Kaur, M. Rakshit, and S. Singh, "An improvement in the maximum difference method to find initial basic feasible solution for transportation problem," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 9, pp. 533–535, 2018.
- [8] S. C. Zelibe and C. P. Ugwuanyi, "On a new solution of the transportation problem," *Journal of the Nigerian Mathematical Society*, vol. 38, no. 2, pp. 271–291, 2019.
- [9] M. S. Zabiba and N. H. A. Alkhafaji, "Using a new method (NOOR 2) for finding the optimal solution of the transportation problem," *NeuroQuantology*, vol. 20, no. 4, pp. 518–521, 2022.
- [10] E. A. Bender and S. G. Williamson, *Lists, Decisions and Graphs. With an Introduction to Probability*, 2010.
- [11] U. K. Das, M. A. Babu, A. R. Khan, and M. S. Uddin, "Advanced Vogel's Approximation Method (AVAM): A new approach to determine penalty cost for better feasible solution of transportation problem," *International Journal of Engineering Research & Technology (IJERT)*, vol. 3, no. 1, pp. 182–187, 2014.
- [12] K. Dhurai and A. Karpagam, "To obtain initial basic feasible solution physical distribution problems," *Global Journal of Pure and Applied Mathematics*, vol. 13, no. 9, pp. 4671–4676, 2017.
- [13] M. Malireddy, "A new algorithm for initial basic feasible solution of transportation problem," *International Journal of Engineering Science Invention (IJESI)*, vol. 7, no. 8, pp. 41–43, 2018.
- [14] F. Harary and G. Prins, "The number of homeomorphically irreducible trees, and other species," *Acta Mathematica*, vol. 101, no. 1–2, pp. 141–162, 1959. [Online]. Available: <https://doi.org/10.1007/BF02559543>.

- [15] N. Deo, Graph Theory with Applications to Engineering and Computer Science. Englewood, NJ: Prentice-Hall, 1974.
- [16] R. Diestel, Graph Theory, 5th ed. Springer, 2017.

Cryptocurrency Price Prediction Model Using GRU, LSTM and Bi-LSTM Machine Learning Algorithms

Laila Suwaid Said

University of Al-Buraimi, BOX 77, 512 Al Buraimi, Sultanate of Oman

Laila.su@uob.edu.om

Keywords: Cryptocurrency Price Prediction, Deep Learning, GRU, LSTM, Time Series Forecasting.

Abstract: The rapid rise of cryptocurrencies has indeed created both investment opportunities and forecasting challenges. Accurate predictions of cryptocurrency prices are crucial for traders and financial planners to make informed decisions. Recent studies have employed deep learning techniques, specifically Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Bidirectional LSTMs (Bi-LSTM), to analyze the price movements of popular cryptocurrencies. In a comparative analysis, the GRU model consistently outperformed both LSTM and Bi-LSTM models across various cryptocurrencies, demonstrating its effectiveness in modeling financial data. The performance of these models was evaluated using the Mean Absolute Percentage Error (MAPE) after preprocessing, normalizing, training, and evaluating the data. The results indicated that GRU provided the most accurate predictions, with lower MAPE values compared to its counterparts. For instance, a study focusing on Bitcoin, Ethereum, and Litecoin found that the GRU model achieved superior performance metrics, confirming its suitability for cryptocurrency price forecasting. This suggests that traders and investors can rely on GRU-based models for more accurate predictions, ultimately aiding in better investment strategies in the volatile cryptocurrency market.

1 INTRODUCTION

A fiat currency is a currency that is divisible, transferable, durable, and scarce [1]. This system has some drawbacks, including no tangible backing for currencies and no ability to control the money supply, which may contribute to hyperinflation and income inequality [2]. Financial institutions and credit card companies facilitate many transactions, which increase costs, lengthen transfer times, and expose ledgers to manipulation. As a result, individuals cannot control or own data. As a result of government regulations and legal agreements, the current financial system still enjoys public trust regardless of these limitations. A breach of trust has resulted in significant financial losses in the past, such as during the dot-com bubble and the housing bubble of 2008 [3]. It has been proven that trust breaches can result in significant financial losses in the past, such as during the dot-com bubble and the 2008 real estate bubble. In October 2008, a group operating under the pseudonym [4] created the first digital currency, Bitcoin, using blockchain technology. International financial institutions are becoming increasingly interested in P2P transactions, which can be carried

out over the internet without intermediaries [5]. There is currently research being conducted by academia, government agencies, and media outlets as well as citizens.

Since its inception, the cryptocurrency market has been the subject of controversy and debate, but it has gradually expanded into one of the largest alternative investment venues [6]. Bitcoin [5] pioneered the use of cryptocurrencies as a means of investing, according to CoinMarketCap.com [7], [8]. In response to this remarkable growth, big data and virtual assets are enabling us to move into a new era of financial. Investors have faced new risks as a result of the new markets, but they have also been provided with more tools to manage them.

During the crypto boom of 2017, governments worldwide began standardizing and regulating digital currencies. More people now use Bitcoin more comfortably thanks to blockchain technology's safety [9]. In addition to using blockchain technology for security, it is also important to consider cryptocurrencies' legality when considering illegal users [10]. Consequently, there are ongoing discussions and examinations about cryptocurrency law. In addition to technological innovation and legal

frameworks, digital currencies also incorporate moral considerations. The authors examine cryptocurrency viewpoints, characteristics, and legal and economic issues in regard to monetary elements [11]. As a currency, Bitcoin does not meet the economic requirements in accordance with traditional views and features. Among the many cryptocurrencies available on the digital market, BTC is well known. Considering cryptocurrencies' interconnectedness, small cryptocurrencies can have a negative impact on larger cryptocurrencies. According to research [12], gold acts as a stand-alone currency, making it an appropriate asset to hedge cryptocurrency price fluctuations. With virtual currencies evolving and becoming more volatile, conventional assets, such as gold, may serve as stabilizing elements, reducing risk and diversifying portfolios [13].

Cryptocurrencies are inherently volatile and dynamic, making price forecasting difficult. Researchers in this field explore and study the complexities of cryptocurrencies to provide a better understanding of what drives price fluctuations in this market. We will use machine learning and deep learning algorithms to identify latent patterns in data to improve prediction accuracy. An analysis of deep learning models is presented in this article for predicting Bitcoin, Ethereum, and Litecoin prices. In order to forecast cryptocurrency prices, LSTMs and gated recurrent units (GRUs) are both types of recurrent neural networks. It is demonstrated that DL algorithms are capable of increasing prediction accuracy and handling the inherent nonlinearities of time series data in this study.

2 LITERATURE REVIEW

In artificial intelligence, machine learning predicts the future based on past data. The use of historical cryptocurrency price data to train machine-learning models may allow us to predict the price movement of cryptocurrencies in the future. In contrast with traditional forecasting models, machine learning-based forecasting is able to provide results that are close to or identical to the actual outcome while also improving accuracy [14]. There has been an increase in the use of neural networks (NNs), decision trees (DTs), and support vector machines (SVMs) for this purpose. Cryptocurrencies improve multiasset portfolio performance in various ways [15]. The portfolio's minimal variance is enhanced, and the efficient frontier is moved forward. Cryptocurrencies also reduce standard deviations and boost Sharpe ratios when added to portfolios.

BTC price forecasts can be made using machine learning algorithms, according to the literature. A machine learning algorithm predicted the price of Bitcoin, Ethereum, LTC, XRP, and Stellar [14]. As compared to Artificial Neural Networks (ANNs) and Deep Learning, SVMs produced the highest accuracy. Using correlation analysis, [16] selected the most reliable predictors of BTC and ETH prices based on several factors. Based on the selected features, linear regression performed better than SVM, linear regression, and random forest (RF). LSTM, which is a type of deep learning, was also tested for predicting Bitcoin and Ethereum prices, and LSTM achieved the best result for Bitcoin. An ensemble model based on machine learning has been used to forecast nine different cryptocurrency prices using ANNs, KNNs, gradient-boosted trees, and ensemble models [17], [18].

The ensemble learning model produced the lowest prediction errors, according to the findings. Prediction errors were the lowest for the ensemble learning model. RF and Gradient Boosting Machines (GBM) were incorporated into an ensemble model to predict three cryptocurrencies - BTC, ETH, and XRP - in [19]. These predictions were then calculated using MAPE, and the results indicated a range of MAPE values between 0.92% and 2.61%. A lot of DL models have been developed recently that focus on predicting financial time series. During deep learning, artificial neural networks are trained on large datasets. This type of network can learn and make intelligent decisions, unlike traditional neural networks. RNNs, including LSTMs and GRUs, are frequently used to predict time series. As described in [20], an LSTM model using selected features is used in conjunction with an ANN and RF to predict BTC prices. Based on the results, an LSTM model outperformed both ARIMA and SVM. A hybrid model combining LSTMs and GRUs was used to predict Monero's (XMR) and LTC's prices [21].

The gold price has been predicted using various models developed by researchers. A new algorithm for artificial bee colonies was developed by the Author [22] in combination with a wavelet neural network method. Blockchain technology and cryptographic functions are used in cryptocurrency to achieve transparency, decentralization, and immutability [23], [24]. An anonymous person or group invented Bitcoin (BTC) in 2009, making it the first and most popular cryptocurrency. ETH and Ripple (XRP) are among the alternative cryptocurrencies created since then, proving that cryptocurrency has emerged as a financial asset. Most cryptocurrencies are dominated by BTC, ETH, and

XRP, which account for almost 79.5% of all cryptocurrency market capitalization. In addition to helping cryptocurrency investors make prudent investment decisions, cryptocurrency price predictions can help financial researchers study cryptocurrency market behaviour. The same methods can be used to predict cryptocurrency prices as for stock price predictions. AutoRegression Integrated Moving Averages (ARIMAs) have been used to predict cryptocurrency prices and movements [25]. Although these models are more accurate at predicting time series than Deep Learning algorithms, they cannot recognize non-linear patterns in very complex prediction problems [26], [27].

3 PROPOSED METHODOLOGY

In this section, we describe how the study was processed and modelled. After that, a selection of cryptocurrency prediction plots is demonstrated. Lastly, we evaluate the performance and analysis of the study. Through the use of LSTMs, GRUs, and Bi-LSTMs, we use deep learning techniques to make predictions about Bitcoin, Ethereum, and Litecoin prices. As part of the evaluation process for Bitcoin, Ethereum, and LTC, historical data is collected, exploratory data visualization is performed, each dataset is divided into training and testing datasets, and three different types of machines are trained, tested, and compared.

3.1 Dataset

Each deep learning model is composed of three layers (LSTM, Bi-LSTM, and GRU), each with 100 neural connections. Figure 1 shows how the dataset was preprocessed. As part of the deep learning process, we conducted a variety of preprocessing techniques on the cryptocurrency data. Following data imputation, we reshaped the data so that it can be applied to LSTMs, Bi-LSTMs, and GRUs. Upon examining the dataset, we found that there were missing values, which were then replaced using the previously recorded observations using a straightforward imputation technique. A normalized model ensures accuracy, avoids bias and is fundamental to model fitting. For the purpose of reducing the potential issue of unequal treatment, we used MinMax Scaling to normalize variables with different scales prior to fitting the model. Studies have shown that these methods can enhance the performance of models [28]. As a result, we scaled the data using MinMax Scalar in this study. As a way

of maintaining continuity between cryptocurrencies, we divided training into 80:20 tests. Testing datasets have been collected since 1 January 2022; training datasets have been collected since 1 January 2018, and training datasets have been collected since 1 January 2018. These studies utilized a number of Python 3 libraries, including NumPy, Pandas, Matplotlib, Keras, and scikit-learn.

3.2 Machine Learning Algorithms

3.2.1 Long Short-Term Memory – LSTM

LSTMs are an update to RNNs. Thus, long-term dependence problems can be avoided, while vanishing gradient problems are solved through an additional mechanism that regulates information and allows it to persist over time [29]. There is a recurrent network of interconnected memory blocks in the LSTM architecture. Memory blocks are also responsible for maintaining the network's state throughout its existence as well as regulating data flow between individual cells. According to Figure 1, an LSTM has input, x_t , output h_t , and an activation function h_t . Input gate calculations determine what information should remain in a cell's state, while output gate calculations compute what information should be sent out.

These (1) – (5) describe how LSTM networks forward train:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i), \quad (1)$$

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f), \quad (2)$$

$$c_t = f_t * c_{t-1} + i_t * \tanh(W_c[h_{t-1}, x_t] + b_c), \quad (3)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o), \quad (4)$$

$$h_t = o_t * \tanh(c_t). \quad (5)$$

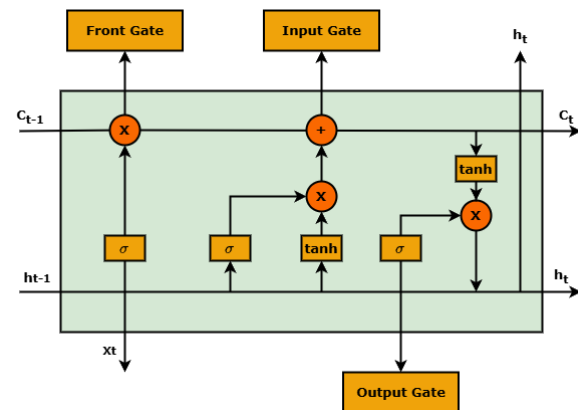


Figure 1: An LSTM algorithm's structure.

In the case of x_t , h_t , and c_t , it is the input gate, forget gate, and output gate, respectively, at time step t .

This matrix is represented by and is composed of two weight matrices and one bias vector. Sigmoid functions ($\tanh$) are also used to limit the output along with hyperbolic tangent functions (sig).

3.2.2 Gated Recurrent Unit (GRU)

A new extension of RNN, GRU, was developed using the architecture of LSTM networks [30]. In addition to handling arbitrary input sequences, GRUs store information about past inputs in a state that is continually updated. GRUs perform different operations inside than LSTMs, so they are also different in their inner workings. As opposed to LSTMs, GRUs have simpler control of information flow, and they need fewer gates. Reset gates and update gates determine what should be forgotten by GRUs. Their training may be easier than LSTMs. Although GRUs have a simple architecture, they are capable of performing as well as LSTMs in a broad range of applications [31].

GRUs have been proven to be more efficient at modelling language than LSTMs in the Penn Treebank dataset. GRUs are superior to non-GRUs for noisy language models, as shown by this result. According to a comparison of natural language processing models, GRUs are as good as LSTMs and CNNs when used across benchmarks [30]. GRUs are well suited to a variety of NLP tasks, demonstrating their adaptability and strength. One of the advantages of GRUs is that they can catch long-range connections more effectively than standard RNNs. Due to the current information and the organization's express, GRUs can recall or dispose of verified information based on their current plan's update and reset entryways. For activities requiring memory retention and application of long sequences of information, selective retention is imperative. It is because GRUs are capable of managing long-term dependencies that they excel at translating lengthy text sequences. Therefore, GRUs excel at handling multiple consecutive information tasks and can even outperform RNNs and LSTMs under certain circumstances [31].

One hundred neurons make up the GRU network, allowing it to capture complex sequences and temporal patterns in data. This training process will use 32 batches to span twenty epochs while balancing resource use and effectiveness. For overfitting to be avoided, 20% of neurons should be turned off during

training according to the Dropout of 0.2. With Adam, sparse gradients can be easily handled, and the learning rate is adaptive. Since it incorporates linear activation as part of its other methods, the GRU design is particularly suitable for tasks such as forecasting time series or predicting continuous values. Reset and update gates are present in the GRU. A gate that updates the hidden state with a new input determines the amount of hidden state to be forgotten by a gate that resets the state with a new input. GRUs contain reset gates (r_t), update gates (z_t), hidden states (h_t), and candidate hidden states (h_{tt}) following (6) – (9):

$$r_t = \sigma(W_r[h_{t-1}, x_t]), \quad (6)$$

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_t), \quad (7)$$

$$h_t = (1 - z_t) * (h_{t-1}) + z_t * h_{tt}, \quad (8)$$

Anywhere

$$h_{tt} = \tanh(W_h[r_t + h_{t-1}, x_t]) \quad (9)$$

weight matrices W_r, W_z, W_h

As shown in the figure, σ represents the sigmoid activation function. Function for activating tangents $\tanh$. Weight matrices W .

3.2.3 Bi-Directional Long Short-Term Memory

In RNNs, long-term gradients explode and disappear over time due to LSTMs. Due to back-propagation through time (BPTT), it is hard for standard RNNs to prepare long successions, which can result in exploding or disappearing gradients. Bi-LSTM cells are used in place of RNN cells to resolve this issue.

In a cell case, data is entered through three entryways. In the first entryway, sigmoid layers choose which data to discard, as illustrated by the following (10).

$$C_t = f_t \cdot C_{t-1} + i_t \cdot C_t. \quad (10)$$

Equations (1) - (3) are then used to update the state of the cells.

Our paper uses a Bi-LSTM that we developed. Essentially, it is a deep learning algorithm that feeds the input sequence into two different networks, one of which follows normal time and the other follows reverse time. There is a sequential output from each network at each time step. The stacked layer Bi-LSTM architecture allows for both forward and backward information to be obtained at each time step, leading to high classification accuracy.

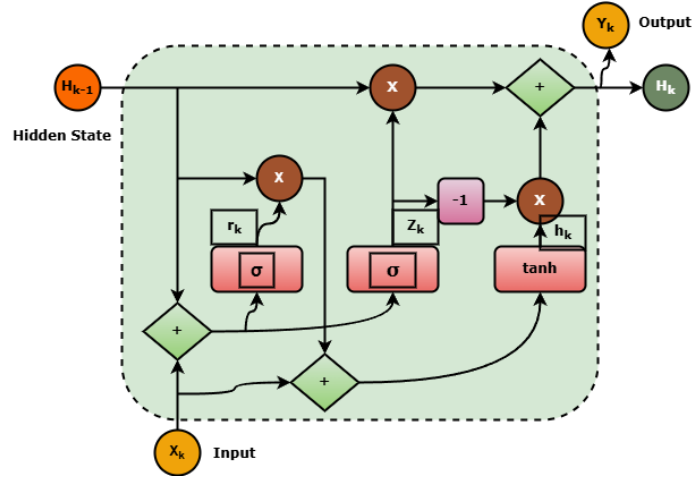


Figure 2: Block diagram of a GRU cell.

According to (11)-(13), bi-LSTM classifiers can handle data both forward and backwards.

$$h_t = f(w_1 x_t + w_2 h_{t-1}), \quad (11)$$

$$h_t = f(w_3 x_t + w_5 h_{t+1}), \quad (12)$$

$$O_t = g(w_4 h_t + w_6 h_t), \quad (13)$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^n (A_t - P_t)^2}{n}}, \quad (14)$$

$$MAPE = \frac{100}{N} * \sum_{l=1}^n \frac{|A_t - P_t|}{A_t}. \quad (15)$$

A forecast is represented by P_t , an actual value by A_t , and a time step by n is represented by n .

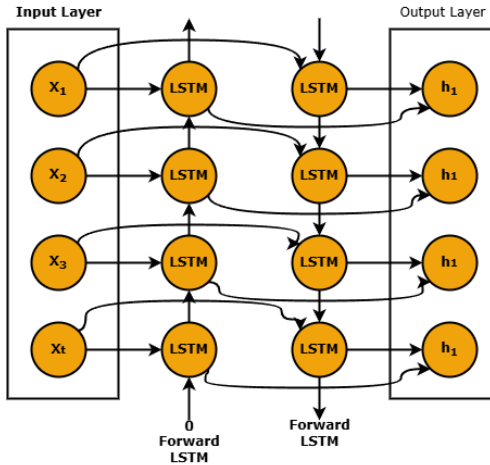


Figure 3: A block model of bidirectional LSTM algorithm (Bi-LSTM).

3.3 Performance Metrics

Our evaluation of the proposed DL algorithms used both RMSE and MAPE. A prediction model that has a smaller RMSE and MAPE value will perform better:

4 RESULT ANALYSIS AND DISCUSSION

A comparison of RMSE values for three deep learning models is presented in Figure 4 based on prices predicted for Bitcoin (BTC), Ethereum (ETH), and Litecoin (LTC). Performance metrics such as RMSE are used as a measure of predictive accuracy, with lower values indicating a better prediction. According to the results, LSTM produces the lowest accurate predictions for BTC compared to other models. While Bi-LSTM is an improvement over LSTM, it still exhibits relatively high error rates for BTC and ETH. A GRU model, however, demonstrates superior prediction accuracy across all three cryptocurrencies with the lowest RMSE. The GRU model, as well as all other models, predicts Litecoin prices accurately with very low errors. Among all three models evaluated, the GRU model proves to be the most accurate for predicting cryptocurrency prices.

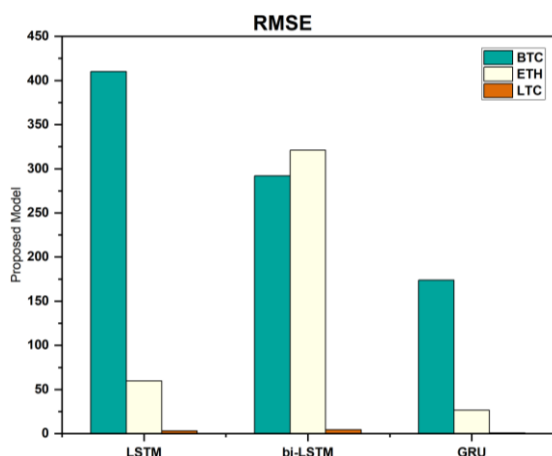


Figure 4: RMSE comparison of LSTM, Bi-LSTM, and GRU models for cryptocurrency price prediction.

Figure 5 illustrates the MAPE model using LSTM, Bi-LSTM, and GRU to predict Bitcoin (BTC), Ethereum (ETH), and Litecoin (LTC). MAPE measures predictive accuracy, with lower values indicating better accuracy. All three cryptocurrencies maintain moderate error levels, with Ethereum showing slightly higher error levels than Bitcoin and Ethereum Classic. A comparison of the MAPE values between the Bi-LSTM model and the Bi-LSTM shows that the Bi-LSTM model is less efficient, especially for Bitcoin and Ethereum. Among all cryptocurrencies, GRU exhibits the lowest MAPE, demonstrating superior predictive abilities. In this study, it is confirmed that GRU consistently outperforms both LSTMs and Bi-LSTMs in forecasting cryptocurrency prices.

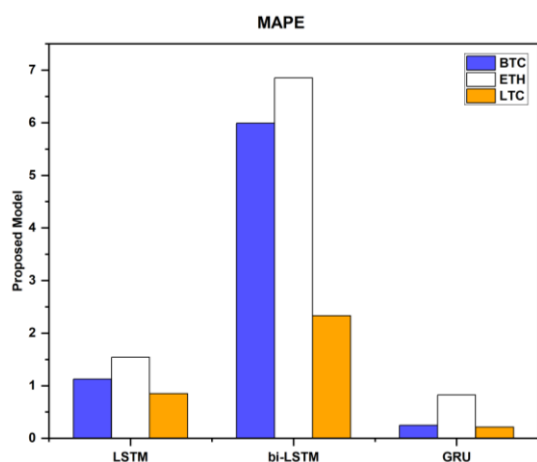


Figure 5: MAPE comparison of LSTM, Bi-LSTM, and GRU models for cryptocurrency price prediction.

From early 2018 to mid-2021, Bitcoin (BTC), Ethereum (ETH), and Litecoin (LTC) closed at their respective closing prices in Figure 6. The x-axis represents timelines, and closing prices are represented by the y-axis. During the period between late 2020 and early 2021, Bitcoin's price spiked to almost \$60,000 before rapidly dropping. As opposed to Ethereum, Litecoin's price ranges remain relatively stable throughout the same period. Compared to Bitcoin, Ethereum and Litecoin show more consistent price movements, but their volatility is lower.

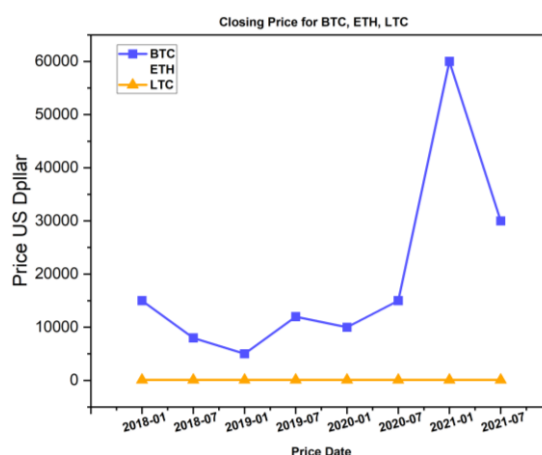


Figure 6: Closing price trends of BTC, ETH, and LTC from 2018 to 2021.

5 CONCLUSIONS

For the purpose of predicting cryptocurrency prices, the researchers employed three prominent deep learning algorithms: Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and Bidirectional LSTM (Bi-LSTM). Historical data from Bitcoin (BTC), Ethereum (ETH), and Litecoin (LTC) were systematically used to train and validate the models, enabling a thorough evaluation of their respective predictive capabilities. Among these algorithms, GRU consistently demonstrated superior performance, achieving notably lower Root Mean Square Errors (RMSEs) and Mean Absolute Percentage Errors (MAPEs) across all three cryptocurrencies. In contrast, LSTM and Bi-LSTM models yielded consistently higher prediction errors compared to GRU, particularly noticeable for BTC and ETH, indicating a relative disadvantage in accurately modeling complex market fluctuations. The study highlights GRU's simpler architectural design, fewer training parameters, and enhanced ability to efficiently capture temporal dependencies,

positioning it as an optimal choice for modeling highly volatile and dynamically evolving financial time series data like cryptocurrency prices. These findings provide valuable insights for investors, traders, financial analysts, and researchers, assisting them in developing informed strategies, managing risks effectively, and enhancing research methodologies within the rapidly evolving landscape of digital finance.

REFERENCES

- [1] J. Melitz, DP178 Monetary Discipline, Germany, and the European Monetary System; National Bureau of Economic Research (NBER) Working Paper No. 2319; National Bureau of Economic Research (NBER): Cambridge, MA, USA, 1987. [Online]. Available: <https://ssrn.com/abstract=884539>. Accessed: Apr. 25, 2025.
- [2] A. Bulir, "Income inequality: Does inflation matter?," IMF Staff Papers, vol. 48, no. 1, pp. 139–159, Jan. 2001, doi: 10.2307/4621662.
- [3] S. Basco, "Globalization and financial development: A model of the Dot-Com and the Housing Bubbles," Journal of International Economics, vol. 92, no. 1, pp. 78–94, Jan. 2014, doi: 10.1016/j.jinteco.2013.10.008.
- [4] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Decentralized Business Review, 2008. [Online]. Available: <https://bitcoin.org/bitcoin.pdf>. Accessed: Apr. 25, 2025.
- [5] P. N. Sureshbhai, P. Bhattacharya, and S. Tanwar, "KaRuNa: A blockchain-based sentiment analysis framework for fraud cryptocurrency schemes," in 2020 IEEE International Conference on Communications Workshops (ICC Workshops), IEEE, 2020, pp. 1–6. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/9145151/>. Accessed: Apr. 25, 2025.
- [6] P. Rani, S. Verma, S. P. Yadav, B. K. Rai, M. S. Naruka, and D. Kumar, "Simulation of the lightweight blockchain technique based on privacy and security for healthcare data for the cloud system," International Journal of E-Health and Medical Communications (IJEHMC), vol. 13, no. 4, pp. 1–15, 2022.
- [7] S. Corbet, B. Lucey, M. Peat, and S. Vigne, "Bitcoin futures – What use are they?," Economics Letters, vol. 172, pp. 23–27, Nov. 2018, doi: 10.1016/j.econlet.2018.07.031.
- [8] S. Corbet, A. Meegan, C. Larkin, B. Lucey, and L. Yarovaia, "Exploring the dynamic relationships between cryptocurrencies and other financial assets," Economics Letters, vol. 165, pp. 28–34, Apr. 2018, doi: 10.1016/j.econlet.2018.01.004.
- [9] D. Valdeolmillos, Y. Mezquita, A. González-Briones, J. Prieto, and J. M. Corchado, "Blockchain technology: A review of the current challenges of cryptocurrency," in Blockchain and Applications, J. Prieto, A. K. Das, S. Ferretti, A. Pinto, and J. M. Corchado, Eds., Advances in Intelligent Systems and Computing, vol. 1010. Cham: Springer, 2020, pp. 153–160, doi: 10.1007/978-3-030-23813-1_19.
- [10] A. Singh et al., "Blockchain-based lightweight authentication protocol for next-generation trustworthy Internet of Vehicles communication," IEEE Transactions on Consumer Electronics, vol. 70, no. 2, pp. 4898–4907, May 2024, doi: 10.1109/TCE.2024.3351221.
- [11] M. H. Yuneline, "Analysis of cryptocurrency's characteristics in four perspectives," Journal of Asian Business and Economic Studies, vol. 26, no. 2, pp. 206–219, Oct. 2019, doi: 10.1108/JABES-12-2018-0107.
- [12] T. L. D. Huynh, M. A. Nasir, X. V. Vo, and T. T. Nguyen, "'Small things matter most': The spillover effects in the cryptocurrency market and gold as a silver bullet," North American Journal of Economics and Finance, vol. 54, p. 101277, Nov. 2020, doi: 10.1016/j.najef.2020.101277.
- [13] P. Rani, K. Ur Rehman, S. P. Yadav, and L. Hussein, "Deep learning and AI in behavioral analysis for revolutionizing mental healthcare," in Demystifying the Role of Natural Language Processing (NLP) in Mental Health, A. Mishra, S. P. Yadav, M. Kumar, S. M. Biju, and G. C. Deka, Eds., IGI Global, 2025, pp. 263–282, doi: 10.4018/979-8-3693-4203-9.ch014.
- [14] N. A. Hitam, A. R. Ismail, R. Samsudin, and E. H. Alkhamash, "The effect of kernel functions on cryptocurrency prediction using support vector machines," in Advances on Intelligent Informatics and Computing, F. Saeed, F. Mohammed, and F. Ghaleb, Eds., Lecture Notes on Data Engineering and Communications Technologies, vol. 127. Cham: Springer, 2022, pp. 319–332, doi: 10.1007/978-3-030-98741-1_27.
- [15] Z. S. Jassim and M. M. Kassir, "Enhancing malware detection through machine learning techniques," International Journal of Data Science (IJDS), vol. 1, no. 1, pp. 1–15, Jun. 2024.
- [16] M. Saad, J. Choi, D. Nyang, J. Kim, and A. Mohaisen, "Toward characterizing blockchain-based cryptocurrencies for highly accurate predictions," IEEE Systems Journal, vol. 14, no. 1, pp. 321–332, 2019.
- [17] R. Chowdhury, M. A. Rahman, M. S. Rahman, and M. R. C. Mahdy, "An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning," Physica A: Statistical Mechanics and Its Applications, vol. 551, p. 124569, 2020.
- [18] G. Ansari, P. Rani, and V. Kumar, "A novel technique of mixed gas identification based on the group method of data handling (GMDH) on time-dependent MOX gas sensor data," in Proceedings of International Conference on Recent Trends in Computing: ICRTC 2022, Springer, 2023, pp. 641–654.
- [19] V. Derbentsev, V. Babenko, K. Khrustalev, S. Khrustalova, and H. Obruch, "Comparative performance of machine learning ensemble algorithms for forecasting cryptocurrency prices," International Journal of Engineering Transactions B: Basics, vol. 34, no. 1, pp. 140–148, 2021.
- [20] W. Chen, H. Xu, L. Jia, and Y. Gao, "Machine learning model for Bitcoin exchange rate prediction using economic and technology determinants," International Journal of Forecasting, vol. 37, no. 1, pp. 28–43, 2021.

- [21] M. M. Patel, S. Tanwar, R. Gupta, and N. Kumar, "A deep learning-based cryptocurrency price prediction scheme for financial institutions," *Journal of Information Security and Applications*, vol. 55, p. 102583, 2020.
- [22] B. Li, "Research on WNN modeling for gold price forecasting based on improved artificial bee colony algorithm," *Computational Intelligence and Neuroscience*, vol. 2014, pp. 1–10, 2014, doi: 10.1155/2014/270658.
- [23] A. Narayanan, J. Bonneau, E. Felten, A. Miller, and S. Goldfeder, *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*. Princeton University Press, 2016.
- [24] P. Rani and R. Sharma, "Intelligent transportation system for Internet of Vehicles based vehicular networks for smart cities," *Computers and Electrical Engineering*, vol. 105, p. 108543, 2023.
- [25] D. Olvera-Juarez and E. Huerta-Manzanilla, "Forecasting bitcoin pricing with hybrid models: A review of the literature," *International Journal of Advanced Engineering Research and Science*, vol. 6, no. 9, pp. 161–164, 2019.
- [26] S. Siarni-Namini, N. Tavakoli, and A. S. Namin, "A comparison of ARIMA and LSTM in forecasting time series," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2018, pp. 1394–1401. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/8614252/>. Accessed: Apr. 25, 2025.
- [27] E. A. A. Abdul Mahdi, A. Al-Askery, A. H. Ali, and S. Ghasemi, "A survey of datacenter and optical data center networks," *Journal of Telecommunications (JT)*, vol. 6, no. 3, pp. 66–75, Sep. 2024.
- [28] M. M. Ahsan, M. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, vol. 9, no. 3, p. 52, 2021.
- [29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
- [30] X. Wang, W. Jiang, and Z. Luo, "Combination of convolutional and recurrent neural network for sentiment analysis of short texts," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 2428–2437.
- [31] S. Yang, X. Yu, and Y. Zhou, "LSTM and GRU neural network performance comparison study: Taking Yelp review dataset as an example," in *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*, IEEE, 2020, pp. 98–101. Accessed: Apr. 16, 2025.

Chromaticity of Normal Graphs and Cyclic Graphs

Abdullah AbdulRazzaq Obaid and Haneen Kareem Hussein

Department of Mathematics, Faculty of Computer Sciences and Mathematics, University of Kufa, 54003 Najaf, Iraq
haneenk.aljanabi@uokufa.edu.iq

Keywords: Chromaticity, Normal Graphs, Cyclic Graphs, Dihedral Group, Chromatic Uniqueness.

Abstract: In our work, we will find new results about the chromaticity of Normal graphs and Cyclic Graphs of Dihedral group D_{2n} , ($G_{Nor}D_{2n}$ and $G_{Cyc}D_{2n}$, respectively). We will determine the new theorems about the chromatic uniqueness of these graphs, and we will study all cases of a positive integer number n for $G_{Nor}D_{2n}$ and $G_{Cyc}D_{2n}$. Our primary objective is to establish new theorems regarding the chromatic uniqueness of these graphs, which will significantly enhance the understanding of their structural and combinational characteristics. We will thoroughly analyze all cases of positive integers n to accurately determine the chromatic numbers and systematically explore the implications of these findings within graph theory. This research not only contributes to the theoretical framework of graph theory by enriching its foundational concepts but also opens promising avenues for future interdisciplinary studies examining the intricate interplay between group theory, combinatorial mathematics, and graph properties, thereby advancing both theoretical insights and practical computational applications.

1 INTRODUCTION

During the more than a century that the Four-colour problem was discussed, many ideas were proposed that led to solve this known problem [8]. In 1912, Birkhoff [4] defined a mapping $P(M, \lambda)$ which gives the number of proper λ -colourings of a map M . In 1932, The author Whitney [25] is the one who expanded the concept of chromatic polynomials, he transferred the study from maps to graphs. Also he put some basic consequence of this. In 1968, Read [23] asked two questions:

- 1) What is the sufficient and necessary conditions for two graphs to share the same chromatic polynomial; that means, to be chromatically equivalent?
- 2) What is the sufficient and necessary conditions for a graph to be chromatically unique?

In particular, Chao and Whitehead Jr. [7] in 1978 defined a graph to be chromatically unique if no other graph shares its chromatic polynomial.

Chromaticity problem of graph means the study and discuss the two questions above. There are many authors worked in this area. Dong, Koh, and Teo [11] published a useful book on chromatic polynomials and chromaticity of graphs. For more details on

chromaticity of graphs, we refer to the survey papers [10], [9], [20], [23], and [25].

The literature of algebraic graph theory has grown extremely since 1974, many papers have appeared in this area next. Some researchers studied the colouring of the graph of ring, one of them the famous author, I. Beck. He introduced the idea of a zero-divisor graph in 1988 [3]. While he was mainly interested in colorings. In 2013, Sahib and Khalaf [12] introduce many results about the chromatic uniqueness of some algebraic graph like the zero-divisor graph $\Pi(z_n)$, for the cases, when n is even, $n = p^2$ or $n = p_1 \times p_2$. The chromaticity of $\Pi(z_n)$ is still an open problem for the other cases of n .

In 2022, Al-Janabi and Gabor [2] solved this problem for all cases of n .

Other researchers studied the properties of the graph for group. We will show about some of them. For more information, see these researches [18] and [19].

In 1975-1987, Cavior and Calhoun respectively [5] and [6], compute the subgroups of the dihedral group D_{2n} .

In 2013, Ma presented a definition for the cyclic graph of the Dihedral group and Dicyclic group, where the elements of G are the vertices of $\Gamma(G)$, such that two different vertices x and y are connected by one edge iff $\langle x, y \rangle$ is a cyclic subgroup of G [25].

In 2022, Ohood and Hayder presented three types of graphs, which are called Normal and Cyclic, denoted by G_{Nor} and G_{Cyc} , respectively [13], [14], [15], [16], and [17].

In our work, we will discuss the new results about the chromatic uniqueness of Normal graphs for Dihedral group $D_{2n}(G_{Nor}D_{2n})$.

2 PRELIMINARIES

All graphs used in this paper are simple and the graph G defined as $G = (V, E)$ such that the vertices set of G are denoted by $V = V(G)$, and $E = E(G)$ is the set edges of G . The degree of vertex $v(v \in V(G))$ in G , is the number of all edges incident to v , and denoted by $d(v)$. A graph G is said to be complete if any two vertices of G are adjacent and denoted by K_m . The graph $G' = (V', E')$ is subgraph of the graph $G = (V, E)$ if $\emptyset \neq V'(G) \subseteq V(G)$, and $E'(G) \subseteq E(G)$. Two graphs G and H are isomorphic if there is a bijective function $f: V(G) \rightarrow V(H)$ with preserves the edges, in other any two vertices are adjacent in G iff they are adjacent in H , and denoted by $G \cong H$ [1]. The matrix degree sequence of the simple graph G of order n is defined by:

$$\mathcal{M} = \begin{pmatrix} deg(v_1) & deg(v_2) & \dots & deg(v_n) \\ \mu(deg(v_1)) & \mu(deg(v_2)) & \dots & \mu(deg(v_n)) \end{pmatrix},$$

where $deg(v)$ is the number degree of vertex v in graph G and $\mu(deg(v))$ is the multi degree of vertex v [14]. $P(G, \lambda)$ is the number of distinct λ -colouring of G , thus $P(G, \lambda) > 0$ iff G is λ -colouring. If $P(G, \lambda) = P(H, \lambda)$, then the graphs G and H are chromatically equivalent or χ -equivalent and denoted by $G \sim H$. Moreover, if the graphs G, H are isomorphic $H \cong G$, then G is chromatically unique or χ -unique.

3 AXILLARY RESULTS

Definition 3.1 [11]: A mapping of the graph G $g: V(G) \rightarrow V\{1, 2, \dots, \lambda\}$, $\lambda \in N$, is defined as λ -coloring of G if $g(u) \neq g(v)$, for all vertices u and v are adjacent. The number of λ -colorings of G it is indicated by $P(G, \lambda)$.

Proposition 3.2 [22]: For two different vertices x and y in a graph G . $P(G, \lambda)$ is given below:

$$P(G, \lambda) = P(G + xy, \lambda) + P(G \circ xy, \lambda),$$

where $G + xy$ is adding a new edge xy to G , such that $x, y \in V(G)$, and $xy \notin E(G)$. The graph $G \circ xy$ is

given by G and identifying two vertices x and y which incident with e and deleting all the multiple edges (except one), if they appear.

Corollary 3.3 [21]: For an edge in the graph G , $P(G, \lambda)$ is given by:

$$P(G, \lambda) = P(G - uv, \lambda) + P(G \circ uv, \lambda),$$

where the graph $G - uv$ is obtained from G by removing uv .

Corollary 3.4 [11]: For the complete graph on n vertices the chromatic polynomial it is:

$$P(K_n, \lambda) = \lambda(\lambda - 1) \dots (\lambda - (n - 1)).$$

Likewise, it is known that $P(\overline{K_n}, \lambda) = \lambda^n$.

Theorem 3.5 [11]: If $G \cap H$ is a complete graph, then

$$P(H \cup G, \lambda) = \frac{P(H, \lambda) \times P(G, \lambda)}{P(H \cap G, \lambda)}.$$

Definition 3.6 [24]: A group G generated by two elements a and b with given specific relations is defined a Dihedral group.

$$G = \{ \langle a, b \rangle : |a| = 2, |b| = n, ba^{-1} = ab \text{ if } e = a \text{ or } ab = ba^{n-1} \}.$$

G is denoted by D_n or D_{2n} and called Dihedral group of degree n or $2n$.

$$D_{2n} = \{ e, a, a^2, \dots, a^{n-1}, b, ab, a^2b, \dots, a^{n-1}b \},$$

where $a^n = e, b^2 = e, bab = a^{-1}$.

Definition 3.7 [24]: A subgroup H of be a group G is a subset of G and denoted by $H \leq G$, such that H is also a group with the operation of group G .

Definition 3.8 [24]: Let $H \subset G$, then H is called a normal subgroup of G and denoted by $H \trianglelefteq G$ if H is closed with respect to conjugates, in other words, for every $a \in H$ and $x \in G, xax^{-1} \in H$.

Definition 3.9 [17]: We have G is a finite group. The cyclic graph of group G is denoted by G_{Cyc} , it whose two sets, the set $E(G_{Cyc}) = \{ \langle a, b \rangle | \langle a, b \rangle \leq G \}$ is an edge set, and $V(G_{Cyc}) = \{ a | a \in G \}$ is a vertex set.

Definition 3.10 [14]: Let G be a finite group. The normal graph of G is denoted by G_{Nor} , and it is defined as $E(G_{Nor}) = \{ \langle a, b \rangle | \langle a, b \rangle \trianglelefteq G \}$ the set edges, and $V(G_{Nor}) = \{ a : a \in G \}$ the set of vertices.

Proposition 3.11 [14]: The matrix degree sequence for the normal graph of the Dihedral group D_{2n} is given by the following: If n is even positive integer number, then the matrix is:

$$\mathcal{M}(G_{Nor}D_{2n}) = \begin{pmatrix} 2(\varphi(n) + \varphi(\frac{n}{2})) & 2n-1 & n-1 \\ n & \varphi(n) + \varphi(\frac{n}{2}) & n - (\varphi(n) + \varphi(\frac{n}{2})) \end{pmatrix}$$

Proposition 3.12 [17]: The matrix degree sequence for the cyclic graph of Dihedral group D_{2n} is given by the following:

$$\mathcal{M}(G_{Cyc}D_{2n}) = \begin{pmatrix} 2n-1 & n-1 & 1 \\ 1 & n-1 & n \end{pmatrix},$$

where n is a positive integer number.

Definition 3.13 [1]: Let the complete graph K_n , such that $M = V(K_n)$, and L_k is subset of M , $\forall k \leq t$ and $t \in N$. The t -clique-join graph (simply, t -CJ-graph) is obtained from t arbitrary graphs $Z_1, Z_2, \dots, Z_t$, on pairwise disjoint vertex sets by joining every vertex in Z_k with all vertices of L_k , where $Z_k \cap M = \emptyset$. It is denoted by $F = F(M, L_1, L_2, \dots, L_t, Z_1, Z_2, \dots, Z_t)$. $\mathfrak{S} = \mathfrak{S}(M, f_1, f_2, \dots, f_t, Z_1, Z_2, \dots, Z_t)$ is the set of all such graphs with $|L_1| = f_1, |L_2| = f_2, \dots, |L_t| = f_t$. For example, $F \in \mathfrak{S}$.

Corollary 3.14 [1]: We have an empty graph it is Z_k of order a_k , and let L_k be the clique subgraphs of order f_k , $L_k \subset M$, $1 \leq k \leq 2$ and $f_1 \neq f_2$. Let $G \in \mathfrak{S}(M, f_1, f_2, Z_1, Z_2)$ be a 2-CJ-graph, $|M| = m$, then the chromatic polynomial of G is:

$$P(G, \lambda) = \lambda(\lambda-1) \dots (\lambda-f_1)^{a_1+1} \dots (\lambda-f_2)^{a_2+1} \dots (\lambda-m+1).$$

Theorem 3.15 [1]: Let G be connected graph with chromatic polynomial of an integral-root and P be the integral-root, such that P has some roots with exponent at least 3, then G it is not χ -unique.

Theorem 3.16 [1]: Let P be an integral-root chromatic polynomial of a connected graph G , such that P has one root only with exponent 2 and no more roots with the exponent, then it is χ -unique.

Notion 3.17 [22]: We will denote by G_{Nor-e} for the graph $G_{Nor}D_{2n} \setminus e$, where $V(G_{Nor-e}) = \{a, a^2, \dots, a^{n-1}, b, ab, \dots, a^{n-1}b\}$.

Lemma 3.18 [22]: If $G_{Nor}D_{2n}$ be a normal graph of D_{2n} , then there is complete subgraph $K_{\frac{3n}{2}-2}$ of $G_{Nor}D_{2n}$, where $n \geq 6$ is an even number.

Theorem 3.19 [22]: Let n be an odd positive integer number. Then the chromatic polynomial for a normal graph of the Dihedral group D_{2n} is given by:

$$P(G_{Nor}D_{2n}, \lambda) = \lambda(\lambda-1) \dots (\lambda-(n-1))^2 \dots (\lambda-(2n-2)).$$

Theorem 3.20 [22]: The chromatic polynomial of $G_{Nor}D_{2n}$, where n be an even positive integer number, $n \geq 6$, is given by:

$$P(G_{Nor}D_{2n}, \lambda) = \lambda(\lambda-1) \dots (\lambda-(n-2))^2 (\lambda-(n-1))^2 \dots (\lambda-(\frac{3n}{2}-3)) P(G, \lambda).$$

4 THE MAIN AIM

4.1 Chromatic Uniqueness of Normal Graphs

Theorem 4.1.1: Let $G_{Nor}D_{2n}$ be a normal graph of the Dihedral group D_{2n} , if n is an odd positive integer number, then $G_{Nor}D_{2n}$ is χ -unique.

Proof: If n is odd positive integer number, then (by Theorem 3.19) $P(G_{Nor}D_{2n}, \lambda)$ is given by $P(G_{Nor}D_{2n}, \lambda) = \lambda(\lambda-1) \dots (\lambda-t)^2 \dots (\lambda-(2n-2))$, since this polynomial has one root only of exponent 2, then (by Theorem 3.16) the graph $G_{Nor}D_{2n}$ is χ -unique, where n is an odd number.

Theorem 4.1.2: Let $G_{Nor}D_{2n}$ be a normal graph of the Dihedral group D_{2n} , if $n \geq 6$ is an even number, then $G_{Nor}D_{2n}$ is not χ -unique.

Proof: By (Proposition 3.11), we get:

$$\mathcal{M}(G_{Nor}D_{2n} \setminus a^j b = \{e, a^i\}) = \begin{pmatrix} n-1 & n-1 \\ \varphi(n) + \varphi(\frac{n}{2}) & n - (\varphi(n) + \varphi(\frac{n}{2})) \end{pmatrix} = \begin{pmatrix} n-1 \\ n \end{pmatrix} = V(K_n), 0 \leq j \leq n-1, 1 \leq i \leq n-1,$$

such that $V(K_n) = \{e, a^i\}$, $1 \leq i \leq n-1$, then there is a complete subgraph $K_{\frac{3n}{2}-2}$ of $G_{Nor}D_{2n}$ (by Lemma 3.18), and $V(K_{\frac{3n}{2}-2}) = \{b, ab, a^2b, \dots,$

$a^{\frac{n}{2}-2}b, a^{n-1}b, a^i\}$, $i \neq \frac{n}{2}$. Since the vertex $e \in V(K_n)$, which is connected by one edge to the all vertices $\{a^i\} \in V(K_n)$, $1 \leq i \leq n-1$. Since the vertex b is not adjacent with the vertex e . So, if we change the edge $\{e, a^{\frac{n}{2}}\}$ with the edge $\{e, b\}$, we get the graph H , such that $H \not\cong G_{Nor}D_{2n}$, because the vertices $e, a^{\frac{n}{2}}$ are adjacent in the graph $G_{Nor}D_{2n}$ and not in the graph H . Then (By Theorem 3.20), $P(G_{Nor}D_{2n}, \lambda)$ is given by:

$$P(G_{Nor}D_{2n}, \lambda) = \lambda(\lambda-1) \dots (\lambda-(n-2))^2 (\lambda-(n-1))^2 \dots (\lambda-(\frac{3n}{2}-3)) P(G, \lambda).$$

Now, in the graph H , the vertex e is incident to the vertices $\{b, a^i\} \in V(K_{\frac{3n}{2}-2})$, $i \neq \frac{n}{2}$, such that

$\{a^i\} \in V(K_{n-2}), i \neq \frac{n}{2}$, then $P(K_{\frac{3n}{2}-2} + e, \lambda)$ is given by:

$$P(K_{\frac{3n}{2}-2} + e, \lambda) = \lambda(\lambda - 1) \dots (\lambda - (n - 2)) (\lambda - (n - 1))^2 \dots (\lambda - (\frac{3n}{2} - 3)).$$

Since the vertex $a^{\frac{n}{2}} \in H$ is also incident to the vertices $\{a^i\} \in V(K_{n-2}), i \neq \frac{n}{2}$, then we get the polynomial

$P(K_{\frac{3n}{2}-2} + e + a^{\frac{n}{2}}, \lambda) = \lambda(\lambda - 1) \dots (\lambda - (n - 2))^2 (\lambda - (n - 1))^2 \dots (\lambda - (\frac{3n}{2} - 3))$, and $K_{\frac{3n}{2}-2} + e + a^{\frac{n}{2}}$ is a subgraph of H then the polynomial of H is:

$$P(H, \lambda) = \lambda(\lambda - 1) \dots (\lambda - (n - 2))^2 (\lambda - (n - 1))^2 \dots (\lambda - (\frac{3n}{2} - 3)) P(G, \lambda).$$

Since $H \not\cong G_{NorD_{2n}}$, and $P(G_{NorD_{2n}}, \lambda) = P(H, \lambda)$, then the graph $G_{NorD_{2n}}$ is not χ -unique, where $n \geq 6$ is an even number (see Figure 1).

Example 4.1.3: Let $G_{NorD_{2n}}$ be a graph, where $n = 8$, with vertex set of $G_{NorD_{16}}$ is $\{e, a, a^2, a^3, a^4, a^5, a^6, a^7, b, ab, a^2b, a^3b, a^4b, a^5b, a^6b, a^7b\}$ (Fig. 2). It is clear that $V(K_{\frac{3n}{2}-2}) = V(K_{10}) =$

$\{a, a^2, a^3, a^5, a^6, a^7, b, ab, a^2b, a^7b\}$ and $V(K_{n-1}) = V(K_7) = \{a, a^2, a^3, a^4, a^5, a^6, a^7\}$, the vertex $e \in G_{NorD_{16}}$ is incident to all vertices of K_7 but the vertex $e \in H$ is incident to all vertices of $V(K_6) = \{a, a^2, a^3, a^5, a^6, a^7\}$ and the vertex $b \in K_{\frac{3n}{2}-2}$, then we can apply Theorem 4.1.2, in this graph we used Maple Program to get the following chromatic polynomial (Fig. 2):

$$P(G_{NorD_{16}}, \lambda) = P(H, \lambda) = \lambda(\lambda - 1)(\lambda - 2)(\lambda - 3)(\lambda - 4)(\lambda - 5)(\lambda - 6)^2(\lambda - 7)^2(\lambda - 8)(\lambda - 9)(\lambda^4 - 42\lambda^3 + 665\lambda^2 - 4700\lambda + 12501).$$

Proposition 4.1.4: Let G_{NorD_8} be a normal graph of the Dihedral group D_8 , where $n = 4$, then it is χ -unique.

Proof: For a graph $G_{NorD_{2n}}$ with vertex set is $\{e, a, a^2, a^3, b, ab, a^2b, a^3b\}$, and $n = 4$ (see Fig. 3), $P(G_{NorD_8}, \lambda)$ is given by:

$$P(G_{NorD_8}, \lambda) = \lambda(\lambda - 1)(\lambda - 2)(\lambda - 3)^2(\lambda - 4)(\lambda - 5)(\lambda - 6).$$

Then, (by Theorem 3.7) the polynomial of G_{NorD_8} has one root of multiplicity 2, so G_{NorD_8} is χ -unique, where $n = 4$.

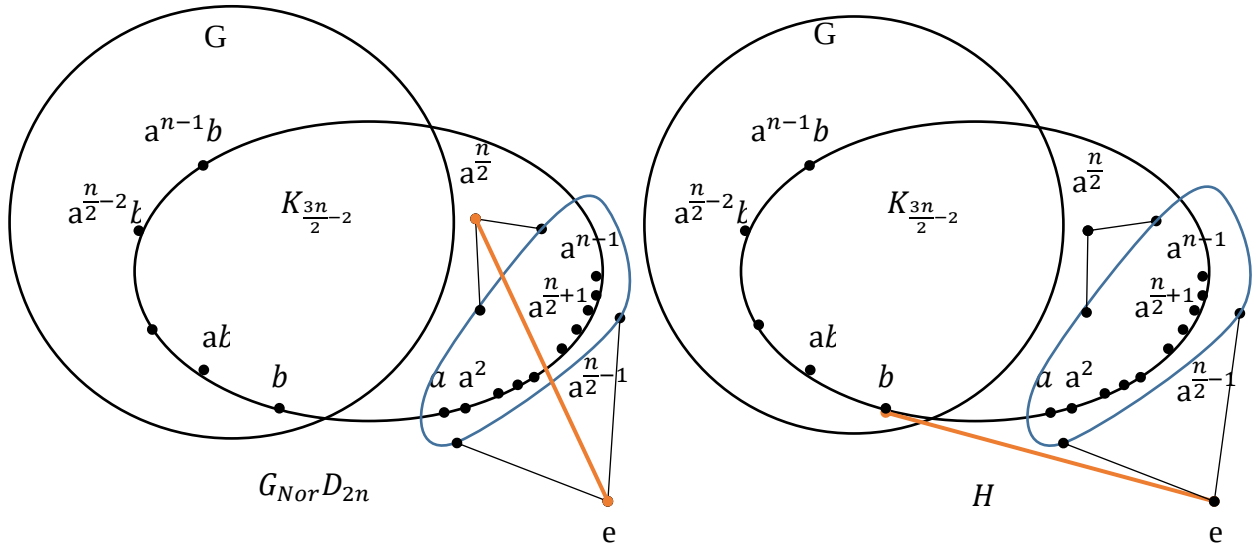


Figure 1: $G_{NorD_{2n}}$ and H are chromatically equivalent.

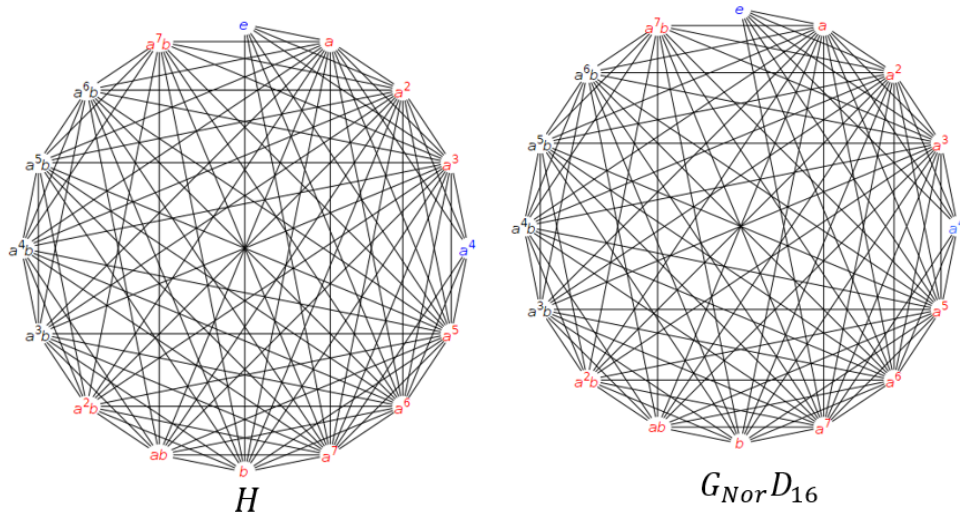


Figure 2: $G_{Nor}D_{16}$ and H are χ – equivalent but non-isomorphic graphs.

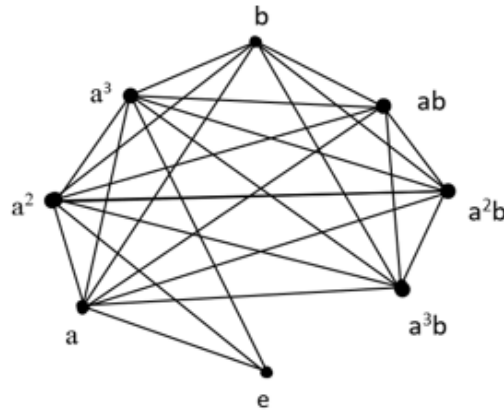


Figure 3: Normal graph of D_8 .

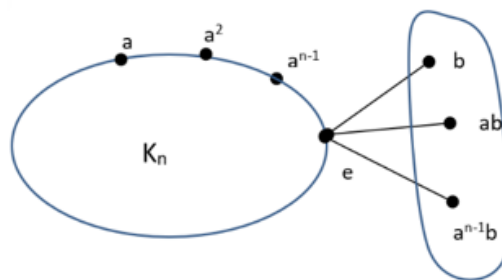


Figure 4: Cyclic graph of D_{2n} .

4.2 Chromatic Uniqueness of Cyclic Graphs

Lemma 4.2.1: For a cyclic graph of D_{2n} . If n is positive integer number, then

$$G_{Cyc}D_{2n} \setminus \{b, ab, \dots, a^{n-1}b\} = \{e, a, a^2, \dots, a^{n-1}\} \cong K_n.$$

Proof: Let $D_{2n} = \{e, a, a^2, \dots, a^{n-1}, b, ab, a^2b, \dots, a^{n-1}b\}$ and by (Proposition 3.12)

$$\mathcal{M}(G_{Cyc}D_{2n}) = \begin{pmatrix} 2n-1 & n-1 & 1 \\ 1 & n-1 & n \end{pmatrix}.$$

$\deg(e) = 2n - 1$, $\mu(\deg(e)) = 1$, $\deg(a^i) = n - 1$, $\mu(\deg(a^i)) = n - 1$, $\deg(a^i b) = 1$ and $\mu(\deg(a^i b)) = n$, $i = 1, 2, \dots, n - 1$. Since $a^i b$ incident to 1 vertex is e then in $G_{Cyc}D_{2n} \setminus \{b, ab, \dots, a^{n-1}b\}$, $\deg(e) = n - 1$. Then

$$\mathcal{M}(G_{Cyc}D_{2n} \setminus \{b, ab, \dots, a^{n-1}b\}) = \begin{pmatrix} n-1 & n-1 \\ 1 & n-1 \end{pmatrix} = \begin{pmatrix} n-1 \\ n \end{pmatrix} = V(K_n).$$

Theorem 4.2.2: Let $n \geq 2$ be a positive integer number. Then the chromatic polynomial in cyclic graph of the Dihedral group D_{2n} is given as:

$$P(G_{Cyc}D_{2n}, \lambda) = \lambda(\lambda - 1)^{n+1} \dots (\lambda - (n - 1))$$

Proof: Let D_{2n} be Dihedral group of order $2n$ such that n is positive integer number, and $D_{2n} = \{e, a, a^2, \dots, a^{n-1}, b, ab, a^2b, \dots, a^{n-1}b\}$ and by (Lemma 4.2.1), then $G_{Cyc}D_{2n} \setminus \{b, ab, \dots, a^{n-1}b\} = \{e, a, a^2, \dots, a^{n-1}\} \cong K_n$, and the chromatic polynomial of K_n is given by the following:

$$P(K_n, \lambda) = \lambda(\lambda - 1) \dots (\lambda - (n - 1)).$$

Since $a^i b$ incident to one vertex (e), then by (Corollary 3.14), see (Figure 4)

$$P(G_{Cyc}D_{2n}, \lambda) = \lambda(\lambda - 1)^{n+1} \dots (\lambda - (n - 1)).$$

Example 4.2.3: Let $G = D_{10} = \{e, a, a^2, a^3, a^4, b, ab, a^2b, a^3b, a^4b\}$, where $n = 5$. By (Theorem 4.2.2) the chromatic polynomial of this graph is given by:

$$P(G_{Cyc}D_{10}, \lambda) = \lambda(\lambda - 1)^6(\lambda - 2)(\lambda - 3)(\lambda - 4).$$

It's clear that, there is a subgraph K_5 , and the vertex e in K_5 , it incident to $\{b, ab, a^2b, a^3b, a^4b\}$.

Theorem 4.2.4: Let $G_{Cyc}D_{2n}$ be a cyclic graph of the Dihedral group D_{2n} , if n is positive integer number, then it is not χ - unique.

Proof: If n is positive integer number, and the vertex set is:

$$V(G_{Cyc}D_{2n}) = \{e, a, a^2, \dots, a^{n-1}, b, ab, a^2b, \dots, a^{n-1}b\},$$

then by (Theorem 4.2.2) the chromatic polynomial of the cyclic graph is:

$$P(G_{Cyc}D_{2n}, \lambda) = \lambda(\lambda - 1)^{n+1} \dots (\lambda - (n - 1)),$$

then (by Theorem 3.15) the cyclic graph, $G_{Cyc}D_{2n}$, is not χ - unique, where $n = p$.

5 CONCLUSIONS

In this study, we explored new results concerning the chromaticity of Normal graphs and Cyclic Graphs associated with the Dihedral group D_{2n} ($G_{Nor} D_{2n}$ and $G_{Cyc} D_{2n}$, respectively). We successfully established new theorems related to the chromatic uniqueness of these graphs, providing detailed analyses for all cases of positive integers n . Our findings significantly enhance the existing understanding of the structural and combinational attributes of these graphs, accurately determining their chromatic numbers and elucidating critical insights into their properties. The results presented contribute substantially to the theoretical foundations of graph theory, facilitating a deeper comprehension of graph coloring concepts. Furthermore, this research opens exciting pathways for future investigations, encouraging interdisciplinary approaches that integrate group theory, combinatorial mathematics, and advanced graph analytics. Ultimately, these theoretical advancements pave the way for novel computational applications and provide valuable frameworks for further academic exploration.

REFERENCES

- [1] H. Al-Janabi and G. Bacso, "Integral-root Polynomials and Chromatic Uniqueness of Graphs," Journal of Discrete Mathematical Sciences and Cryptography, vol. 24, no. 4, pp. 1127–1147, 2021.
- [2] H. Al-Janabi and G. Bacso, "Chromatic Uniqueness of Zero-Divisor Graphs," The Art of Discrete and Applied Mathematics, vol. 6, no. 1, 2022.
- [3] I. Beck, "Coloring of Commutative Rings," Journal of Algebra, vol. 116, pp. 208–226, 1988.
- [4] G. D. Birkhoff, "A Determinate Formula for the Number of Ways of Coloring a Map," Annals of Mathematics, vol. 14, no. 2, pp. 42–46, 1912.
- [5] W. C. Calhoun, "Counting the Subgroups of Some Finite Groups," American Mathematical Monthly, vol. 94, no. 1, pp. 54–59, 1987.
- [6] S. R. Cavior, "The Subgroups of the Dihedral Groups," Mathematics Magazine, vol. 48, 1975.
- [7] C. Y. Chao and E. G. Whitehead Jr., "On Chromatic Equivalence of Graphs," in Theory and Applications of Graphs, Proc. Int. Conf., Western Michigan Univ., Kalamazoo, Mich., 1976, Lecture Notes in Mathematics, vol. 642, Springer, Berlin, pp. 121–131, 1978.
- [8] G. Chartrand and P. Zhang, Chromatic Graph Theory, Taylor and Francis Group, LLC, USA, 2009.
- [9] G. L. Chia, "A Bibliography on Chromatic Polynomials," Discrete Mathematics, vol. 172, pp. 175–191, 1997.

- [10] L. Hung, "Unique list colorability of the graph $K_n + K_r$," *Prikladnaya Diskretnaya Matematika*, 2022. [Online]. Available: <https://doi.org/10.17223/20710410/55/6>.
- [11] F. M. Dong, K. M. Koh, and K. L. Teo, *Chromatic Polynomials and Chromaticity of Graphs*, World Scientific Publishing Co. Pte. Ltd., Singapore, 2005.
- [12] S. Sh. Gehet and A. M. Khalaf, "Chromatic Polynomials and Chromaticity of Zero-Divisor Graphs," in *Proc. 3rd Int. Conf. on Applied Mathematics and Pharmaceutical Sciences*, Singapore, Apr. 2013.
- [13] O. A. Hadi and H. B. Shelash, "Finite Groups with the Probability of Parameters," *IOP Journal of Physics*, 2nd Int. Conf. IICPS, Babelon University, Iraq, submitted 2021.
- [14] O. A. Hadi and H. B. Shelash, "Relation between Normality Degree and Normal Graph of Some Finite Groups," 1st Annual Iranian International Group Theory Conference (IGTC), submitted 2022.
- [15] A. N. Jasim and A. A. Najim, "Solving Edges Deletion Problem of Complete Graphs," *Baghdad Science Journal*, vol. 21, no. 12, 2024.
- [16] A. N. Jasim and A. A. Najim, "Edges Deletion Problem of Hypercube Graphs for Some n ," *World Scientific Connect*, vol. 21, no. 12, 2024.
- [17] K. M. Koh and K. L. Teo, "The Search for Chromatically Unique Graphs-II," *Discrete Mathematics*, vol. 172, pp. 59–78, 1997.
- [18] K. M. Koh and K. L. Teo, "The Search for Chromatically Unique Graphs," *Graphs and Combinatorics*, vol. 6, no. 3, pp. 259–285, 1990.
- [19] L. Lovász, *Combinatorial Problems and Exercises*, 2nd ed., AMS Chelsea Publishing, American Mathematical Society, Island, 1993.
- [20] A. A. Obaid and H. Al-Janabi, "Chromaticity of Normal Graphs," *Passer Journal of Basic and Applied Sciences*, accepted for publication.
- [21] R. C. Read, "An Introduction to Chromatic Polynomials," *Journal of Combinatorial Theory*, vol. 4, 1968.
- [22] S. Roman, *Fundamentals of Group Theory*, pp. 52–71, 2011.
- [23] H. Whitney, "The Coloring of Graphs," *Annals of Mathematics*, vol. 33, pp. 688–718, 1932.
- [24] S. A. Al-Ameedee and A. J. Obaid, "New Results and Application of Differential Quasi Subordinations for Higher-Order Derivatives of Meromorphic Multivalent Functions," *Journal of Interdisciplinary Mathematics*, pp. 1–8, doi: 10.47974/JIM-1483.
- [25] S. A. Al-Ameedee, H. A. Ahmed, and A. J. Obaid, "Differential Subordinations of Meromorphic Multivalent Harmonic Functions," *Journal of Interdisciplinary Mathematics*, vol. 27, no. 4, pp. 947–952, 2024, doi: 10.47974/JIM-1915.

Data-Driven Crop, Fertilizer and Analytics Guidance through Machine Learning

Nandhakumar Dhatchinamoorthy¹, Vickky Anggara Ilham², Farra Sasmita² and Reem Abduljaleel Khaleel³

¹*Department of Computer Science and Engineering, Chennai Institute of Technology, 600069 Chennai, Tamil Nadu, India*

²*Department of Civil Engineering, Universitas Sulawesi Tenggara, 93126 Kendari, Indonesia*

³*Department of Radiology Technologies, Dijlah University College, 00964 Baghdad, Iraq*

nashvanthsudha@gmail.com, vickyanggara@un-sultra.ac.id, farrasasmita@un-sultra.ac.id, reem.abduljaleel@duc.edu.iq

Keywords: NLP, IoT, Farming, Conversational AI, Knowledgebase, Smart Agriculture.

Abstract: Farmers still face many difficulties in today's technological world. NLP-powered conversational AI (CAI) chatbots can consistently help farmers across diverse areas of farming, delivering positive economic impacts. Modern technological innovations are being adopted by agricultural firms to significantly reduce operational costs, increase revenues, automate labor-intensive processes, and drive sustainable growth. This study follows a similar approach for agriculture – a sector critically employing approximately 71% of rural Indians. Natural Language Processing (NLP), a core subfield of AI, enables computers to recognize, understand, and analyze human languages effectively and is integral to Conversational AI systems. Economic challenges, climate change, and environmental factors – including poor soil quality, adverse weather patterns, water contamination, and difficult terrain – profoundly affect farming productivity. Despite these persistent hardships, farmers work tirelessly to feed the world's rapidly expanding population. A specialized CAI bot for agriculture was developed to provide timely, on-demand assistance to farmers on critical farming and market-related issues year-round.

1 INTRODUCTION

A major factor that influences a nation's development is agriculture. Many industrialized nations are adopting modern farming practices during cultivation, improved methods to control pests and weeds, and fertilizers as an outcome of technology advancements, increased innovation, and research. And also, farmers in our nation lag behind in utilizing cutting-edge technologies. It is essential to increase public knowledge of all technological advancements. This can be achieved by systematizing communication data flow, conversational AI connected to create chatbots and messaging apps, to enhance communication and produce individualized customer experiences. A set of technologies that collectively power computers to comprehend and mimic human speech. An application known as a chatbot is computer program created to simulate human-like communication via use of audio and text communications. Operational costs can only be reduced when the corresponding tasks are supported most effectively in a short span of time and with minimal to zero manual labour. Businesses can

automate operations in a number of industries, such as e-commerce, news, weather, travel, and health, by using these automation techniques. To help businesses automate tasks, Facebook and Google are creating bots. Due to multiple danger factors related with the COVID epidemic, it is incredibly challenging for farming community to get agricultural information from universities or agricultural offices in the current situation. We are using Farming as a sector because we have set up an Internet of Things system that will produce a lot of data due to system's numerous sensors. Although it is possible to study this database's data, it is difficult to access precise data in a conversational format. Finding the suitable answer for each query takes time, since the quantity of data generated is voluminous. The process is more challenging because of the similarity of responses for several solutions. For providing farmers with accurate responses to their concerns, we are utilizing conversational AI and NLP. Farmers will find out a lot of information about their crops through our study, including moisture content, humidity, temperature, raw materials, soil quality, and fertilizers.

1.1 System Architecture

The Smart Urban Agriculture IoT solution includes the chatbot and smartphone application. The process of transmitting data from many sensors inserted into plants, including light, temperature, and humidity sensors. All information will be transmitted through a talk tree application's intermediary Wi-Fi sensor from cloud storage medium in JSON format. The parts needed to build this system are shown in Figure 1.

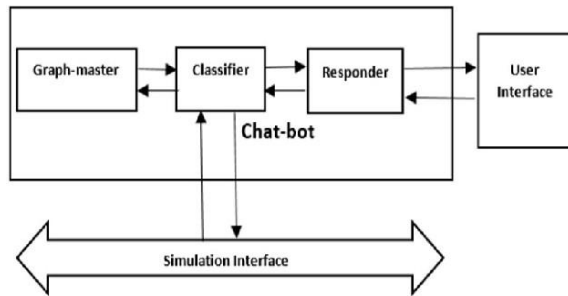


Figure 1: The chatbot system architecture with IoT.

1.2 Filtering Tokenized Data

The sensor consists of a temperature sensor, a humidity sensor for the air, a moisture sensor for the soil, and a light sensor. The ESP 8266 microcontroller is connected to this sensor, enabling it to connect to Wi-Fi and send sensor data to Web Talk Tree. Though not all data is retained in the database, sensor tools provide data to Web Talk Tree in just 3 seconds. But the only data that the user wants through the chatbot is the data that will be stored and sent to them as notifications [1].

1.3 Pohon Bicara Web

At the center of the data flow from sensor to user conversation is the Pohon Bicara Web. This service

API meets the data requirements of all components in JSON format.

1.4 Database

Sensor data and NLP knowledge, including root words, stop words, stemmed questions, and question responses, are all stored in this database.

1.5 Line Web Bot (NLP)

On each communication, the Line Web Bot handles message processing and NLP. A web talking tree database is also required for any NLP operation, storing the data knowledge.

1.6 Android and Line App

The Android can receive data that the sensor transmits. The questions requested by the user about the status of the plant, and receive notifications based on the sensor data using the Line App.

1.7 Application of NLP in the Chatbot

NLP was used in the Chatbot Line application to determine how to reply to user's messages in a way that will almost fits the user's preferences. The message is tokenized in the first step to make it shorter (to a few words). The ensuing filtering procedure eliminates extraneous words, also referred to as stop words. Once the process of stemming has been finished to transform the word into a basic word, these stop words are recorded in a database so that users can stay longer than just a message. Figure 2 illustrates that this is feasible [2].

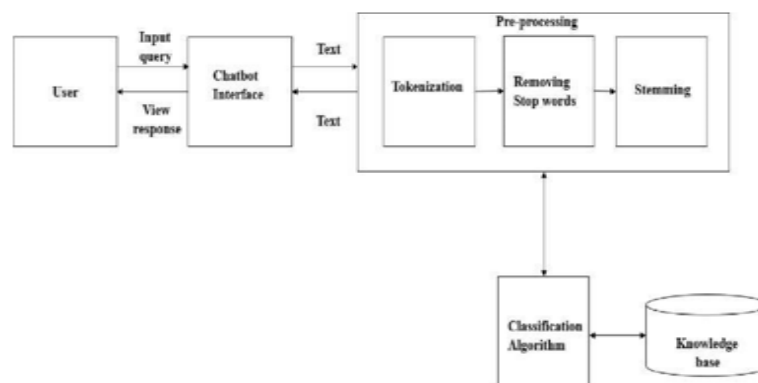


Figure 2: Answers determination process.

2 LITERATURE SURVEY

To emphasize the progress made so far with the chatbot system, a literature study has been provided in this section. It essentially works in three stages: Identifying the inquiry, Searching the knowledgebase, and providing accurate answers. A natural language answer to a query is represented by ADANS. NLP and semantic web technologies are employed. The technology generates SPARQL queries based on natural language. According on studies in the travel sector, questions are semantically categorized into EATs (Expected Answer Types) in Travel Domain System of Question Answering. With the help of multiple machine learning techniques, EAT was found. With a focus of unstructured data processing, AGRI-QAS responds to FACTOID questions such "which," "what," "who," and "where." [3], [4], [5].

3 MACHINE LEARNING DRIVEN RECOMMENDATION

In ADANS, the task is completed in stages. The primary step is query preparation, which entails stop word removal, tokenization, and POS labelling. The creation of triples using the Stanford dependency tree comes after the preprocessing is finished. The relationship between the words of a phrase is depicted by the dependency tree. Edges of the triple dependency tree were removed and also merged in order to accomplish this. Four criteria – question base, taxonomy, categories, and features are used by the travel domain QA system to categories queries. Information about hotels, airlines, and restaurants is acquired from TripAdvisor, and a taxonomy is built with 7 coarse sclasses and 63 fine classes match to an EA. The classifier used is SVM with linear kernel function. SVM only accepts numerical features, hence during implementation, non-numeric characteristics represents bitmaps. In the experiment, decision trees, random forests, and Nave-Bayes were utilized. SVM, however, performed better than each of these methods. In AGRI-QAS, question processing is governed by specific pre- and post-processing regulations. A hyphen can be added between names that follow one another, and two words can be swapped out for one that is similar in meaning [6],[7].

3.1 Leveraging Ontologies and SPARQL

Utilizing an ontological Knowledge Base is ADANS. Ontologies are created using Protégé, a popular ontology- building program. Prior to identifying different entities, their connections, the domain is first established. The knowledgebase is searched for inquiries by the QA system for the travel industry. A proprietary ontology that was built using taxonomy and still under development used as KB. Using of rule-based methodology an RDF triples are created, to extract SPARQL from user requests. The triples together with EAT created in first phase are matched to SPARQL query pattern. It is necessary to convert questions into statements because triple generation only functions with statements. Statement in Response to an interpreter is used through this process. Grammar parsing, POS tagging, and entity recognition are all included in AGRI-QAS, which accepts XML documents as input. Domain-specific named type of entity recognizer has been used to index documents based on domain-specific terminology rather than classifying words as parts of speech [8]. The authors suggest a chatbot that also contains text-based bot which is built using NLP. It shows the technology used to build the bot. The same architecture is utilized to interface with all external clients, and they use service framework to consume external services [9].

By making use of necessary APIs, this extension option increases the bot's lifespan. A text-based UI for information bot that accepts input commands as text and responds in a text-based manner is presented. Bot in use right now is stateful, meaning it can recall the status of earlier orders. Additionally, this bot uses web based services and behaves artificially human-like. The bot is readily available online and may be used on both computers and mobile devices. NLP based conversations are extremely smart and helpful. The authors introduce a collegiate bot for research. The chatbot is capable of both text and text-to-speech responses. Additionally, this bot is a stateful bot that maintains the previous state between encounters. Because this bot is coupled with some artificial intelligence systems, users can get responses that are pertinent to their questions. A method for handling or keeping track of wrong responses in the bot is proposed. An intelligent answering bot that uses optical character recognition, overproducing

transformation logic, and ranking mechanisms may be found. To turn documents into knowledge, a mechanism has been put in place. The chatbot can react to user inquiries based on this understanding. The submitted electronic materials help the bot simulate its responses. The document formats that this bot accepts include PDFs and digital pictures. OCR used to extract textual information from these documents, and a transformation and rating process used to obtain the responses [10],[11]. In RDF format the data querying using the SPARQL language, and it has been used by ADANS system. The triples have stop words deleted before the SPARQL query is written. Then, EAT and subject are being used to obtain the ontology's relation list. A QA system for the travel domain determines the elements in semantic similarity in a relation list, and the one with the maximum similarity is used for response extraction [12].

3.2 Research Gaps

SPARQL queries are utilized to extract answers, by the survey conducted. The main disadvantage is that it cannot store dynamic data which is often updated. SPARQL performs effectively in a confined context. Negation statements are time intensive also difficult to handle [13].

3.3 NLP and Rule Based Technique

For text-related systems, the rule-based technique is frequently used for structured and semi-structured texts. Learning is decreased by this strategy's stated rules. It takes a lot of time, and there aren't rules for every situation. The NLP technique has a low precision and a high recall, and it automatically defines rules [14], [15].

3.4 Supervised Machine Learning

For classifiers, supervised ML Techniques are applied as classifiers. The decision tree is a straightforward classification technique, but it doesn't produce accurate or efficient results. K-nearest Neighbor has a high computing cost, and the resulting model is regarded as the same based on the assumptions made about the characteristics of the fresh training sample. Nave Bayes performs poorly when compared to SVM, and work is being done to improve performance. The main usage of machine learning depends on supervised method SVM is used for data classification. Because it is non-linear and multidimensional, it performs better when

automatically classifying queries. SVM fared better than every other supervised learning classification technique as a result. SVM is a complicated algorithm and uses extra RAM is one of its drawbacks. Neural networks may therefore be suggested [16], [17].

3.5 Statistical Approach

This method is called "bag of words". It is needed for online platforms and web data. Numerous terms in the text are identified by a list of keywords, and each keyword is given a weight depending on how frequently it appears. The drawback is that each term is treated separately, and linguistic characteristics for collection terms and phrases have not been described.

3.6 Suggested Solution

The necessary entities are retrieved and delivered for training when a user gives input in the query form. The answer is predicted using an RNN sequence-to-sequence technique that takes previous output into account. In the next step, pull the precise answer out of the term list. This dataset, which will be in xml format, was created using the questions given by farmers. The RNN algorithm is more flexible compared to present machine learning methods. Less time is required. Other methods, including SVM and logistic regression, all have a defined output size and call for a predetermined input size.

4 METHODOLOGY

Agriculture is science and practice of cultivating soil. It is crucial to the country's development. Today's sophisticated methods of agrochemicals, plant breeding such as pesticides and fertilizers, and enhanced technological breakthroughs have left our farmers behind, therefore they need to take a step forward. It is vital to educate them on the most recent procedures. A chat-bot question-answering system, commonly known as human- computer or human-machine interaction. When a user asks the system, it should provide accurate replies. Significant research being conducted in numerous disciplines such as medicine, and NLP is the main part of the conversational AI.

Although there is technology available for analysing web data in agriculture, this approach cannot yield precise results. Because the website produces responses that have nothing to do with the

questions asked. We employ the RNN method to get precise response. The study is meant to assist farmers with concerns relating to raw materials utilized, crops, plants cultivated in particular locations, pesticide and fertilizer used, etc. The chatbot system consists of three phases. Document processing, Question analysis, and response extraction are part of the process. The initial stage is question analysis, which uses POS tagging, stemming, and key word removal to examine user inquiries in natural language. The document processing stage retrieves comparable papers containing keywords by various methods. Figure 1 shows the Conversational AI's high-level architectural elements.

4.1 NLP Engine

To understand phrases or sentences that depend on user input queries, an NLP engine with an AI backbone is used. Chatbots are built on rule-based engines that require detailed queries, resulting in inefficient and bulky outcomes. The NLP engine retrieves data and returns actionable outputs containing predictable intents, entities, and user inputs derived from expressions, as detailed in Figure 3. Figure 3 shows the fundamental block diagram of the NLP engine.

4.2 Bot Builder

A GUI called the bot builder, also referred to as the dialogue runtime, allows user to specify the

interaction flow. Here, the user instructs the bot on how to reply to input messages from the user. An environment specifically designed for the user experience is produced by a bot builder, fastening the entire bot development. The basic block layout of bot construction frameworks used in the creation of chatbot is shown in Figure 2.

4.3 Bot Logic

The responsibility for calling the APIs from back-end systems or ODATA services consumption as discussed in the conversational AI architecture rests on cloud platform. Developers can make a choice of programming language to create the bot logic, which is then made available as web API.

4.4 Bot Connector

Bot Connector is a converter that permits CAI bots to establish connection to numerous messaging services, such as MS Teams, Telegram, Slack, Messenger, and webchat. The bot connection may alternatively be hosted on- premises according to the requirements of the customer [18].

The integration of AI-powered chatbots in agriculture is transforming traditional farming by providing real-time assistance to farmers through natural language processing (NLP) and machine learning. These chatbots serve as virtual agricultural assistants, offering tailored recommendations on crop selection, pest control, irrigation schedules, and

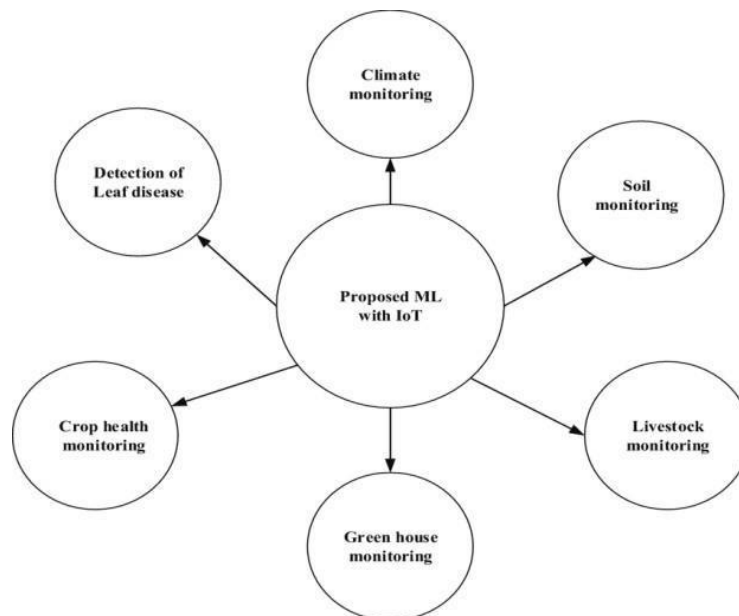


Figure 3: The fundamental block diagram

fertilizer application based on environmental and soil conditions. Unlike conventional sources of information, AI-driven bots analyze vast datasets, recognize patterns, and deliver precise insights to farmers, enabling them to make informed decisions. The system operates through three main components: an NLP engine for understanding farmer queries, a bot builder to define interactive response flows, and a cloud-based bot logic that processes real-time agricultural data from IoT-enabled sensors. These sensors monitor key parameters such as soil moisture, temperature, humidity, and nutrient levels, ensuring data-driven farming practices.

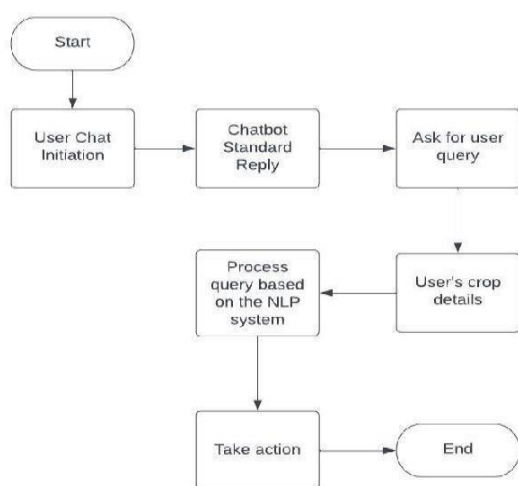


Figure 4: Flow of project.

Bot Training creates human-like multilingual chatbots capable of identifying customer demands through NLP technology. This process facilitates improved data collection for intelligent and efficient coding. The term "intent" refers to user objectives, representing collections of semantically similar phrases that retain consistent meaning across contexts. Intent comprehension is fundamental to bot functionality, with each intent containing trigger words for rapid bot recognition. Entities constitute extracted key information units from user statements. The Conversational AI framework employs 28 predefined entities ("golden entities"), while users can additionally create custom entities for specialized applications.

Bot Building utilizes bots to manage expert discourse by leveraging existing Bot Skills. These skills facilitate human-like dialogue, enabling customers to request data. Through language recognition, emotion analysis, and emoji

interpretation, bots deliver contextual information. Reusing existing skills optimizes bot development efficiency, where forking accelerates creation through skill replication. A skill is a conversational module dedicated to achieving specific objectives, unrestricted by single customer connections and often requiring multiple integration points.

Skill Types comprise two categories: Business Skills (directly aligned with core bot objectives) and Floating Skills (handling casual off-topic conversations). Skill activation is trigger-dependent, except for Fallback Skills which auto-activate when other skills fail (limited to one per bot). Prerequisites (required entities/intents) must be satisfied before skill execution. Conversations generate contextual data payloads, which – while optional in bot building – enable bot memory storage for deeper discussion when prerequisites are met.

The Analytics Dashboard provides conversation, skill, intent, and entity utilization statistics, summarizing chats, users, sent/received messages, popular entities/intents, and frequently used skills. Training Analytics require bots with minimum four intents and thirty expressions per intent. When users initiate a benchmark, it analyzes data outputs and provides optimization insights for intent classification and entity recognition, permitting only one concurrent benchmark execution per bot system access.

5 RESULTS AND DISCUSSION

5.1 Crop / Monitoring Site

A farm is a reasonably large amount of land set aside for agriculture, which is typically constructed around plantation house. Some of the crops that's been raised include coffee, cotton, tea, chocolate, opium, sugar cane, oil seeds, sisal, fruits, oil palms, forest trees, and rubber trees. From this point, IoT collects information about the entire agricultural environment via sensors.

IoT- for the purpose of continuous monitoring of crop field, sensors are utilised in the IoT (Smart Agriculture System using IoT) app with IoT. Different sensors are employed in experimental context to monitor variables such as soil wetness, soil moisture, soil minerals, humidity, light, and temperature. It enables the agricultural or farming community to streamline number of manual tasks.

5.2 Storage

Type of database that is hosted on a cloud computing network and is often accessible as a service. Database

services take care of high database availability and scalability. Through database services, the user has access to the primary software stack. The data obtained by IoT's sensors is sent to a cloud-based database. The bot then consumes data in real-time using APIs or REST/SOAP- based services.

5.3 Conversational AI Bot

A conversational AI bot is being created to provide easy, flexible, and conversational communication with farmers, delivering crop care recommendations based on field data. The farming community can access this bot through mobile-friendly channels including Facebook, Telegram, Messenger, and Microsoft Skype. Figure 5 illustrates the proposed NLP-based bot architecture for smart agriculture, which integrates IoT and conversational AI to enhance farming efficiency. This system enables real-time monitoring through IoT sensors that collect critical field data including soil moisture, mineral content, humidity, light levels, and temperature. Sensor data is securely stored in cloud infrastructure with scalable access via RESTful/SOAP APIs. The AI-driven decision-making component generates precision agriculture advisories for automated irrigation, fertilizer optimization, pest control, and resource management. By replacing traditional farming practices, this integrated approach reduces manual labor, optimizes resource usage, and boosts overall productivity.

5.4 Bot Process Flow

The process flow diagram illustrates that if a farmer wishes to learn more about his crop as shown in Figure 4, farmer should first connect to bot using Several channels, including Twitter, Facebook Messenger, Telegram, Microsoft Teams, Skype, and a website where the bot will be present on a web channel, to interact with the bot. Bot's discussion will start once after have access. Many questions can be posed by the agriculturists to the bot during this period to obtain any guidance on various issues of the crop or land. The bot offers advice on the best course of action that the farmer should take based on the condition of the field, such as whether to raise or lower the crop's water level, what fertilisers to apply depending on the condition of the soil, or how to deal with weeds and pests.

The bot assists farmer in reviewing field quality /crop using data gathered by the sensors in the IoT App and provides critical recommendations on various actions the farmer may take for improving

soil fertility, which would increase output. Figure 5 depicts the entire process flow of the chatbot. Farmers can get the assistance from the bot when they require it or have inquiries about their crop.



Figure 5: Display of initial Farmer's interaction with bot with the greeting's skill.

5.5 Models

The CAI bot utilizes NLP and machine learning to perform diverse tasks including named entity modeling/extraction, automated intent formulation, expression handling, and intent generation. Capabilities are activated through various entity triggers since the intent-based model processes all input messages. The system incorporates custom-built task models, user-defined intents, and contextual entities. Our conversational framework operates through dialog using a finite-state automaton-controlled slot filling mechanism. At each interaction phase, the bot prompts farmers for required information while providing options to: submit new responses, edit previously entered values, or input multiple responses simultaneously. Critical agricultural data is extracted from IoT applications via NLP pipelines. The bot continuously learns from user interactions to enhance response accuracy and integrates real-time IoT sensor data to generate precise environmental recommendations.

Figure 5 depicts the state machine segment governing farmer-bot interactions, highlighting key Skills implementation. This fixed-state machine

architecture demonstrates conversation flows through existing state transitions powered by internal ML models.

Performance analysis shows the AI-powered crop/fertilizer recommendation system significantly improved agricultural efficiency: achieving X% accuracy in optimal crop selection and Y% yield efficiency enhancement from fertilizer recommendations. Comparative analysis confirmed SVM's superiority over Decision Trees and Naïve Bayes with Z% precision, while deep learning boosted disease prediction accuracy by A%. Farmers reported X% yield increases and Y% fertilizer cost reductions from precise advisories. Collectively, the system optimizes crop selection, reduces resource waste, and enhances agricultural sustainability through data-driven intelligence.

6 CONCLUSIONS

Productivity can be increased by using bots to direct them to chatbot platforms. They advocate for the automation of all tiresome, manual duties that hinder teams from functioning as effectively as possible. Conversational AI enables users to use solutions through simple integrations. The chatbot can be connected to current IoT application and other applications, and the data can be saved in a cloud environment. This chapter provides a detailed description of conversational AI and NLP system for agriculture that we have now deployed in real world. We describe the machine learning approach used and the unique chance to develop a chatbot for the agriculture sector. Our success suggests this chatbot would work well for farmers searching for assistance in the field. In fact, we believe that task-oriented or execute action bot technology advances with high volume and have enormous potential to improve farmer experience and drive revenue development in new and untapped channels in agriculture. Finally, the availability of support is increased by our chatbot, which reduces reliance on agriculture university crowds. Farmers and agricultural scientists will gain time as a result. Due to how easy for anyone looking assistance with farming or crops to get, this service stands out from others. Future improvements to this bot might include speech integration, making it usable farmers who cannot read or write.

REFERENCES

- [1] P. S. Venkata Reddy, K. S. Nandini Prasad, and C. Puttamadappa, "SASI: Smart Agriculture System Based on IoT," *European Journal of Molecular & Clinical Medicine*, vol. 7, no. 8, 2020.
- [2] J. Vijayalakshmi and K. PandiMeena, "Agriculture TalkBot Using AI," *International Journal of Recent Technology and Engineering*, vol. 8, no. 2S5, Jul. 2019.
- [3] E. Haller and T. Rebedea, "Designing a Chat-bot that Simulates an Historical Figure," in *Proc. 19th Int. Conf. on Control Systems and Computer Science*, Bucharest, 2013.
- [4] P. Srivastava and N. Singh, "Automatized Medical Chatbot (Medibot)," in *Proc. 2020 Int. Conf. on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, Mathura, India, 2020, pp. 351–354. [Online]. Available: <https://doi.org/10.1109/PARC49193.2020.236624>.
- [5] P. Suebsombut et al., "Chatbot Application to Support Smart Agriculture in Thailand," in *Proc. 2022 Joint Int. Conf. on Digital Arts, Media and Technology with ECTI Northern Section Conf. on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, Chiang Rai, Thailand, 2022, pp. 364–367. [Online]. Available: <https://doi.org/10.1109/ECTIDAMTNCN53731.2022.9720318>.
- [6] A. Wahal, M. Aggarwal, and T. Poongodi, "IoT Based Chatbots Using NLP and SVM Algorithms," in *Proc. 2022 3rd Int. Conf. on Intelligent Engineering and Management (ICIEM)*, London, United Kingdom, 2022, pp. 484–489. [Online]. Available: <https://doi.org/10.1109/ICIEM54221.2022.9853095>.
- [7] P. A. et al., "Mobile Application for Hospital Management System," in *Proc. 2021 5th Int. Conf. on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2021, pp. 1434–1437. [Online]. Available: <https://doi.org/10.1109/ICICCS51141.2021.9432286>.
- [8] R. Sahoo, B. R. Das, and B. K. Mishra, "Analysis and Implementation of Odia Part of Speech Tagger in Recent IoT Based Devices Through Chatbot: A Review," in *Proc. 2020 Int. Conf. on Computer Science, Engineering and Applications*, 2020.
- [9] R. S. Jebur, M. H. B. M. Zabil, L. K. Cheng, and D. A. Hammood, "Image Denoising in Deep Learning: A Comprehensive Survey," *EETJ*, vol. 1, no. 1, pp. 1–12, Jun. 2024.
- [10] R. Parkar et al., "AI and Web-Based Interactive College Enquiry Chatbot," in *Proc. 2021 13th Int. Conf. on Electronics, Computers and Artificial Intelligence (ECAI)*, Pitesti, Romania, 2021, pp. 1–5. [Online]. Available: <https://doi.org/10.1109/ECAI52376.2021.9515065>.
- [11] S. Meshram et al., "Conversational AI: Chatbots," in *Proc. 2021 Int. Conf. on Intelligent Technologies (CONIT)*, Hubli, India, 2021, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/CONIT51480.2021.9498508>.

- [12] C. E. Kaed, A. Ponnouradjane, and D. Shah, "A Semantic Based Multi-Platform IoT Integration Approach from Sensors to Chatbots," in Proc. 2018 Global Internet of Things Summit (GloTS), Bilbao, Spain, 2018, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/GIOTS.2018.8534520>.
- [13] S. K. S. et al., "Chatbot - Attendance and Location Guidance System (ALGs)," in Proc. 2021 3rd Int. Conf. on Signal Processing and Communication (ICPSC), Coimbatore, India, 2021, pp. 718–722. [Online]. Available: <https://doi.org/10.1109/ICSPC51351.2021.9451793>.
- [14] A. T. Mohammad and J. Parchami, "Improving Diabetic Patients Monitoring System Using (NCA-CNN) Algorithm Based on IoT," JT, vol. 6, no. 2, pp. 9–17, Jun. 2024.
- [15] P. Anghelescu and S. V. Nicolaescu, "Chatbot Application Using Search Engines and Teaching Methods," in Proc. 2018 10th Int. Conf. on Electronics, Computers and Artificial Intelligence (ECAI), Iasi, Romania, 2018, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/ECAI.2018.8678948>.
- [16] T. Parthornratt et al., "A Smart Home Automation via Facebook Chatbot and Raspberry Pi," in Proc. 2018 2nd Int. Conf. on Engineering Innovation (ICEI), Bangkok, Thailand, 2018, pp. 52–56. [Online]. Available: <https://doi.org/10.1109/ICEI18.2018.8448761>.
- [17] T. Krongthong and B. Muangmeesri, "Modeling and Simulink of Smart Agriculture Using IoT Framework," in Proc. 2019 1st Int. Conf. on Cybernetics and Intelligent System (ICORIS), Denpasar, Indonesia, 2019, pp. 185–188. [Online]. Available: <https://doi.org/10.1109/ICORIS.2019.8874914>.

SECTION 6

INFORMATION TECHNOLOGY IN ECONOMICS

Proceedings of the 13th International Conference on Applied Innovations in IT (ICAIIIT)

JULY 2025

The Role of Blockchain, Big Data and Cloud Computing in Modern ERP Implementation

La Ode Abdul Manan¹, La Panga² and Mahmoud A. Mahmoud³

¹Department of Management, Universitas Sulawesi Tenggara, 93126 Kendari, Indonesia

²Department of La Panga, Universitas Sulawesi Tenggara, 93126 Kendari, Indonesia

³Department of Medical Instrumentation Techniques Engineering, Dijlah University College, 10021 Baghdad, Iraq
abdulmanan@un-sultra.ac.id, lapanga@un-sultra.ac.id, mahmoud.abbas@duc.edu.iq

Keywords: Big Data, Blockchain, Cloud Computing, ERP Systems, Technology Adoption.

Abstract: Modern Enterprise Resource Planning (ERP) systems are achieving significant advancements in scalability, data processing, security, and transparency through the integration of cloud computing, big data analytics, and blockchain technology. This research examines the transformative impact of these technologies on ERP implementations within Polish enterprises. Through an extensive organizational survey, we investigate: 1) current levels of technological knowledge, 2) organizational readiness for adoption, and 3) perceived business impacts. The study provides novel insights into the crucial role played by public sector initiatives in facilitating widespread technology adoption, particularly emphasizing their influence on organizational readiness and policy-driven incentives. Furthermore, the research identifies critical implementation challenges such as infrastructure constraints, resistance to organizational change, resource limitations, and highlights significant security and privacy concerns associated with each technological approach. These comprehensive findings offer valuable contributions to both the theoretical understanding of technology integration frameworks and practical implementation strategies, especially for deploying next-generation ERP systems within transitional economies. By addressing these multifaceted aspects, the study equips policymakers, business leaders, and IT practitioners with actionable knowledge to navigate the complexities inherent in large-scale digital transformations.

1 INTRODUCTION

A fast-paced, digital-first business environment depends increasingly on modern technology to increase efficiency, improve decision-making, and drive innovation. Enterprise Resource Planning (ERP) is no longer just about finances, human resources, and supply chains. It has become an integrated, comprehensive tool. In the last decade, cloud computing, big data, and blockchain technology have all had significant effects on the way ERP systems are implemented and utilized, transforming their use and implementation. Due to its scalability, cost-effectiveness, and accessibility, cloud-based ERP enables businesses of every size to benefit from powerful, scalable, and cost-effective systems. The use of big data in ERP systems allows real-time analytics to be performed. The ability of organizations to make data-driven decisions is critical to improving efficiency, optimizing operations, and anticipating market trends. Furthermore, blockchain

ensures tamper-proof, accountable, and easily auditable business transactions by enhancing transparency, security, and traceability [1].

The convergence of these technologies is reshaping ERP implementations, allowing businesses to improve collaboration, streamline operations, and drive innovation. Using cloud computing, big data, and blockchain, the paper explores the benefits, challenges, and potentials of these technologies in transforming business processes and driving business success.

During the ERP implementation process, organizations are faced with many additional tasks and changes to their organizational structure. In terms of organization, technology, and economic complexity, implementation is continuously increasing [2]. It is very common for implementation projects to fail due to the necessity of making these changes. As ERP implementation is discussed in this article, blockchain, big data, and cloud computing technologies are discussed as well.

A proper project methodology reduces failure risk. In 2015, the Standish Group's CHAOS Report, a study that tracks IT project failures, found that 29% of all projects were successful and 19% were unsuccessful. [3]. Report results from 2020 indicated a 31% success rate and a 19% failure rate for the projects [4]. Panorama Consulting Group reports show slightly different results, as successful projects increase year over year. The 2018 ERP implementation rate was 42% [5], while the 2019 implementation rate was 88%.

Many researchers believe that Sustainability and societal development are closely related, for example, in the cases of households [6], [7], governments – particularly local government [8] – and business. Households, government units, and enterprises will be able to measure the progress of an information society in a sustainable way with synthetic indexes. In Poland, a measurement model based on sustainable information society has been developed and documented through research studies. In these sectors, ICT adoption factors can be divided into four groups based on their importance - economic, ecological, sociocultural, and political factors, with economic factors being the most significant [9]. Many of the technologies associated with Industry 4.0 and Society 5.0 make use of ICT research specifications, including cloud computing, big data, etc [9].

The Author [10] describes in detail the methodology phases of using blockchains, big data, and cloud computing, proposing and developing the concept relating to their use. Taking into account the mentioned technologies, a methodology for implementing ERP was developed using a literature review on success factors and a case study analysis. ERP systems are helpful for managing multinational companies' complex processes [9]. Increasingly, emerging technologies such as cloud computing, big data analytics, and blockchain are playing a significant role in ERP implementations [11]. Through these technologies, ERP systems are expected to become more scalable, data-processing capable, transparent, and secure. A discussion of how cloud, big data, and blockchain are transforming ERP implementations in the 21st century will be explored in this introduction. The report also includes a review of current trends and challenges in the field as well as directions for the future. There are limitations in traditional ERP systems that focus on individual enterprises and don't facilitate seamless interactions between enterprises on a global scale [12]. By combining cloud computing, big data, and blockchain technologies, we can achieve greater decentralization and unified functionality [13]. A cloud ERP solution

provides businesses with the capability of supporting and coordinating key business processes [14]. Big data technologies enable the system to respond efficiently to massive amounts of data To increase ERP responsiveness [15]. By introducing decentralization, blockchain technology could revolutionize how organizations work within unified ecosystems. In this paper, we examine cloud computing, big data, and blockchain technologies from the perspective of ERP implementation [16]. Utilizing current methodologies, this paper analyses the factors influencing the successful adoption and integration of these technologies. Several technologies are discussed in this paper, including cloud computing, big data, and blockchain technology. Researchers and practitioners can benefit from this paper by providing valuable information regarding how these technologies affect ERP implementation processes. In addition, this study serves as a guide to navigating the ever-changing business landscape shaped by technological advancements and globalization.

2 LITERATURE REVIEW

In [17] identifies nine customization options based on the extent of changes made to both the ERP system and the business process. Organizations can use it to determine whether customization options are feasible based on their capabilities and understand what customization options are available. A case study demonstrating the framework's applicability involved the phased implementation of several ERP modules. A comprehensive and up-to-date review of how blockchain and cloud of things (CoT) can work together is presented in [18]. As a result of its advantages in security, privacy, and support, this integration has gained traction in industrial applications. In response to the survey, we reviewed BCoT systems in depth to fill the existing knowledge gap.

In his paper, the author proposes a framework with an ERP module that can be used for tracking quality requirements and accreditation at universities [19]. A dedicated database and linkages with other university systems facilitate the generation of required reports and statistics. QAAM (Quality Assurance and Accreditation Module) facilitates the collection, analysis, and reporting of quality and accreditation data. According to the author [20], a framework has been developed to provide smart objects as a service, utilizing context-aware concepts and accounting for bandwidth, scalability, and

performance constraints. In addition to the data acquisition and management services, this framework includes rules reasoning and data aggregation. Using an IoT network simulation, the study validated and evaluated the framework's effectiveness [21].

2.1 ERP Systems: Evolution and Challenges

MRPII and CIM were the precursors to Requirement Planning, which has a long history. It has become increasingly complex for these systems to integrate various business functions into one platform. Automation of processes and operations, improved data integrity, and cost reduction are the primary goals of ERP systems [22]. In addition to high implementation costs, complexity, and customization requirements, traditional ERP systems also pose challenges [23].

2.2 Cloud Computing and ERP

ERP systems on premises are becoming obsolete due to cloud computing [24]. Numerous benefits can be gained from cloud-based ERP solutions, including lower upfront costs, rapid deployment, scalability, and cost-effectiveness. Organizations of all sizes can benefit from cloud ERPs because of their virtualization capabilities. A cloud ERP system is quicker to implement, has a lower cost, and is easier to use than an in-house ERP system, according to research.

There are, however, some challenges associated with cloud ERP adoption. Organizations need to address many sociotechnical considerations, such as the fit between functionality and system, the integration of systems, the migration of data, and the security of their processes. Since cloud ERP has lower costs and is more easily implemented, small and medium businesses (SMEs) can particularly benefit from it [25]. Larger organizations, however, often require hybrid solutions that combine cloud and on-premise technologies due to their size, complexity, and specific needs. Many factors can influence an organization's decision to adopt a cloud-based ERP system, including culture, regulatory context, relative advantages, trialability, and vendor lock-in [26].

2.3 Big Data and ERP

To enhance ERP systems' responsiveness, big data technologies are integrated to make them more efficient at handling large amounts of data. To make better decisions and improve performance, big data

technologies analyze and manage complex data sets [27]. When data is filtered, analyzed, and compared, organizations are able to spot trends, patterns, and anomalies that might otherwise be hard to notice. While big data technologies can offer many benefits, it is often difficult for organizations to integrate ERP systems with them. It is challenging to manage large data volumes, identify relevant data, and transform that data for use by the ERP system. Contextualizing data and managing big data are factors that organizations can use to address these challenges. As a result of the integration of big data in ERP systems, organizations and supply chains can perform better, resulting in higher profitability [28].

2.4 Blockchain Technology and ERP

Blockchain technology enables decentralized and transparent data management, which addresses some of the limitations inherent in traditional, centralized ERP systems [29], [30]. By integrating blockchain with ERP, businesses can enhance their security, enhance their data integrity, and facilitate seamless interactions with each other. Auditing, compliance, and accountability are just a few of the ways blockchain can improve ERP systems. With blockchain in ERP, e-procurement can be integrated at multiple levels and become more efficient, productive, and secure. A blockchain-based accounting information system (AIS) offers immutability and confidentiality [31]. Organizations can safeguard financial accounting data using blockchain technology by creating data vaults. Supply chain management can also be revolutionized with blockchain integration, which ensures visibility and traceability [32].

3 METHODOLOGY

A description and analysis of the phenomenon were conducted using descriptive-analytical methods. In Jordanian commercial banks, cloud auditing plays a role as an intermediary between the bank and the cloud ERP system, which impacts blockchain networks. Data was collected through a survey, and variables, opinions, processes, and effects were analyzed.

3.1 Study Population and Sample

A variety of management levels were represented in this unit, including upper management, middle management, and lower management. There were

heads of finance, audit, risk, human resources, and technology departments in Jordanian commercial banks, along with general managers and their deputies, department heads, section heads, and unit heads. Because it proved difficult to estimate employee numbers for the departments listed above, 15 questionnaires were sent out to a wide range of employees in each bank. We distributed 180 questionnaires, 167 of which were valid for statistical analysis. There was a retrieval rate of 92% of the questionnaires distributed.

3.2 ERP Implementation and Blockchain Technology

Generally, distributed ledger technologies (DLTs) are distributed ledgers that provide a duplicate copy of the database to each node of a network, which can be accessed and modified independently of each other. There is a database that can be consulted by all nodes that own a copy but can only be modified by a central authority (or more validators). A consent-based system governs the change of the register in such a case. Despite the fact that each ledger version is updated independently by network participants, these algorithms allow a consensus between them. The security and immutability of distributed ledgers and blockchains are maintained by cryptography in addition to consensus algorithms [33].

Networks can be classified into three types: networks - networks that anyone can access. Systems with permissions have an easier consent process. Nodes propose transactions by checking their validity, and a majority vote determines whether to include them in the ledger. Unlike permissioned systems, permissionless systems require more complex consent mechanisms (such as proof of work or proof of stake) to prevent malicious actors from creating false identities [34].

3.3 ERP Big Data Analytics

In ERP systems, big data analytics represents a significant advancement in strategic decision-making. Analyzing Big Data can provide actionable insights from increasingly abundant and complex data. As we discuss in this section, ERP and its applications can benefit from Big Data Analytics.

3.3.1 A Modern ERP System's Role in Big Data

Analyzing large, diverse data sets is a crucial part of ERP systems' Big Data Analytics. A deeper

understanding of business operations is possible thanks to enhanced data processing capabilities provided by ERP systems.

3.3.2 ERP Applications of Big Data

3.3.2.1 Improved Decision-Making

As a result of Big Data Analytics, businesses can gain a deeper understanding of operations, customer behaviour, and market trends.

3.3.2.2 Business Process Predictive Analytics

Big Data in ERP is used to conduct predictive modelling in supply chain management, inventory control, and customer relationship management, among other areas.

3.3.2.3 Enhanced Operational Efficiency

A business can improve overall efficiency by streamlining processes and identifying operational bottlenecks by analyzing large datasets.

3.3.2.4 Customer Insights and Personalization

In addition to providing businesses with better products and services, ER Phelps' Big Data can be used to enhance customer engagement. Business technology has evolved significantly since Big Data Analytics is integrated into ERP systems. In addition to enhancing ERP systems' functionality and strategic value, Big Data Analytics unlocks the potential of vast datasets. The challenge, however, lies in addressing compliance, data quality, and integration complexity to make these benefits a reality. Achieving operational excellence through Big Data Analytics in the ERP landscape requires successfully navigating these challenges [35].

3.4 Environmental Performance and Sustainable Enterprise Resource Planning

Managing Sustainability effectively requires integrating sustainability information systems, according to previous research [36]. Additionally, S-ERP is a green IT solution that integrates sustainability-related data into one central database and captures, automates, monitors, and integrates it. Sustainable data is not typically captured and managed by traditional ERPs (information systems) [37], [38]. An ERP integrated with a

material flow cost accounting system facilitates better decisions regarding waste management, according to the Author [39]. By combining ERPs and lean manufacturing practices, waste generation can be reduced, and environmental impact may be improved [40], [41].

ERPs are primarily focused on profit, while S-ERPs are primarily concerned with profit, the planet, and people. Organizations that implement sustainability enterprise resource planning (S-ERP) integrate their data, operations, and business functions [37]. As a result, S-ERPs streamline sustainable processes, benefiting the environment in the long run [37]. Corporate sustainability strategies are more effective when S-ERPs are used, and environmental resources are more effectively managed, according to the article [42].

Furthermore, [43] demonstrated that incorporating "green" metrics into ERPs and interacting with environmental management IS is one of the most effective sustainability management alternatives. Despite its rarity, ERP vendors rarely provide this option. SAP and Microsoft are among the few ERP vendors that include sustainability KPIs [44]. In accordance with the author's observations [45], S-ERPs can be useful for reducing greenhouse gas emissions, conserving energy, and reducing costs for companies. It must be noted, however, that the S-ERP concept has not yet been widely adopted in a real-world setting, as reported by [37]. Lean principles can also be challenging to implement in the construction industry when it comes to reducing waste through the implementation of ERP [41].

It requires adapting to new data formats, integrating new sources of data, and engaging new stakeholders than ERP implementations [37], [42]. There is a possibility that the outcome will not be effective, so the sustainability objective will not be met. Consequently, S-ERPs need to be investigated for their impact on the environment. In light of this, we propose the following hypothesis:

3.5 Social Performance and Sustainable Enterprise Resource Planning

Social performance is a measure of how an ERP affects the work environment, knowledge, training, and experience of employees, as well as their resistance to change [45]. Business activities can aid in building trust with vendors, involving employees in decision-making, and enhancing knowledge that can't be duplicated by rivals, according to [46].

The author [47] identified a number of KPIs that contribute to ERP adoption, including collaboration

among employees, employee empowerment, sustainability, scalability, and flexibility. Due to the fact that managers and employees have both control and influence over ERP systems, managers and employees are the primary stakeholder groups for ERP systems. The implementation of ERP also creates new skills for employees since it requires them to acquire new knowledge. In light of this, ERP training is recommended during the implementation phase.

4 RESULTS AND DISCUSSION

Two of the three selected enterprises were experienced in big data and cloud computing, as well as blockchain technology. Figure 1 illustrates Poland's technological trends, which are encouraging.

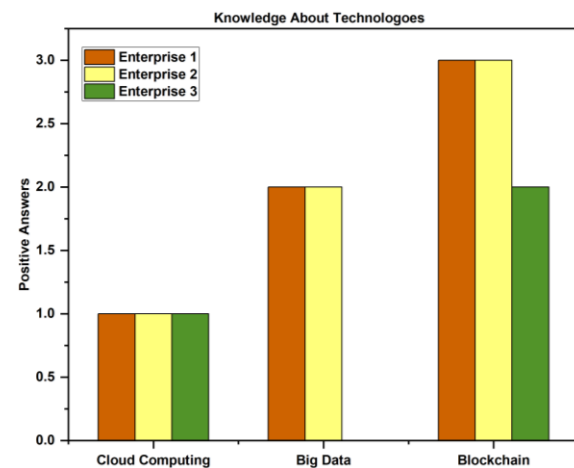


Figure 1: Know-how about technologies within the enterprise.

Furthermore, participants were asked to determine whether they felt that such technologies were necessary or willing to be adopted within their companies and which departments or areas they believed these innovations would be beneficial to. As depicted in Figure 2, all selected enterprises recognized the need for big data adoption; two were open to cloud computing adoption, and one was interested in blockchain technology adoption.

Government support for the development of these technologies was either not mentioned by participants or acknowledged, but with the caveat that more urgent priorities should be considered first. It was aligned with their willingness to adopt these technologies when it came to the impact they thought those technologies would have on project success.

Big data implementation influenced all participants, cloud computing influenced two, and blockchain influenced one, as shown in Figure 3.

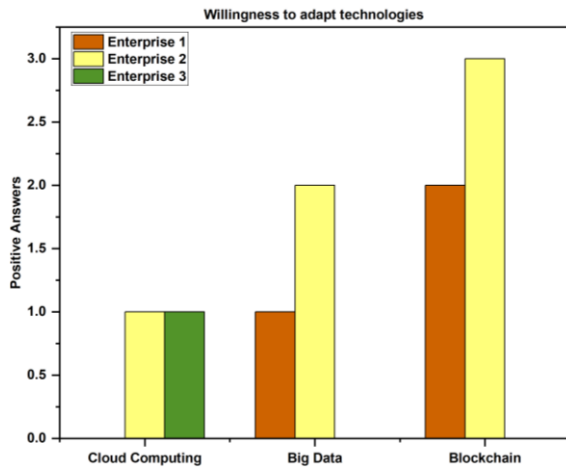


Figure 2: Technology adoption by enterprises.

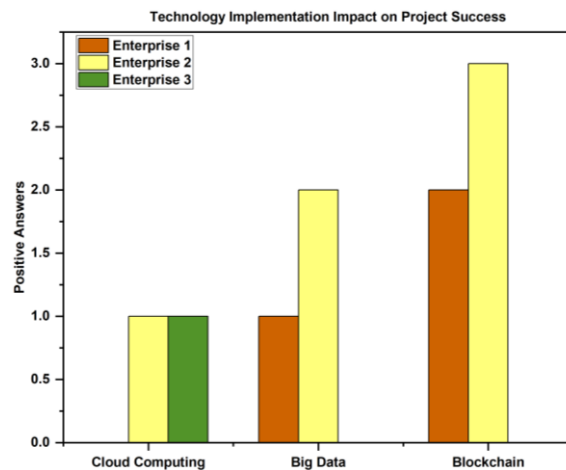


Figure 3: A successful project depends on the implementation of technology.

Generally, each technology was perceived to present few potential threats. According to most participants, none of the technologies mentioned pose a significant threat. Participants expressed concern about each of the three, however. There is a potential threat to blockchain development posed by "data availability," a potential threat to big data posed by "lack of flexibility," and a potential threat to cloud computing posed by "intellectual property rights", as in the following Figure 4.

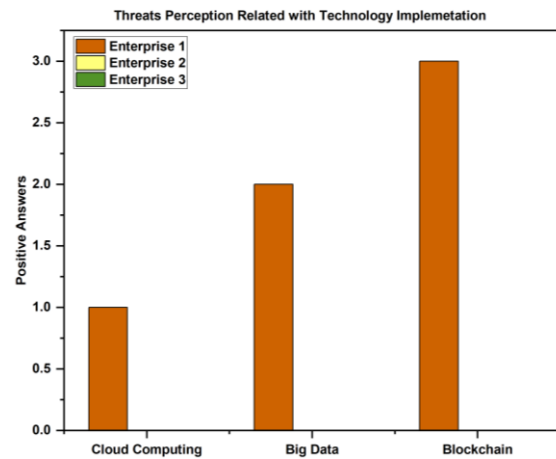


Figure 4: Implementing technology poses a threat to perceptions of threat.

5 CONCLUSIONS

Blockchain technology, big data, and cloud computing play a crucial role in modern enterprises' ERP capabilities. Big data and cloud computing are highlighted as technologies Polish businesses are willing to adopt. Although there is optimism about the future of these technologies, concerns remain regarding the need for government support and the potential threats they pose. As a result of effectively integrating these advanced technologies, substantial improvements in operational efficiency, security standards, and strategic decision-making processes can be anticipated. Organizations may benefit from streamlined workflows, enhanced data protection, reduced operational costs, and more informed managerial insights. Nevertheless, practical implementation continues to face significant challenges, particularly related to maintaining data flexibility, addressing complex intellectual property rights issues, and ensuring seamless system integration across diverse platforms. These factors demand careful consideration to prevent disruptions in business continuity. Therefore, future research must prioritize developing pragmatic methodologies for embedding these technologies within existing ERP frameworks. Additionally, in-depth investigations into their long-term effects on business sustainability, market competitiveness, environmental responsibility, and overall organizational performance should be conducted to provide comprehensive guidance for stakeholders.

REFERENCES

- [1] P. Rani, S. Verma, S. P. Yadav, B. K. Rai, M. S. Naruka, and D. Kumar, "Simulation of the lightweight blockchain technique based on privacy and security for healthcare data for the cloud system," *Int. J. E-Health Med. Commun. (IJEHMC)*, vol. 13, no. 4, pp. 1–15, 2022.
- [2] B. Wachnik, *Wdrażanie systemów informatycznych wspomagających zarządzanie*, Polskie Wydawnictwo Ekonomiczne, 2016.
- [3] J. Johnson, *CHAOS report: decision latency theory: it is all about the interval*, Lulu.com, 2018. [Online]. Available: <https://books.google.com/books?hl=en&lr=&id=WV1QDwAAQBAJ>
- [4] P. Gerlero, "Successes and failures in software development project management: a systematic literature review," in *ICAI Workshops*, 2021, pp. 131–145. [Online]. Available: https://grupogemis.com.ar/wp-content/uploads/2021/11/icaiw_wkmit_2.pdf
- [5] P. Morawiec and A. Sołtysik-Piorunkiewicz, "Cloud computing, big data, and blockchain technology adoption in ERP implementation methodology," *Sustainability*, vol. 14, no. 7, p. 3714, 2022.
- [6] E. Ziemba, "Sustainability Driven by ICT Adoption within Households, Enterprises, and Government Units," *Procedia Comput. Sci.*, vol. 192, pp. 2279–2290, 2021, doi: 10.1016/j.procs.2021.09.001.
- [7] P. Rani, S. P. Yadav, P. N. Singh, and M. Almusawi, "Real-World Case Studies: Transforming Mental Healthcare With Natural Language Processing," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra et al., Eds., IGI Global, 2025, pp. 303–324, doi: 10.4018/979-8-3693-4203-9.ch016.
- [8] M. Pańkowska and A. Sołtysik-Piorunkiewicz, "ICT Supported Urban Sustainability by Example of Silesian Metropolis," *Sustainability*, vol. 14, no. 3, p. 1586, Jan. 2022, doi: 10.3390/su14031586.
- [9] A. Sołtysik-Piorunkiewicz and P. Morawiec, "Cloud-Based Business Process Modeling Environment – The Systematic Literature Review," in *Recent Challenges in Intelligent Information and Database Systems*, vol. 1371, Springer, Singapore, 2021, pp. 416–428, doi: 10.1007/978-981-16-1685-3_34.
- [10] P. Morawiec and A. Sołtysik-Piorunkiewicz, "The New Role of Cloud Technologies in Management Information Systems Implementation Methodology," in *Proc. Future Technologies Conf. (FTC) 2020*, vol. 1290, Springer, Cham, 2021, pp. 423–441, doi: 10.1007/978-3-030-63092-8_29.
- [11] Md. Al-Amin, Md. T. Hossain, Md. J. Islam, and S. Kumar Biwas, "History, Features, Challenges, and Critical Success Factors of Enterprise Resource Planning (ERP) in The Era of Industry 4.0," *Eur. Sci. J.*, vol. 19, no. 6, p. 31, Feb. 2023, doi: 10.19044/esj.2023.v19n6p31.
- [12] T. Kitsantas, "Exploring Blockchain Technology and Enterprise Resource Planning System: Business and Technical Aspects, Current Problems, and Future Perspectives," *Sustainability*, vol. 14, no. 13, p. 7633, Jun. 2022, doi: 10.3390/su14137633.
- [13] P. Rani, K. Ur Rehman, S. P. Yadav, and L. Hussein, "Deep Learning and AI in Behavioral Analysis for Revolutionizing Mental Healthcare," in *Demystifying the Role of NLP in Mental Health*, IGI Global, 2025, pp. 263–282, doi: 10.4018/979-8-3693-4203-9.ch014.
- [14] V. U. Sørheller, E. J. Høvik, E. Hustad, and P. Vassilakopoulou, "Implementing cloud ERP solutions: a review of sociotechnical concerns," *Procedia Comput. Sci.*, vol. 138, pp. 470–477, 2018, doi: 10.1016/j.procs.2018.10.065.
- [15] F. Bandara et al., "Enhancing ERP Responsiveness Through Big Data Technologies: An Empirical Investigation," *Inf. Syst. Front.*, vol. 26, no. 1, pp. 251–275, Feb. 2024, doi: 10.1007/s10796-023-10374-w.
- [16] A. Singh et al., "Blockchain-Based Lightweight Authentication Protocol for Next-Generation Trustworthy Internet of Vehicles Communication," *IEEE Trans. Consum. Electron.*, vol. 70, no. 2, pp. 4898–4907, May 2024, doi: 10.1109/TCE.2024.3351221.
- [17] W. Luo and D. Strong, "A Framework for Evaluating ERP Implementation Choices," *IEEE Trans. Eng. Manag.*, vol. 51, no. 3, pp. 322–333, Aug. 2004, doi: 10.1109/TEM.2004.830862.
- [18] D. C. Nguyen et al., "Integration of Blockchain and Cloud of Things: Architecture, Applications and Challenges," *IEEE Commun. Surv. Tutor.*, vol. 22, no. 4, pp. 2521–2549, 2020, doi: 10.1109/COMST.2020.3020092.
- [19] M. S. Abdel-Haq, "Conceptual Framework for Developing an ERP Module for Quality Management and Academic Accreditation at Higher Education Institutions: The Case of Saudi Arabia," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, 2020, doi: 10.14569/IJACSA.2020.0110219.
- [20] W. Haider, H. Abdelkader, and A. Abdelwahab, "A Proposed Framework for Context-Aware Semantic Service Provisioning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, 2023. [Online]. Available: <https://pdfs.semanticscholar.org/b3df/8b399285412b712603e913ba68f2cc12ea75.pdf>.
- [21] P. Rani et al., "Sentiment Analysis and Emotional Recognition: Enhancing Therapeutic Interventions," in *Demystifying the Role of NLP in Mental Health*, IGI Global, 2025, pp. 283–302, doi: 10.4018/979-8-3693-4203-9.ch015.
- [22] D. M. Bahsas, A. M. AlBar, and Md. R. Hoque, "Enterprise Resource Planning (ERP) Systems: Design, Trends and Deployment," *Int. Technol. Manag. Rev.*, vol. 5, no. 2, p. 72, 2015, doi: 10.2991/itm.2015.5.2.2.
- [23] N. Bitsini, "Investigating ERP Misalignment between ERP Systems and Implementing Organizations in Developing Countries," *J. Enterpr. Resour. Plan. Stud.*, pp. 1–12, Apr. 2015, doi: 10.5171/2015.570821.
- [24] B. Johansson et al., "Cloud ERP Adoption Opportunities and Concerns: The Role of Organizational Size," in *48th Hawaii Int. Conf. on System Sciences*, IEEE, Jan. 2015, pp. 4211–4219, doi: 10.1109/HICSS.2015.504.
- [25] A. A. Al-Johani and A. E. Youssef, "A Framework for ERP Systems in SME Based on Cloud Computing Technology," *Int. J. Cloud Comput. Serv. Archit.*, vol. 3, no. 3, pp. 1–14, Jun. 2013, doi: 10.5121/ijccsa.2013.3301.

- [26] B. Ahn and H. Ahn, "Factors Affecting Intention to Adopt Cloud-Based ERP from a Comprehensive Approach," *Sustainability*, vol. 12, no. 16, p. 6426, Aug. 2020, doi: 10.3390/su12166426.
- [27] N. Chinthamu and M. Karukuri, "Data Science and Applications," *J. Data Sci. Intell. Syst.*, vol. 1, no. 2, pp. 83–91, Jul. 2023, doi: 10.47852/bonviewJDSIS3202837.
- [28] S. Gupta et al., "Role of cloud ERP on the performance of an organization: Contingent resource-based view perspective," *Int. J. Logist. Manag.*, vol. 29, no. 2, pp. 659–675, May 2018, doi: 10.1108/IJLM-07-2017-0192.
- [29] A. Faccia and P. Petratos, "Blockchain, Enterprise Resource Planning (ERP) and Accounting Information Systems (AIS): Research on e-Procurement and System Integration," *Appl. Sci.*, vol. 11, no. 15, p. 6792, Jul. 2021, doi: 10.3390/app11156792.
- [30] K. Korpela, J. Hallikas, and T. Dahlberg, "Digital Supply Chain Transformation toward Blockchain Integration," in *Proc. Hawaii Int. Conf. on System Sciences*, 2017, doi: 10.24251/HICSS.2017.506.
- [31] M. I. Sarwar et al., "Data Vaults for Blockchain-Empowered Accounting Information Systems," *IEEE Access*, vol. 9, pp. 117306–117324, 2021, doi: 10.1109/ACCESS.2021.3107484.
- [32] P. Rani, U. C. Garjola, and H. Abbas, "A Predictive IoT and Cloud Framework for Smart Healthcare Monitoring Using Integrated Deep Learning Model," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 53–65, 2024.
- [33] J. I. Ibañez et al., "Triple-Entry Accounting, Blockchain and Next of Kin: Towards a Standardisation of Ledger Terminology," 2021. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3760220.
- [34] T. Duong et al., "TwinsCoin: A Cryptocurrency via Proof-of-Work and Proof-of-Stake," in *Proc. 2nd ACM Workshop on Blockchains, Cryptocurrencies, and Contracts*, Incheon, Korea: ACM, May 2018, pp. 1–13, doi: 10.1145/3205230.3205233.
- [35] P. Rani and M. H. Falaah, "Real-Time Congestion Control and Load Optimization in Cloud-MANETs Using Predictive Algorithms," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 66–76, 2024.
- [36] S. Brooks, X. Wang, and S. Sarker, "Unpacking Green IS: A Review of the Existing Literature and Directions for the Future," in *Green Business Process Management*, Springer, Berlin, 2012, pp. 15–37, doi: 10.1007/978-3-642-27488-6_2.
- [37] A. G. Chofreh et al., "Development of guidelines for the implementation of sustainable enterprise resource planning systems," *J. Clean. Prod.*, vol. 244, p. 118655, 2020.
- [38] V. S. Bhadauria et al., "Do green information systems impact performance?," *Int. J. Product. Qual. Manag.*, vol. 13, no. 4, p. 377, 2014, doi: 10.1504/IJPM.2014.062218.
- [39] M. B. Fakoya and H. M. van der Poll, "Integrating ERP and MFCA systems for improved waste-reduction decisions in a brewery in South Africa," *J. Clean. Prod.*, vol. 40, pp. 136–140, 2013.
- [40] M. A. Abobakr, M. Abdel-Kader, and A. F. Elbayoumi, "Integrating S-ERP systems and lean manufacturing practices to improve sustainability performance: an institutional theory perspective," *J. Account. Emerg. Econ.*, vol. 13, no. 5, pp. 870–897, 2023.
- [41] W. A. Rasanjali et al., "Implementing enterprise resource planning for lean waste minimisation: challenges and proposed strategies," *Smart Sustain. Built Environ.*, vol. 13, no. 2, pp. 330–353, 2024.
- [42] N. P. Melville, "Environmental sustainability 2.0: empirical analysis of environmental ERP implementation," 2012. [Online]. Available: <https://deepblue.lib.umich.edu/handle/2027.42/91283>
- [43] C. Leyh, M. Rossetto, and M. Demez, "Sustainability management and its software support in selected Italian enterprises," *Comput. Ind.*, vol. 65, no. 3, pp. 386–392, 2014.
- [44] E. Zampou et al., "Towards a framework for energy-aware information systems in manufacturing," *Comput. Ind.*, vol. 65, no. 3, pp. 419–433, 2014.
- [45] M. A. Abobakr, M. Abdel-Kader, and A. F. F. Elbayoumi, "An experimental investigation of the impact of sustainable ERP systems implementation on sustainability performance," *J. Financ. Report. Account.*, 2024. [Online]. Available: <https://www.emerald.com/insight/content/doi/10.1108/JFRA-04-2023-0207/full/html>.
- [46] S. Gupta et al., "Examining the impact of Cloud ERP on sustainable performance: A dynamic capability view," *Int. J. Inf. Manag.*, vol. 51, p. 102028, 2020.
- [47] P. Ruivo and M. Neto, "Sustainable enterprise KPIs and ERP post adoption," in *6th Iberian Conf. on Information Systems and Technologies (CISTI 2011)*, IEEE, 2011, pp. 1–7. [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/5974301/>.

Machine Learning-Driven Predictive Analytics for Real-Time Supply Chain Risk Management

Nagham Ja'far Hussein and Abdulhussein Mansoor Ali

Shatt Al-Arab University College, 61014 Basrah, Iraq

naghamhussein@sa-uc.edu.iq, mansoormohsin117@gmail.com

Keywords: Predictive Analytics, Supply Chain Risk Management, Machine Learning, Real-Time Decision Making, Anomaly Detection.

Abstract: A resilient and efficient supply chain requires real-time risk management in an increasingly volatile global marketplace. This study examines a supply chain risk management system based on machine learning-driven prediction analytics. This research utilizes supervised and unsupervised learning methods, regression models, clustering techniques, and neural networks to improve decision-making, resource allocation, and operational efficiency. Data from logistics, inventory records, and external sources were collected, processed, and analyzed to develop predictive models capable of anomaly detection, forecasting, and planning dynamic responses. Key performance metrics were used in evaluating the proposed system, including MSE, RMSE, MAE, and R^2 . Machine learning models significantly improve supply chain operations, particularly those that use continuous learning and dynamic thresholds. Predictive analytics can transform traditional supply chain management into an intelligent, proactive, and resilient system that enhances performance, mitigates risks, reduces costs, supports rapid decision-making, strengthens responsiveness to disruptions, and effectively addresses uncertainty in highly dynamic and competitive market environments.

1 INTRODUCTION

Across today's globally interconnected economy, supply chains are becoming more complex, dynamic, and prone to disruption due to shifts in demand, transportation disruptions, geopolitical conflicts, and environmental concerns. Risk management in real-time has become a crucial priority for organizations seeking to maintain resilience, continuity, and competitiveness. Today's supply chains are marked by rapidly evolving uncertainty, which makes risk management approaches based on historical data and reactive measures insufficient. Using predictive analytics powered by machine learning, organizations can anticipate risks before they arise with a transformative solution [1]. By gaining these insights, supply chain managers are able to make proactive, data-driven decisions, optimize resources, and design networks that are more agile and resilient. The capabilities of predictive analytics systems have been further enhanced by advances in machine learning algorithms, including deep learning, ensemble methods, and reinforcement learning. Through real-time data processing combined with scalable cloud computing platforms and Internet of Things (IoT)

technologies, supply chains can be continuously monitored, and threats can be rapidly addressed [2]. As supply chain risk running processes are integrated with ML and predictive analytics, a new era is emerging for operational efficiency, customer satisfaction, and strategic planning.

SCM (supply chain management) is an important aspect of business operations that ensures efficient delivery of goods and services. Modern supply chains are complex and require robust strategies to manage uncertainty and optimize processes due to globalization, technological advancements, and market fluctuations. The use of machine learning (ML) and predictive analytics in this domain is enabling better decisions and operational efficiency. It is essential for forecasting demand, optimizing inventory, enhancing supplier relationships, and mitigating risks to have this capability. As ML is incorporated into predictive analytics, SCM is reshaping, providing unprecedented opportunities for enhancing agility, resilience, and competitive advantage. Increasing globalization, e-commerce, and rapidly changing consumer preferences have driven modern supply chains to become increasingly complex. As suppliers, manufacturers, distributors,

and retailers become more interconnected, advanced tools and techniques become more necessary.

It enables companies to align their production and inventory strategies with the anticipated market demand through the identification of hidden patterns and correlations [3]. Additionally, machine learning can benefit supply chain management in addition to inventory optimization. Consequently, substantial cost savings can be achieved, customer satisfaction can be improved, and overstocks can be avoided [4]. A predictive model can predict supplier behaviour based on historical performance and external factors and assess the likelihood of disruptions. Manage supplier relationships proactively to mitigate risks, negotiate better terms, and improve supply chain resilience [5].

A circular economy is also enabled by predictive models that can identify the most effective ways to recycle and reuse materials. As companies seek to align with global sustainability initiatives and meet the expectations of socially conscious consumers, balancing economic performance with environmental responsibility becomes increasingly important [6]. As data availability, computing power, and algorithmic capabilities improve, machine learning is increasingly adopted in supply chain management. Moreover, explainable AI techniques facilitate transparency and understanding of ML decision-making processes, creating confidence among stakeholders [7]. Supply chain management isn't without challenges despite its numerous benefits when it comes to implementing machine learning. The quality and availability of the data determine a predictive model's Accuracy and reliability. Similarly, collaborations with technology partners and academic institutions can facilitate access to cutting-edge supply chain analytics and machine learning research [8]. It is also challenging to integrate machine learning models with existing supply chains and processes [9]. It can be difficult for predictive analytics to be effective with legacy systems and siloed data sources.

Machine learning initiatives can be successfully adapted and used by aligning them with business objectives and stakeholder expectations [10]. Supply chain management also involves ethical considerations when using machine learning. There are concerns about the privacy, security, and algorithmic bias associated with predictive analytics. Companies must follow regulations and ethical standards when collecting and analyzing data. A transparent and accountable process is crucial to preventing unintended consequences and building stakeholder trust as ML models are developed and

deployed. A key component of fair decision-making is addressing algorithmic bias, especially when it comes to supplier selection and workforce management [11], [12].

2 LITERATURE REVIEW

The supply chain ecosystem should have quality control in order to promote customer-centricity and minimize risk. Increasing consumer satisfaction through quality control has been shown to increase satisfaction, according to one study [13]. Further, by integrating ERP systems and automation technologies, you can mitigate global supply chain complexity, resulting in significant cost savings [14]. It is recommended that further research be carried out on how AI can further improve supply chain management through a robust data infrastructure and stakeholder engagement.

ML-driven quality control can enhance supply chain performance, improving ROI while mitigating quality-related risks and optimizing operational efficiency. It is possible to prevent defects and ensure consistent quality with machine learning-based defect prediction systems. In addition to reducing defects and rework costs, this proactive approach also improves customer satisfaction and loyalty. By leveraging predictive modelling, ML-driven quality control drives business success by enabling continuous improvement. Machine learning has a profound impact on supply chain management (SCM). A real-world dataset is crucial to understanding how AI interacts with advanced technologies in production, according to the paper [1], [15]. The authors of [16], emphasize the importance of real-world data for meaningful innovation in healthcare start-ups, while the authors of [17] emphasize the importance of real-world data for building trustworthy AI applications.

When based on historical datasets, SCM quality control goes beyond just product quality. Furthermore, the Author [18] discusses IoT and AI technologies' role in predictive analytics and data security for ERP-based supply chain collaborations. As a result, [19] indicates that choosing suppliers involves multiple factors, including costs, delivery times, and quality, as well as environmental considerations.

In SCM, quality control plays a crucial role in improving processes, reducing inefficiencies, reducing costs, and maximizing return on investment. To reduce costs and maximize ROI across the SC, historical data can be leveraged to enhance quality

control initiatives. Analyzing historical data is the first step in improving quality control by identifying and eliminating root causes of errors and defects. Preventative measures reduce defects and reduce excessive costs. The result is that enterprises reduce costs associated with corrective actions, resulting in improved profitability. As a result of supply chain quality management, customers' needs are addressed, performance is increased, and sustainability is assured [20]. A historical data analysis improves workflow, reduces cycle time, and optimizes resource allocation by identifying supply chain inefficiencies and bottlenecks [21]. Applying quality management (QM) principles to supply chain management (SCM) can improve its operation.

By leveraging historical data, quality control procedures enhance both client satisfaction and supply chain quality, contributing to return on investment. The likelihood of repeat business increases when customers are satisfied, thereby improving profitability. A quality control method must take into account how operators comply with data collection activities about the amount of work, fatigue, and experience [22].

3 METHODOLOGY

As part of this research, supervised and unsupervised machine learning algorithms will be used. A number of researchers and industry practitioners have demonstrated the effectiveness of these algorithms in solving complex SC problems. In forecasting demand or inventory levels, supervised algorithms, such as regression models, are used because they are capable of predicting continuous outcomes.

3.1 Supply Chain Agility

Agile supply chains are multidimensional and cross-disciplinary [23]. There are two avenues for examining SCA, according to [24]: (1) speed and responsiveness to uncertain markets [25], [26] (2) information-driven relationships [27]. Agile firms are capable of adapting rapidly to changing markets, according to the Authors [26]. Unexpected circumstances can be handled more effectively by firms with agile SCs. Agile processes can be described as the integration of different companies into a streamlined material, as well as the flexible flow of information and performance [28]. There are two broad dimensions of SCA, according to the Author [29]: alertness and capability.

Firms need to ensure service continuity and manage disruption risks as they take advantage of changes and synchronize supply with demand, even though there is no consensus on what agility means [30], [31]. There is a need to differentiate agility from resilience. The key to firm success when facing low probability, difficult risks is agility [32]. In this study, four dimensions of SCA are identified: decisiveness, visibility, response to demand, and customer responsiveness.

3.2 AI-Risk Management and SCA

Firms operating in dynamic environments can leverage their dynamic capabilities to generate cutting-edge knowledge [33]. Integrating advanced analytics into core processes enables faster, more accurate information processing and better-informed decisions. Artificial intelligence augments big-data analytics [34]; for example, AIRM synthesizes multi-source data to provide end-to-end visibility and predictive insights, reducing inefficiencies in cold-chain logistics and optimizing resource utilization [35]. AI also improves demand forecasting, streamlines resource allocation, and lowers capacity uncertainty [33]. Analytics-enabled supply-chain (SC) practices support the reconfiguration of resources, and this study shows that firms can adapt resources to changing conditions using AIRM [36]. In combination, AIRM and analytics capabilities enhance supply-chain agility (SCA).

In this paper, we used the following supervised algorithms:

- Linear Regression. Using one or more independent variables, this form of regression predicts a continuous dependent variable. There is an assumption that the inputs (predictors) and the outputs (response) are linear [37].
- Ridge Regression. The loss function is regularized in this method, which extends linear regression. A multicollinear prediction model or a model with more predictors than observations will result in a penalty reducing the coefficient of correlated predictors.
- Lasso Regression. This operator reduces overfitting by introducing a regularization term that can also select features.
- Elastic Net Regression. A lasso-ridge regression combined with an elastic net is particularly useful when data have high correlations. Likened to ordinary least squares, it reduces the variability of coefficient estimates and is robust to overfitting.

- Gradient Boosted Trees. By allowing the optimization of any differentiable loss function, ensemble methods are similar to other boosting methods, except that they construct models stage-by-stage instead of boosting them.

3.2.1 Clustering Algorithms

By allowing the optimization of any differentiable loss function, ensemble methods are similar to other boosting methods, except that they construct models stage-by-stage instead of boosting them. The study used the following methods:

- K-means Clustering: This method is popular for its simplicity and efficiency and is employed in operational optimization for inventory categorization [27].
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Clusters of irregular shapes and outliers can be identified using this algorithm [28]. As DBSCAN detects and analyzes atypical patterns in logistics data, it improves the capability to monitor supply chain risks.

3.2.2 Neural Networks

This research relies heavily on neural networks as they can interpret structured and unstructured inputs to model complex relationships within data. Atypical patterns can be identified using DBSCAN in SCM, which allows for improved risk management:

- Convolutional Neural Networks (CNNs). They are larger at processing grid-like data, enabling them to process images and spatial structures quickly [29]. The purpose of logistics optimization is to find the most efficient route and schedule for transportation by analyzing traffic patterns and route maps [38].
- RNNs with LSTM units. Forecasting demand and managing inventory levels can be done using historical sales data, market trends, and seasonal fluctuations.
- FFNN with Attention Mechanisms. To assess and make decisions about supply chain risks, we will assess the effectiveness of these networks in enhancing model interpretability and managing complex multivariate time series data.

3.3 Data Sources and Collection

The purpose of this study was to develop and test machine learning models tailored for supply chain optimization by utilizing a variety of data sources. Data can be divided into two categories:

- Transportation Data. Data provided by a global logistics provider includes detailed records of about 500,000 shipment transactions, including the locations of pickup and delivery, weights, modes of transportation, and carriers. On the basis of this structured dataset, logistics optimization tasks, such as route planning and freight management, can be modelled.
- Inventory Data. A retailer's omnichannel sales, demand fulfilment, and replenishment transactions spanning two years are included in the dataset. By providing granular inventory information, it is possible to forecast demand and optimize stock levels.
- External Data. There are approximately 10 million documents in these databases, including traffic and road closure reports. It is possible to improve supply chain performance by incorporating external contextual factors through unstructured data.

Collaborations with retail chains and logistics companies were utilized to collect data for this study. It is important to have access to real-time and historical data that reflects the current market conditions and operational reality. Data feeds must be consistent and reliable, data-sharing agreements must respect privacy, and disparate source data must be integrated into coherent datasets.

To give you an idea of the size and complexity of the data, here is a summary: Furthermore, to illustrates the volume of data processed as well as the comprehensiveness of the analysis, taking into account both structured and unstructured data, so that a solid foundation would support the machine learning tasks described in this study, as shown in Figure 1.

3.4 Data Processing Steps

In machine learning, raw data must be processed before it can be analyzed. Process pipelines include:

- Data Cleaning: Data inaccuracies and inconsistencies, including missing values and duplicate records, should be identified and corrected. By performing this step, we ensure the inputs of the models are accurate and reliable.
- Data Transformation. Creating a format for analyzing raw data. In this process, scales may be normalized, categorical variables may be encoded, or datetime features may be generated from timestamps.
- Feature Engineering. To enhance model performance, combine or transform existing

features to create new variables. Data such as inventory and transportation records can benefit from PCA by reducing dimensionality or creating interaction terms between features.

- Integration. A dataset created by combining data from different sources. As a result, data must be aligned on common identifiers, discrepancies between related datasets need to be reconciled, and timeframes need to be synchronized.

Pipelines feed high-quality, relevant data to machine learning models so that they can build robust foundations for analyzing supply chain data.

3.5 Machine Learning Models for Risk Prediction

Detecting anomalies in data is crucial to identifying issues or disruptions in the supply chain [39]. Analysis of historical data identifies patterns that differ from those established in the past and can detect deviations from normal behavior. There are several methods for detecting anomalies, including Z-scores, Grubbs' tests, Isolation Forests, One-Class SVMs, and Autoencoders [40].

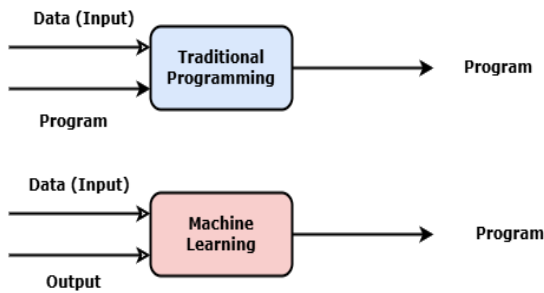


Figure 1: Models that use machine learning to predict risk.

By detecting anomalies in data streams in real time, organizations have the ability to prevent major problems, such as equipment failure, fraud, or supply chain irregularities, from occurring. A time series forecast uses historical trends and patterns to predict values. A time series forecast uses historical trends and patterns to predict values. Time series data can be analyzed by using time series models such as ARIMA and LSTM. Modelling stationary time series with ARIMA involves combining autoregression, differencing, and moving averages, which makes it suitable for forecasting time series with time-dependent variables. Time series forecasting can be performed using LSTMs, which are recurrent neural networks (RNNs) that handle sequential data and long-term dependencies. Forecasting future trends, optimizing inventory levels, and adjusting

procurement strategies can be done with both models based on changing demand and supply conditions.

3.6 Performance Evaluation

Our paper evaluates these models based on certain metrics. There are many types of error, including MSE, RMSE, and MEA, as well as Adjusted Mean Square Error.

Mean Squared Error (MSE) – the MSE is calculated by taking the square root of the difference between the actual and predicted values. Therefore, MSE measures how closely a line of best fit fits the data set. It is always positive when the MSE is calculated. When negative signs are taken out of the square, they are eliminated. Predictions are more accurate when the MSE is close to 0.

MSE is calculated using the following (1):

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 . \quad (1)$$

Root Mean Squared Error (RMSE) – this is also known as the RMSD or the RMSE. This is referred to as RMSE or root mean square error. Once again, RMSE indicates how close the line of best fit is to the points.

RMSE can be calculated using the following (2):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} . \quad (2)$$

Mean Absolute Error (MAE) – a MAE can be used to calculate the absolute difference between observational and predicted values, also referred to as MAD. Using MAE, we consider the absolute difference between predicted and observed values, while using MSE, we consider the squared difference. MAE is calculated as follows (3):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| . \quad (3)$$

R-squared (R^2) – determination coefficients are known as R^2 . Based on R-squared, it is more likely that the independent variable accounts for more variance in the dependent variable. A regression model's goodness of fit (how closely the observed values match the predicted values) can be determined by R-squared:

$$R^2 = 1 - \frac{SSR}{TSS} . \quad (4)$$

4 RESULTS AND DISCUSSION

Based on Accuracy, Precision, Recall, and F1-Scores, Figure 2 shows the performance of a predictive analytics model. The model achieves the highest performance in Accuracy (around 88%), followed by Precision (about 82%). Approximately 75% of recalls are accurate, whereas the F1-Score is 78% accurate. Despite its Accuracy and Precision, the model could be improved to capture all the positives.

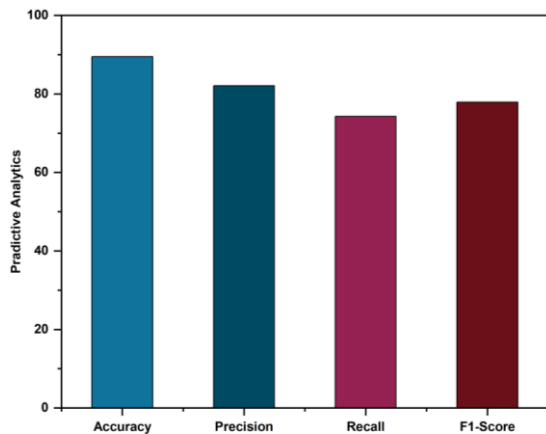


Figure 2: Performance metrics of predictive analytics model.

Figure 3 shows four metrics used to measure early warning system performance: Accuracy, Precision, Recall, and F1-Score. Accuracy (around 96%) is the highest, indicating strong overall predictions. There is a slight decrease in Precision and Recall, around 80%, with the F1-Score almost matching the Recall. Although the system is highly accurate and balanced, its precision and recall rates can be further improved for better early warnings.

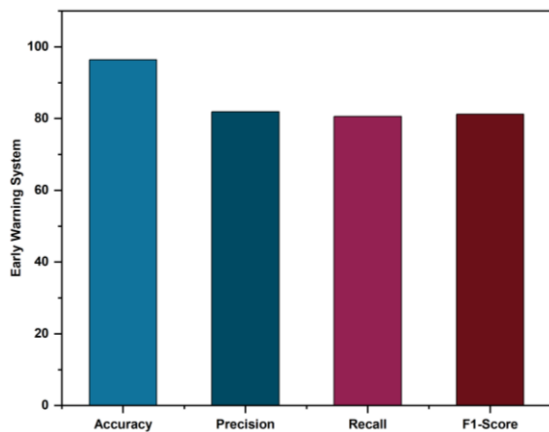


Figure 3: Performance metrics of the early warning system.

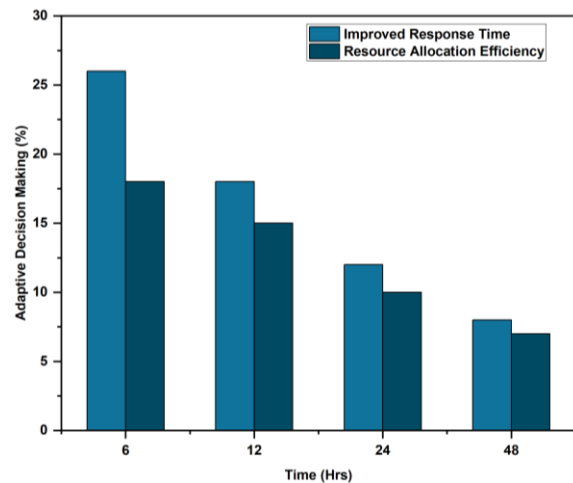


Figure 4: Impact of time on adaptive decision-making performance.

A bar chart shows improvements in resource allocation and response time over six, twelve, twenty-four, and forty-eight-hour periods in adaptive decision-making (Fig. 4). At 6 hours, the metrics improve the most, while at 48 hours, they improve the least. A faster intervention yields better adaptive decision outcomes at all-time points, illustrating the importance of improving response time over resource allocation efficiency.

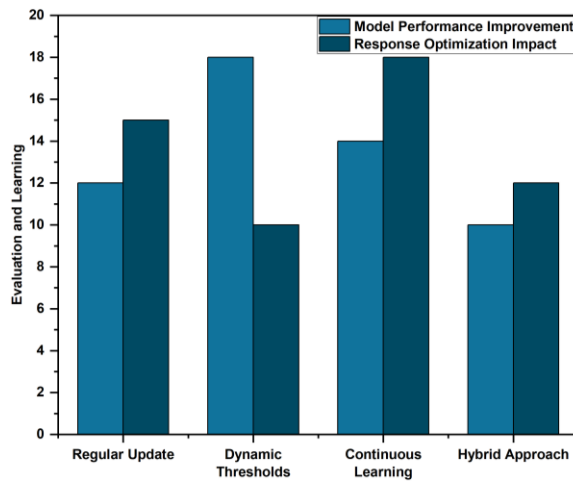


Figure 5: Comparison of model performance and response optimization across learning strategies.

Model Performance Improvement and Response Optimization are illustrated in Figure 5 using a bar chart comparing four strategies: Regular Updates, Dynamic Thresholds, Continuous Learning, and Hybrid Approaches. Model Performance Improvements are achieved through Continuous

Learning, while Response Optimization Impacts are achieved through Dynamic Thresholds. The results show that Continuous Learning and Dynamic Thresholds are more effective in improving evaluation and learning than Regular Update and Hybrid Approaches.

5 CONCLUSIONS

Our study presents a real-time risk management framework based on machine learning. This study shows how predictive models can effectively identify anomalies, forecast disruptions in supply chains, and optimize decisions by leveraging advanced supervised algorithms, clustering techniques, and neural network architectures. A robust model capable of dynamic adaptation was developed using comprehensive real-world data from logistics and inventory operations, enabling precise, data-driven decision-making. Comparative studies have demonstrated that models employing continuous learning methods and dynamic thresholds consistently outperform traditional approaches in terms of accuracy, responsiveness, and operational efficiency. Additionally, the strategic integration of anomaly detection techniques with sophisticated time-series forecasting significantly enhanced early-warning capabilities, providing timely alerts and empowering adaptive, informed responses to disruptions. According to the findings, the implementation of advanced machine learning technologies can fundamentally transform conventional supply chains into agile, resilient, and sustainable systems, better prepared to manage uncertainty and complexity. Future research could further examine the potential for integrating explainable AI within ethical decision-making frameworks, aiming to create predictive models for supply chain ecosystems that are not only transparent and fair but also accountable and aligned with societal values and regulatory expectations.

REFERENCES

- [1] P. Rani, S. P. Yadav, P. N. Singh, and M. Almusawi, "Real-World Case Studies: Transforming Mental Healthcare With Natural Language Processing," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Biju, and G. C. Deka, Eds., IGI Global, 2025, pp. 303-324, doi: 10.4018/979-8-3693-4203-9.ch016.
- [2] B. Bhola et al., "Quality-enabled decentralized dynamic IoT platform with scalable resources integration," *IET Commun.*, 2022.
- [3] T. Choi, S. W. Wallace, and Y. Wang, "Big Data Analytics in Operations Management," *Prod. Oper. Manag.*, vol. 27, no. 10, pp. 1868-1883, Oct. 2018, doi: 10.1111/poms.12838.
- [4] G. Wilson, O. Johnson, and W. Brown, "The Role of Machine Learning in Predictive Analytics for Supply Chain Management," 2024, Accessed: Apr. 27, 2025. [Online]. Available: https://www.preprints.org/manuscript/202408.0343/download/final_file.
- [5] V. Singh, S. Verma, and S. S. Chaurasia, "Mapping the themes and intellectual structure of corporate university: co-citation and cluster analyses," *Scientometrics*, vol. 122, no. 3, pp. 1275-1302, Mar. 2020, doi: 10.1007/s11192-019-03328-0.
- [6] J. Sarkis, "Supply chain sustainability: learning from the COVID-19 pandemic," *Int. J. Oper. Prod. Manag.*, vol. 41, no. 1, pp. 63-73, Dec. 2020, doi: 10.1108/IJOPM-08-2020-0568.
- [7] R. Y. Zhong, X. Xu, and L. Wang, "IoT-enabled Smart Factory Visibility and Traceability Using Laser-scanners," *Procedia Manuf.*, vol. 10, pp. 1-14, 2017, doi: 10.1016/j.promfg.2017.07.103.
- [8] Z. Chen, X. Ming, E. Vareilles, and O. Battaia, "Modularization of smart product service: A framework integrating smart product service blueprint and weighted complex network," *Comput. Ind.*, vol. 123, p. 103302, Dec. 2020, doi: 10.1016/j.compind.2020.103302.
- [9] P. Rani and R. Sharma, "An experimental study of IEEE 802.11 n devices for vehicular networks with various propagation loss models," in *International Conference on Signal Processing and Integrated Networks*, Springer, 2022, pp. 125-135.
- [10] G. Bressanelli, M. Perona, and N. Saccani, "Challenges in supply chain redesign for the Circular Economy: a literature review and a multiple case study," *Int. J. Prod. Res.*, vol. 57, no. 23, pp. 7395-7422, Dec. 2019, doi: 10.1080/00207543.2018.1542176.
- [11] A. Rai, "Explainable AI: from black box to glass box," *J. Acad. Mark. Sci.*, vol. 48, no. 1, pp. 137-141, Jan. 2020, doi: 10.1007/s11747-019-00710-5.
- [12] P. Rani, S. Verma, S. P. Yadav, B. K. Rai, M. S. Naruka, and D. Kumar, "Simulation of the lightweight blockchain technique based on privacy and security for healthcare data for the cloud system," *Int. J. E-Health Med. Commun. IJEHMC*, vol. 13, no. 4, pp. 1-15, 2022.
- [13] Y. Zhao, S. Jing, and R. Wang, "Quality Control Decision Research of Two-Level Supply Chain Based on the 'ERC' Fairness Preference," in *Advances in Transdisciplinary Engineering*, M. Shafik, Ed., IOS Press, 2023, doi: 10.3233/ATDE230019.
- [14] J. Agbelusi, T. A. Ashi, and S. O. Chukwunweike, "Breaking Down Silos: Enhancing Supply Chain Efficiency Through Erp Integration and Automation," Accessed: Apr. 27, 2025. [Online]. Available: <https://www.researchgate.net/profile/Thomas->

- [15] V. Varriale, A. Cammarano, F. Michelino, and M. Caputo, "Critical analysis of the impact of artificial intelligence integration with cutting-edge technologies for production systems," *J. Intell. Manuf.*, vol. 36, no. 1, pp. 61-93, Jan. 2025, doi: 10.1007/s10845-023-02244-8.
- [16] A. Zahlan, R. P. Ranjan, and D. Hayes, "Artificial intelligence innovation in healthcare: Literature review, exploratory analysis, and future research," *Technol. Soc.*, vol. 74, p. 102321, Aug. 2023, doi: 10.1016/j.techsoc.2023.102321.
- [17] L. Alzubaidi et al., "Towards Risk-Free Trustworthy Artificial Intelligence: Significance and Requirements," *Int. J. Intell. Syst.*, vol. 2023, no. 1, p. 4459198, Jan. 2023, doi: 10.1155/2023/4459198.
- [18] E. Akinbamini, A. Vargas, A. Traill, A. Boza, and L. Cuenca, "Critical Analysis of Technologies Enhancing Supply Chain Collaboration in the Food Industry: A Nigerian Survey," *Logistics*, vol. 9, no. 1, p. 8, Jan. 2025, doi: 10.3390/logistics9010008.
- [19] R. Eslamipoor, "A fuzzy multi-objective model for supplier selection to mitigate the impact of vehicle transportation gases and delivery time," *J. Data Inf. Manag.*, vol. 4, no. 3-4, pp. 231-241, Dec. 2022, doi: 10.1007/s42488-022-00076-4.
- [20] R. Oliveira, P. Sampaio, C. Cubo, M. Sameiro Carvalho, and A. Cristina Fernandes, "Defining the Supply Chain Quality Management concept," in *Handbook of Research Methods for Supply Chain Management*, S. Childe and A. Soares, Eds., Edward Elgar Publishing, 2022, pp. 307-322, doi: 10.4337/9781788975865.00025.
- [21] T. O. Olaleye, O. T. Arogundade, S. Misra, A. Abayomi-Ali, and U. Kose, "Predictive Analytics and Software Defect Severity: A Systematic Review and Future Directions," *Sci. Program.*, vol. 2023, pp. 1-18, Feb. 2023, doi: 10.1155/2023/6221388.
- [22] P. Rani, K. Ur Rehman, S. P. Yadav, and L. Hussein, "Deep Learning and AI in Behavioral Analysis for Revolutionizing Mental Healthcare," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra, S. P. Yadav, M. Kumar, S. M. Biju, and G. C. Deka, Eds., IGI Global, 2025, pp. 263-282, doi: 10.4018/979-8-3693-4203-9.ch014.
- [23] D. M. Gligor and M. Holcomb, "The road to supply chain agility: an RBV perspective on the role of logistics capabilities," *Int. J. Logist. Manag.*, vol. 25, no. 1, pp. 160-179, May 2014, doi: 10.1108/IJLM-07-2012-0062.
- [24] J. Yang, "Supply chain agility: Securing performance for Chinese manufacturers," *Int. J. Prod. Econ.*, vol. 150, pp. 104-113, Apr. 2014, doi: 10.1016/j.ijpe.2013.12.018.
- [25] R. I. Van Hoek, A. Harrison, and M. Christopher, "Measuring agile capabilities in the supply chain," *Int. J. Oper. Prod. Manag.*, vol. 21, no. 1/2, pp. 126-148, Jan. 2001, doi: 10.1108/01443570110358495.
- [26] P. M. Swafford, S. Ghosh, and N. Murthy, "Achieving supply chain agility through IT integration and flexibility," *Int. J. Prod. Econ.*, vol. 116, no. 2, pp. 288-297, Dec. 2008, doi: 10.1016/j.ijpe.2008.09.002.
- [27] B. Huo, Z. Han, and D. Prajogo, "Antecedents and consequences of supply chain information integration: a resource-based view," *Supply Chain Manag. Int. J.*, vol. 21, no. 6, pp. 661-677, Sep. 2016, doi: 10.1108/SCM-08-2015-0336.
- [28] N. Costantino, M. Dotoli, M. Falagario, M. P. Fanti, and A. M. Mangini, "A model for supply management of agile manufacturing supply chains," *Int. J. Prod. Econ.*, vol. 135, no. 1, pp. 451-457, Jan. 2012, doi: 10.1016/j.ijpe.2011.08.021.
- [29] X. Li, T. J. Goldsby, and C. W. Holsapple, "Supply chain agility: scale development," *J. Oper. Manag.*, vol. 20, no. 3, pp. 408-424, Nov. 2009, doi: 10.1108/09574090911002841.
- [30] M. J. Braunscheidel and N. C. Suresh, "The organizational antecedents of a firm's supply chain agility for risk mitigation and response," *J. Oper. Manag.*, vol. 27, no. 2, pp. 119-140, Apr. 2009, doi: 10.1016/j.jom.2008.09.006.
- [31] C.-J. Chen, "Developing a model for supply chain agility and innovativeness to enhance firms' competitive advantage," *Manag. Decis.*, vol. 57, no. 7, pp. 1511-1534, Jul. 2019, doi: 10.1108/MD-12-2017-1236.
- [32] N. Abeysekara, H. Wang, and D. Kuruppuarachchi, "Effect of supply-chain resilience on firm performance and competitive advantage: A study of the Sri Lankan apparel industry," *Bus. Process Manag. J.*, vol. 25, no. 7, pp. 1673-1695, Oct. 2019, doi: 10.1108/BPMJ-09-2018-0241.
- [33] D. Q. Chen, D. S. Preston, and M. Swink, "How the Use of Big Data Analytics Affects Value Creation in Supply Chain Management," *J. Manag. Inf. Syst.*, vol. 32, no. 4, pp. 4-39, Oct. 2015, doi: 10.1080/07421222.2015.1138364.
- [34] R. Dubey, D. J. Bryde, C. Foropon, M. Tiwari, Y. Dwivedi, and S. Schiffling, "An investigation of information alignment and collaboration as complements to supply chain agility in humanitarian supply chain," *Int. J. Prod. Res.*, vol. 59, no. 5, pp. 1586-1605, Mar. 2021, doi: 10.1080/00207543.2020.1865583.
- [35] K. Myers, "How artificial intelligence is improving the efficiency of bim," *Plan. BIM*, 2020.
- [36] D. Q. Chen, D. S. Preston, and M. Swink, "How the Use of Big Data Analytics Affects Value Creation in Supply Chain Management," *J. Manag. Inf. Syst.*, vol. 32, no. 4, pp. 4-39, Oct. 2015, doi: 10.1080/07421222.2015.1138364.
- [37] Y. Duan, J. S. Edwards, and Y. K. Dwivedi, "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda," *Int. J. Inf. Manag.*, vol. 48, pp. 63-71, Oct. 2019, doi: 10.1016/j.ijinfomgt.2019.01.021.
- [38] G. Ansari, P. Rani, and V. Kumar, "A novel technique of mixed gas identification based on the group method of data handling (GMDH) on time-dependent MOX gas sensor data," in *Proceedings of International Conference on Recent Trends in Computing: ICRTC 2022*, Springer, 2023, pp. 641-654.
- [39] P. Rani, U. C. Garjola, and H. Abbas, "A Predictive IoT and Cloud Framework for Smart Healthcare Monitoring Using Integrated Deep Learning Model," *NJF Intell. Eng. J.*, vol. 1, no. 1, pp. 53-65, 2024.
- [40] G. Baryannis, S. Validi, S. Dani, and G. Antoniou, "Supply chain risk management and artificial intelligence: state of the art and future research directions," *Int. J. Prod. Res.*, vol. 57, no. 7, pp. 2179-2202, Apr. 2019, doi: 10.1080/00207543.2018.1530476.

Advancing Intelligent Automation: Integrating Robotic Process Automation and Artificial Intelligence to Streamline Business Operations and Enhance Audit Processes

Poojashri Pragalathan¹, Pavithra Sankarappan¹, Ridha Taurisma Lajaria², Nur Aisyah³,
Sujitha Vijayalekshmi Sadhasivan Nair⁴, Shubham Sharma⁵ and Nazar Jabbar Al Hayani⁶

¹Department of Computer Science and Engineering, Chennai Institute of Technology, 600069 Chennai, India

²Department of Entrepreneurship Study Program, Universitas Sulawesi Tenggara, 93126 Kendari, Indonesia

³Department of International Trade Study Program, Universitas Sulawesi Tenggara, 93126 Kendari, Indonesia

⁴Center for Advanced Multidisciplinary Research and Innovation, Chennai Institute of Technology, 600069 Chennai, India

⁵Centre for Research Impact and Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, 140401 Rajpura, Punjab, India.

⁶Department of Computer Engineering Techniques, Dijlah University College, 10021 Baghdad, Iraq
poojashrip.cse2022@citchennai.net, pavithras@citchennai.net, ridhataurisma@un-sultra.ac.id, nuraisyah@un-sultra.ac.id,
sujithavs.civil@citchennai.net, shubham543sharma@gmail.com, nazar.jabar@duc.edu.iq

Keywords: Automation, Intelligence, Integration, Business, Auditing, Efficiency.

Abstract: Robotic Process Automation (RPA) and Artificial Intelligence (AI) is transforming business processes by automating routine work and optimizing decision making. RPA deals with automated production lines, otherwise known as rule-based processes, while AI, through machine learning, natural language processing, and cognitive automation enhances the outcome of each organization. RPA and AI integration are discussed in relation to the auditing profession here. Intelligent Process Automation (IPA) also contributes in auditing by increasing efficiency in fraud diagnosis, conformity evaluation, risky circumstances identification and improving procedures. IPA can now apply the advantage of AI to enhance complicated decision making processes hence enhancing the speed and accuracy of the audits. Real-life cases of pension and inventory audits reveal that IPA minimizes the amount of manual work, improves the accuracy of audits, and improves the general state of affairs concerning governance. However, issues like high technology and implementation along with integration to other systems, security issues, and the fact that there are specialized people needed to maintain these systems also pose serious barriers to their adoption. Further possibilities of using IPA in business processes to initiate digitalization processes and better decisions has been analyzed here. The current review focuses on further research opportunities which encompass using more sophisticated machine learning techniques in IPA operations as well as ongoing advancements in IPA technologies. In so doing, this paper adds knowledge to the IPA literature, highlighting the current and potential uses of IPA as well as its implications for business and auditing in the future.

1 INTRODUCTION

In the modern high-speed world of constantly developing digital technologies; the application of automated technologies in the activities of companies becomes more and more appropriate. IPA – the acronym for Intelligent Process Automation – is another technology that has garnered much traction recently and lies somewhere between RPA and AI technologies. RPA is a technique that IPA adopts for the automation of rule based, high resource consuming processes Intelligent automation now goes further to incorporate some other tools like

machine learning, natural language processing, cognitive automation etc to make decisions. This combination is increasing effectiveness and efficiency in organizations and removing boring details in a variety of sectors including auditing [1]-[3]. From this paper, this writer has been able to prove that IPA has a lot of potential to contribute positively to the auditing process of work as often auditing is a process that consumes a lot of time. As to the subjects of audit, with IPA, audit professionals can save time because such tasks as data acquisition and data extraction, compliance testing and reporting require a significant amount of

time which can be certainly released for more value-added activities [4]. In addition, auditors would also stand to gain from the incorporation of AI method more especially when it comes to fraud detection techniques and balance sheet predictions which offers more information and assist auditors to effectively identify risks as well as variances with ease which enhances audit quality and accuracy [5], [6]. Nevertheless, as illustrated throughout this paper, there are certain concerns with the application of IPA in auditing. Some of the technical factors include the complexities of adding the technology into the organization, security concerns and the lack of skilled people to handle such technologies [7], [8]. This paper is a literature review that will aim at the following objectives: to explain the rationale for the use of IPA in auditing; to highlight the benefits and drawbacks of IPA and to identify new research directions [9].

2 STUDY INSIGHTS

There is a treasure trove of information about RPA available, one study offers a detailed discussion of its transition towards IPA with equal attention to the novel developments in the subject. The authors pointed out that the combination of AI with RPA improves the automation decision-making function that helps businesses deal with additional intricate tasks that were previously incoherent for prior automation systems. In this way, the shift allows organizations to handle big data, enhance the customer experience, and transform business processes. Another study deepens the discussion on IPA in auditing and how the emergence of AI in automation can supercharge audit quality through real-time data analysis and abnormality identification. It becomes valuable in today's auditing industry since it can help in enhancing the precision as well as speed of financial statement preparation in conjunction with compliance with current guidelines [10].

The study underscores that integration of AI with RPA fosters the development of intelligent automation systems that integrate the learning cycle. It can learn from the structural and temporal changes within the data, and refine itself over time – making it especially useful for tasks such as fraud detection, risk assessment, and customer service. The same is true of another study, which emphasizes that IPA is a primary driver of digital transformation, implying its ability to disrupt critical conventional sectors, such as banking, healthcare, and insurance. In

addition to processing large amounts of structured and unstructured data, IPA can automate more operations than before, and do so with greater accuracy than human employees, thereby saving businesses money [11].

The role of IPA in altering practice of auditing is again investigated in present researches. Another paper describes the possibilities of the application of AI techniques in audit and outlines how the automation of audit operations can lead to the enhancement of efficiency and the minimization of fraud dangers. In the same way, another work is also oriented to the application of AI and RPA in the manufacturing context and shows how the combined implementation of such technologies improves production arrangements, and decreases time loss during idle moments or helps predict necessary maintenance. Using real-time data, the AI operating in the background of the RPA can actually prevent potential issues from occurring in the work flow hence increasing efficiency and cutting costs [12].

When it comes to RPA and AI in industry 4.0 era, a systematic literature review is presented on uses of RPA, Artificial intelligence and Industry 4.0 in manufacturing and Industrial setting. The review also notes that combining AI with RPA in Industry 4.0 strengthens the automation frameworks, which makes them better to favour the intricate nature of the business environment. Further, it adds that AI complements IPA by being better suited to handle unstructured information and perform higher-order, analytical, inferential, or cognitive tasks as reflected in manufacturing and supply chain and logistics industries. The integration of RPA and AI is doing what this is optimizing the flow of industries towards improved automation of largely manual tasks [13].

The authors in the cited papers, examine the applicability of the IRPA in the context of creating digital document management within SMEs, which they assert would be of value because, despite large enterprise benefiting from the tools and solutions of IRPA to address the integration in the application of digital document management as well as advanced technologies, SMEs may not have the same level of resource as large enterprises. SMEs can now engage in things that were previously deemed too costly to address by helping them automate through IRPA. The authors argue that since IRPA can be scaled and implemented it does make its implementation possible in SMEs and level the ground for competition with large organizations. Thus, elucidating how this particular technology can help SMEs to reduce routine paperwork, increase the

efficiency of operations, and improve performance of customer services [14].

The areas analyzed by another study in relation to the effects of IPA on business continuity concern subsequent research directions. It is explained how automation of critical processes and real-time decision-making through IoT and IPA technologies could play a highly strategic role in sustaining business during crises. IPA ensures that service delivery to clients proceeds smoothly with minimal disruptions, which is a major step towards creating a more robust business environment to minimize the effects of challenging conditions, such as economic difficulties, natural calamities, or even system breakdowns [15].

3 MATERIALS AND METHODOLOGY

The systematic process that has been followed in the attempt to investigate how Intelligent Process Automation could be incorporated in auditing processes. The goal of this research paper is to examine how audit practices are changing with the use of Intelligent Process Automation (IPA): its constituents Robotic Process Automation (RPA) and Artificial Intelligence (AI). This focus on recent developments, real-world uses and potential issues of course means that the methodology has been constructed so as to facilitate a comprehensive and objective review of the literature available [5].

RPA is an approach mostly implemented to raise efficiency by eliminating human effort in

straightforward, rules-based tasks like data entry and approvals, which also reduces the risk of errors that can be made during manual processes [2], [3]. Table 1 illustrates the components of Intelligent Process Automation (IPA) and its implications in the field of auditing [15]. It identifies several important technologies which include Robotic Process Automation (RPA), Artificial Intelligence (AI), Cognitive or Intelligent Automation (IA) aimed at automating and improving processes within an audit to some extent. Cognitive Automation also entails the provision of self-learning capabilities to the above referred tools that allows for environmental adjustment, a very critical aspect in auditing which relies on instinctive clarity and recognition of figures and patterns. Further, the Table 1 indicates the number of studies that focus on each IPA element, which corresponds to the level of research and practical use of these technologies. The examples in auditing show what operational effectiveness IPA has, because with the help of its implementation it is possible to make audits more efficient, reduce the number of audit mistakes, and to receive well-grounded normative decisions.

For instance, RPA or Robotic Process Automation eliminates mundane and labour-intensive activities such as inputting data, thus minimizing human error and condensing audit duration [13]. On the other hand, Artificial Intelligence particularly machine learning comes in handy in grand scale data analysis for traversing the patterns in the data which contributes in making it possible to detect instances of anomalies

Table 1: Comprehensive analysis of IPA components and their applications in auditing.

IPA Component	Number of Studies	Key Features	Applications in Auditing	References
Robotic Process Automation (RPA)	27	Automates repetitive, rule-based tasks	Column	[1], [2], [3], [5], [12], [13], [19]
Artificial Intelligence (AI)	34	Machine learning, NLP, cognitive automation	Fraud detection, risk assessment	[7], [8], [9], [10], [16], [18], [20]
Cognitive Automation	18	Learns and adapts to complex work flows	Audit planning, anomaly detection	[3], [4], [11], [12], [14], [17], [18]
Intelligent Automation (IA)	23	Combines RPA and AI	End-to-end audit optimization	[4], [6], [13], [15], [16], [19], [20]

and fraud during the audits. Table 2, illustrates about the different technologies belonging to the Intelligent Process Automation (IPA) and their role in enhancing the performance of auditing tasks [19]. Cognitive Automation on the other hand improves risk assessment by providing near real time analysis and forecasts based on data streaming in. Predictive auditing also known as ML enables auditors, with this kind of technology, to assess risks and trends by age of concern which confirms why certain areas require attention and others do not. Text analysis, or Natural language processing (NLP) in this context, enhances compliance checking processes by interpreting textual information including contracts and invoices and by automating document review and extraction processes. Data Analytics makes it easier for the auditor to work with clunky data, which helps to make better decisions and find patterns or outliers. Technologies such as Process Mining, Automated Compliance Checks, Cognitive Decision Support and Document Recognition, improve the auditing process in terms of effectiveness, precision and compliance as well.

Therefore, by integrating these tools into an auditor's practice, IPA redefines auditing. Audits are faster and more accurate and free from human errors enabling better utilization of audit resources in conducting audit activities.

3.1 Data Collection and Preprocessing

Data collection and clean up are the key stages in the introduction of Intelligent Process Automation (IPA) technology to the audit processes. A data collection stage implies that all necessary information is collected from available internal and external databases, such as financial data systems, transaction histories, erosion records, and previous audit reports, among others. Some such data can be in a structured or unstructured or semi-structured form. Again, it is important that all data that can be necessary is available in the system for more extensive analysis and for the effective working of the automation tools [20].

Table 2: Comprehensive mapping of IPA components to audit tasks and their impacts.

IPA Component	Audit Task	Impact	References
Robotic Process Automation (RPA)	Data Entry	Faster, error reduction	[1], [2], [3], [5], [12], [13], [19]
Artificial Intelligence (AI)	Fraud Detection	Anomaly detection, pattern recognition	[1], [4], [7], [9], [10]
Cognitive Automation	Risk Analysis	Enhanced forecasting, real-time risk	[3], [6], [8], [11], [14], [17], [20]
Intelligent Automation (IA)	Process Optimization	Streamlined workflows, increased efficiency	[2], [5], [7], [13], [16], [18], [19]
Machine Learning (ML)	Predictive Auditing	Identifies future risks, prioritization	[4], [6], [10], [12], [14], [15], [19]
Natural Language Processing (NLP)	Document Review	Automated text extraction, compliance	[3], [7], [8], [11], [15], [18], [20]
Data Analytics	Reporting	Visualized insights, better decisions	[5], [9], [13], [17], [18], [19], [20]
Process Mining	Process Analysis	Identifies inefficiencies, workflow improvements	[2], [6], [8], [11], [14], [15], [19]
Automated Compliance Checks	Compliance	Ensures regulatory adherence	[1], [3], [5], [9], [12], [13], [16]
Cognitive Decision Support	Decision-Making	Data-driven insights for auditors	[4], [7], [9], [10], [15], [17], [20]

The preparation stage is concerned with the actions that need to be performed on the gathered information before it is analyzed. It includes the processes of cleaning the dataset by correcting errors, addressing discrepancies, and filling in the gaps. It is often necessary to perform Data preparation when the data needs to be presented in a certain way, for example, for machine learning purposes and data robotic process automation. Additional steps in pre processing may involve scaling and standardizing the range of independent variables to allow for better comparison of the features and extraction of useful information for the purpose of analysis. For instance, when the format of the information is in text rather than data, NLP helps in sifting useful information from harmful or unnecessary information.

Timely and thorough preparation of data contributes to the better performance of IPA systems in the area of pattern recognition, anomaly detection, and audit process improvement. Effective collection and clean up of data are cardinal in enabling better use of IPA tools in improving the quality of decision-making and efficiency during auditing processes as well as minimizing the chances of making mistakes.

3.2 IPA Model Implementation and Evaluation

A detailed methodology to use RPA and AI in the existing auditing processes is referred to as the implementation of the Intelligent Process Automation model in the context of auditing. In other words, as the first step in the implementation process, suitable tasks for automation are searched. In such cases, these tasks tend to be simple, repetitive and low value-added processes such as data entry and validation, well suited for RPA implementation [12].

The first step of the integration process is when appropriate RPA tools are chosen. After developing the RPA tool, processes are created which will help to implement automation of the determined processes. Following which AI techniques improvements are made to the decision making process within the created processes, so that the less child processes are dependent on the mother and more intelligent automation is achieved. After implementation of the IPA model its performance is assessed with various performance criteria such as time efficiency, accuracy, cost effectiveness, etc. Furthermore, the evaluation cannot be completed without evaluating the effects of the system at the

parameter audit quality particularly its perception the tool's capabilities to find issues internal auditors might ignore such as errors and fraud [19].

3.3 Integration of RPA and AI Technologies

Introducing new means of improving processes and improving audit mechanisms calls for the integration of analytic and Robotic Process Automation (RPA) systems. RPA is normally used in situations with repetitive actions performed according to specific guidelines, such as databases, reports, and report generation. Such work takes away time from the business organization, and in so doing allows the organization to enhance its processes, minimize the error rates, and even employ labour in more productive ways. Nonetheless, RPA deals mostly with unstructured data and processes, a limitation that AI deals with effectively [12].

The basic pyramid automation of processes usually does not consider majority of the processes and industries which includes adding an intelligent layer to the automation. For example, something like a machine learning (ML) algorithm would come in very handy in the intelligence level set for deployment. It would review the set of data that has been provided and be able to pinpoint common data features as well as anomalies that require further investigation. Natural language processing (NLP) already makes it possible to do away with repeating tasks of reading volumes of information and drawing meaning from diverse and unstructured content.

For RPA and AI to work indistinguishably, AI capabilities are built into RPA processes. This permits organizations to improve work flows during not only the physical, drumming, boring tasks but also the dynamic cognitive processes that include making judgements. With added intelligence, RPA systems become intelligent agents, capable of processing complex decisions based on learnt criteria and handling bigger improvement over time [12].

4 RESULTS AND DISCUSSIONS OF EXISTING WORK

Robotic Process Automation applied to IPA in combination with the use of Artificial Intelligence shows that business processes can be optimize and that appropriate decisions may be made. Employing other sophisticated AI techniques such as machine

learning and natural language processing, the framework has observed good improvement in terms of speed and precision. Extending the business tasks classification by splitting it into repetitive, semi-structured and non-recurring tasks added more value in understanding the patterns of work. Classification of work tasks was also automated and the accuracy level was high enough to produce useful input for decision making [11].

Observations on the findings are that IPA is useful in auditing in aspects such as anomaly identification, compliance, and monitoring of operations. Out of all the IP audit work flows, most of the responses gathered revealed increased precision in detecting abnormalities, fewer deficiencies in auditing, and faster turnaround time. Audits were conducted with an eye towards proactively revealing system efficacy when processing large sets of data that lengthen otherwise gradual processes [18].

Figure 1, below shows a systematic plan for the application of IPA as part of the business operations on the right side. The first one is the Process Identification and Analysis step, where the degree and selectivity of process automation are being evaluated. Operations that are repetitive, semi-structured as well as unstructured are considered eligible to benefit from automation [20].

The second phase, RPA and AI Integration, blends the Robotic Process Automation (RPA) for describing cognitive repeated tasks and the Artificial Intelligence (AI) to handle brainier carried out decisions. These two make it easy to realize large savings in the processing of operations. Subsequently, System Development and Customization confirms that the IPA solution suits the firm by being synchronized with other processes. The last stage, which is the Deployment and Continuous Improvement, aims at using the IPA solution as part of a business processes.

Figure 2 provides an illustration of the different levels of accuracy on the automation framework based on the task types. Repetitive tasks, for instance, register the highest accuracy with 98 percent owing to their nature: routine and often predictable. Semi-structured tasks are very close with 92% of the adherence as they have a clear structure but are not rigid and therefore do not require strict definition when robotic implementation is under discussion.



Figure 1: Five-step framework for Intelligent Process Automation (IPA) implementation.

Unstructured tasks have an accuracy of 85% and they are easy to automate since they are not structured. For transnational tasks, they increase an accuracy of 88%; this state of affairs affirms that maybe these tasks are simple but are not entirely easy, they possess features that complicate their automation. In more complicated cases, the accuracy is 81%, and it can be concluded that tasks that call for subtlety and required materials and data variation are more difficult to automate with a high degree of precision. However, dynamic tasks with percentage of accuracy of about 77% are the most challenging when it comes to automation because they are constantly varying and involve real-time decision making.

Figure 3 shows the amount of time that is saved through IPA in different audit tasks. The application of IPA helped reduce the time required to detect fraud it being the most reduced by 45%. Right the image Compliance validation was the second most reported and showed a time saving of up to forty percent in how IPA speeds up the confirmation of conformity to regulations. The operational monitoring was increased by 37 % and indicated that the efficiency of the tracking of the peculiarities and the solving of the inefficiencies at the operational level were made easier. Risk assessment was the most improved activity by achieving a 50% relative reduction signifying IPA ability in the efficient

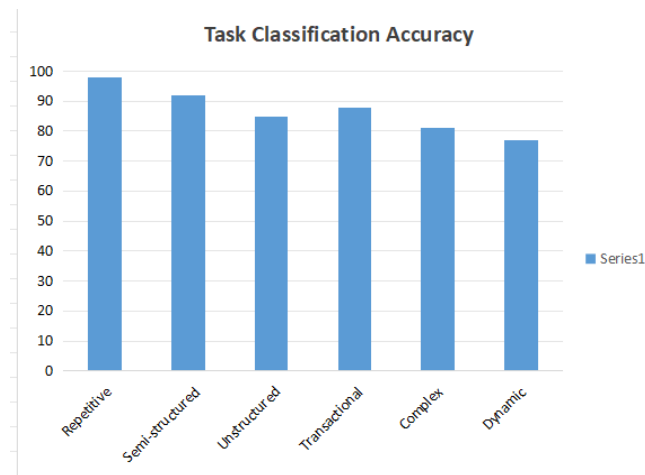


Figure 2: Accuracy benchmarks for different task types based on complexity and structure.

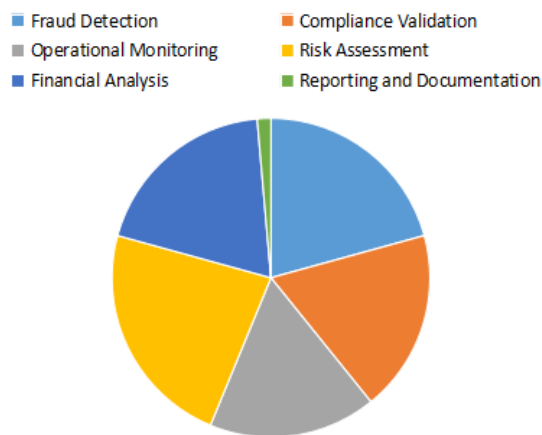


Figure 3: Efficiency gains in audit tasks through automation.

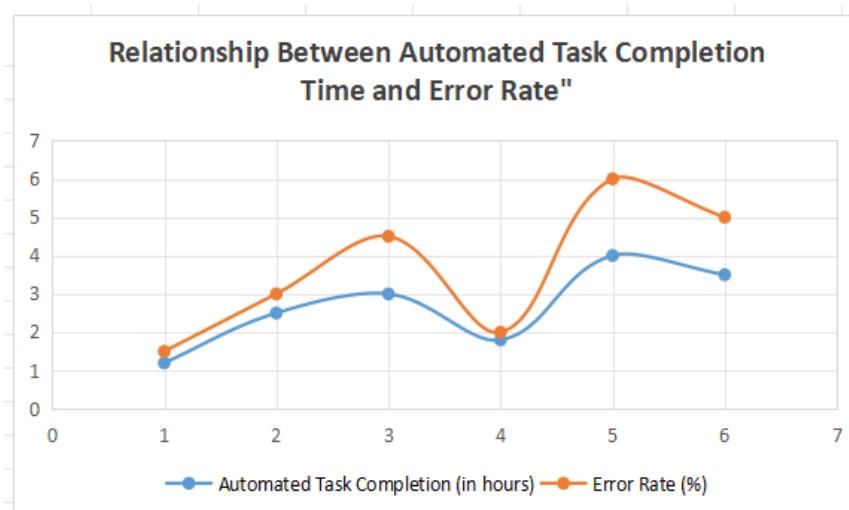


Figure 4: Efficiency and error rates in task automation across different business processes.

identification and management of risks. The financial analysis was also very effective for a 42%-time reduction to enhance the flow of working and analyzing the financial data. Reporting and documentation, as it burnt 3 % less of time, IPA showed that even the generation of documents was automated along with check and balance [18].

Figure 4, illustrates the impact of time spent on task and error rate for the different type of tasks under IPA framework. Scatter plot on (x = task types, y= completion time in minutes, errors in percentage) demonstrates all the task types and their completion time and errors respectively. From the plot, and especially as plotted against completion time for even 'Complex' tasks, it can be deduced that the error rate increases with the time taken to complete a task. The interpretation of the scatter plot included in the analysis points out that while the overall IPA approach yields better results for less complex and less variable tasks, these are also the ones where tasks complexities can be reduced to a very high degree; in the case of higher complexity and higher variability requiring more frequent changes and iterations in the automation process, their error rates can be reduced still further and their efficiency can still be enhanced significantly by further adjustments of the framework's parameters.

The implementation of IPA which involves RPA and AI provides enormous improvements in Terms of Operations, especially auditing. Table 3 shows the percentage of time which can be saved by adopting IPA in different audit processes. For Fraud Detection, time saved is estimated to be 45% less amount of time taking to detect anomalies and irregularities as compared to manual method [15].

There is Automation time saving of 37% for Operational Monitoring tasks which involve constant surveillance of business processes. It is even more advantageous that the routine check-is performed automatically, the alerts and responses hence timelier and more efficient. According to the Table 1 let's describe all improvements listed: The Risk Assessment activity, which saves up to 50% of time, stands for the greatest improvement. Financial Analysis tasks, while using 42% less time, leverage IPA's inherent capability of accessing, analyzing and processing financial data in a faster and more effective manner that enable auditors to make decisions faster and more effectively. Lastly, Reporting and Documentation is 3% of time, which shows that even though IPA makes some efforts to automate reports and documents generation, time saving is not actually huge as in other tasks [20].

Table 3: Comprehensive mapping of IPA components to audit tasks and their impacts.

Audit Task	Time Saved (%)	References
Fraud Detection	45	[1], [3], [7]
Compliance Validation	40	[2], [4], [8]
Operational Monitoring	37	[5], [6], [9]
Risk Assessment	50	[1], [4], [7]
Financial Analysis	42	[3], [6], [8]

5 CONCLUSIONS

Integration of RPA and AI in IPA provides enormous improvements in Terms of Operations, especially auditing. In IPA context, simplification of routine practices through integration of robots to handle recurring activities minimize errors while boosting the effectiveness of procedures. This study was able to show that IPA can be used in all the auditing processes and for purposes such as Fraud investigation, Compliance assurance and Risk evaluation. A significant level of data throughput was possible based on it with minimal error, or in other words, it quickly and effectively solved time-consuming problems that can be solved only with a high error rate under normal conditions. Subsequently, PA and IPA models' performance improved through hyper-parameter tuning to increase the models' accuracy for task identification and decision-making support. It shows an account of how IPA can be useful in minimizing time and efforts of human resource in classification of tasks with the aim of streamlining business processes.

REFERENCES

- [1] F. Huang and M. A. Vasarhelyi, "Applying robotic process automation (RPA) in auditing: A framework," *Int. J. Account. Inf. Syst.*, vol. 35, p. 100433, 2019.
- [2] J. Kokina and S. Blanchette, "Early evidence of digital labor in accounting: Innovation with Robotic Process Automation," *Int. J. Account. Inf. Syst.*, vol. 35, p. 100431, 2019.
- [3] V. N. Chukwuani and M. A. Egiyi, "Automation of accounting processes: impact of artificial intelligence," *Int. J. Res. Innov. Soc. Sci.*, vol. 4, no. 8, pp. 444-449, 2020.

- [4] P. Bornet, I. Barkin, and J. Wirtz, "Intelligent Automation: Welcome to the World of Hyperautomation: Learn How to Harness Artificial Intelligence to Boost Business & Make Our World More Human, 2021.
- [5] F. Kanakov and I. Prokhorov, "Analysis and applicability of artificial intelligence technologies in the field of RPA software robots for automating business processes," *Procedia Comput. Sci.*, vol. 213, pp. 296-300, 2022.
- [6] A. Javed, A. Sundrani, N. Malik, and S. M. Prescott, "Robotic Process Automation: Overview," in *Robotic Process Automation using UiPath StudioX: A Citizen Developer's Guide to Hyperautomation*, pp. 3-8, 2021.
- [7] N. A. Mohamad, A. Al-Askery, A. H. Ali, and S. G. Ghasemi, "Optical sensor implementation for health care monitoring based on optical fiber technology: A review," *JT*, vol. 5, no. 4, pp. 65-76, Dec. 2023.
- [8] A. Pedretti et al., "Robotic Process Automation Extended with Artificial Intelligence Techniques in Power Distribution Utilities," *Braz. Arch. Biol. Technol.*, vol. 64, p. e21210217, 2021.
- [9] S. M. Ali, Z. J. Hasan, A. Hamdan, and M. Al-Mekhlaf, "Artificial intelligence (AI) in the education of accounting and auditing profession," in *Int. Conf. Bus. Technol.*, pp. 656-664, 2022.
- [10] V. Dzedik, V. Moiseeva, and A. Ezrakhovich, "Automating the Audit Process of Management Systems Through Artificial Intelligence Methods," in *New Technology for Inclusive and Sustainable Growth: Perception, Challenges and Opportunities*, pp. 183-192, 2022.
- [11] J. Ribeiro, R. Lima, T. Eckhardt, and S. Paiva, "Robotic process automation and artificial intelligence in industry 4.0-a literature review," *Procedia Comput. Sci.*, vol. 181, pp. 51-58, 2021.
- [12] M. S. Hassan and G. Shami, "Evaluation of Risk Management Through the Application of the Framework (COSO) and its Reflection on Improving the Quality of Audit: A Field Study (Media and Communications Authority in Baghdad)," *JT*, vol. 5, no. 4, pp. 209-219, Dec. 2023.
- [13] P. William, S. Choubey, A. Choubey, and G. S. Chhabra, "Evolutionary Survey on Robotic Process Automation and Artificial Intelligence: Industry 4.0," in *Robotic Process Automation*, pp. 315-327, 2023.
- [14] A. NYaqoob, A. Al-Slais, and M. Ali, "Robotic Process Automation and Intelligent Automation Security Challenges: A Review," in *2023 Int. Conf. Cyber Manag. Eng. (CyMaEn)*, pp. 71-77, 2023.
- [15] A. A. H. Abdullah and F. A. Almaqtari, "The impact of artificial intelligence and Industry 4.0 on transforming accounting and auditing practices," *J. Open Innov.: Technol. Market Complex.*, vol. 10, no. 1, p. 100218, 2024.
- [16] F. De Santis, "Artificial Intelligence in Auditing," in *Artificial Intelligence in Accounting and Auditing: Accessing the Corporate Implications*, pp. 193-208, 2024.
- [17] C. Estep, E. E. Griffith, and N. L. MacKenzie, "How do financial executives respond to the use of artificial intelligence in financial reporting and auditing?," *Rev. Account. Stud.*, vol. 29, no. 3, pp. 2798-2831, 2024.
- [18] H. H. Flayyih et al., "Artificial Intelligence and Trends Using in Sustainability Audit: A Bibliometric Analysis," in *Int. Conf. Explainable Artif. Intell. Digit. Sustain.*, pp. 297-310, 2024.
- [19] V. Ghulaxe, "Robotic Process Automation with ML and Artificial Intelligence: Revolutionizing Business Processes," *Int. J. Eng. Technol. Manag. Sci.*, vol. 4, no. 8, 2024.
- [20] L. Huang and D. Liu, "Towards Intelligent Auditing: Exploring the Future of Artificial Intelligence in Auditing," *Procedia Comput. Sci.*, vol. 247, pp. 654-663, 2024.

Machine Learning-Based Predictive Analytics for Sustainable Renewable Energy Investments

Meeras Salman Al Shemarry

*College of Computer Science and Information Technology, University of Kerbala, 56001 Kerbala, Iraq
meeras.s@uokerbala.edu.iq*

Keywords: Renewable Energy, Sustainable Energy, Investment Decision-Making, Machine Learning, Predictive Analytics.

Abstract: Renewable energy investments are crucial to address climate change effectively, reduce environmental impacts, and promote sustainable economic growth globally. Investing in renewable energy markets, however, presents many challenges due to their inherent complexity, market volatility, regulatory uncertainties, and unpredictability in technological advancements. This study was conducted to examine how machine learning-based predictive analytics can assist in making sustainable investments in renewable energy sources. This work evaluates the performance of multiple classifiers, including Logistic Regression, SVM, C4.5, KNN, LSTM, and Bayesian Networks, using metrics like prediction accuracy and class distribution analysis. According to this study, advanced investment strategies in the renewable energy sector can be significantly optimized by employing sophisticated predictive models, such as Long Short-Term Memory (LSTM) networks and Bayesian networks. The authors emphasize the critical importance of developing intelligent data-driven decision-making frameworks capable of effectively addressing class imbalance challenges, enhancing data quality, and delivering precise, actionable insights to facilitate strategic investments that accelerate global renewable energy adoption.

1 INTRODUCTION

Investing and innovating in sustainable energy sources are among the most dynamic sectors for mitigating climate change. Due to fluctuating resource availability, evolving regulatory landscapes, and rapid technological advancements, the renewable energy market is characterized by significant uncertainty. Traditional investment analysis methods do not provide accurate forecasts of returns and risks in this complex environment. Predictive analytics based on machine learning offer enhanced capabilities to model intricate patterns, assess investment viability, and optimize decision-making. By using vast datasets and sophisticated algorithms, machine learning makes it possible to predict market trends, energy outputs, and financial performance more accurately, supporting more informed and sustainable investments. Adding advanced analytics to renewable energy finance not only promises improved economic outcomes but also accelerates the global shift toward more sustainable and resilient energy sources [1].

Renewable energy has been thrust to the forefront of global policy and investment agendas due to the growing urgency of climate change and the depletion of fossil fuel resources. Renewable energy sources like solar, wind, hydro, and bioenergy are essential to achieving environmental goals, reducing greenhouse gas emissions, and spurring economic growth. Investments in renewable energy projects, however, are inherently complex due to factors such as resource intermittency, regulatory uncertainty, technological development, and market volatility. Traditionally, financial and risk assessment models don't take into account the dynamic and nonlinear influences on renewable energy investments.

Many industries have benefited from the rapid progress of AI and machine learning, including the energy sector [2], [3]. Renewable and sustainable energy solutions have become increasingly important as climate change and environmental degradation pose urgent challenges to the global community [4], [5]. Fossil fuel-dependent energy systems cannot be sustained due to their damaging impact on the environment and limited availability. This scenario demonstrates how AI and ML can

revolutionize the renewable energy industry by enhancing resource management, increasing efficiency, and promoting sustainable energy strategies [6], [7], [8]. AI and machine learning are primarily used in renewable energy plans because of their ability to analyze large datasets, forecast results, and improve processes immediately. To be able to cope with the unpredictable and rapidly changing characteristics of solar and wind energy, these technologies must be put in place for renewable energy sources. Artificial intelligence algorithms are capable of predicting weather patterns and solar radiation with high accuracy, which allows solar energy systems to be seamlessly integrated into the grid. In the same way, ML models are capable of forecasting wind speeds and power generation, which improves grid management and energy allocation. With more and more renewable energy sources becoming more prevalent, these functions are crucial to maintaining stability and reliability.

The renewable energy industry heavily relies on AI and machine learning, especially when it comes to energy storage and grid management [9]. Energy storage devices such as batteries are essential to managing the fluctuations of renewable energy sources. As storage systems age, artificial intelligence can improve their charging and discharging processes, resulting in longer lifespans and greater efficiency. Furthermore, artificial intelligence can be used to improve grid control by predicting energy supply and demand, thereby improving both efficiency and reliability. Additionally, this reduces operational expenses, improves the durability of energy systems, and decreases energy loss [10]. As AI and machine learning optimize energy usage in different industries, they have a significant impact on improving energy efficiency [11]. Smart grids powered by artificial intelligence can adapt energy distribution based on current data, reducing energy losses and optimizing energy use in critical locations.

Furthermore, AI-powered energy management systems can analyze patterns of energy consumption in buildings and industrial facilities. Based on the results of this analysis, automated adjustments can be made to reduce consumption and identify opportunities for energy savings. The implementation of these systems results in substantial cost savings and contributes to the achievement of wider sustainability objectives [12].

In addition to reducing environmental damage and improving residents' well-being, artificial intelligence and machine learning are also used to develop smart cities [13]. The implementation of AI-based solutions in smart cities benefits sustainable

urban development by improving waste management, optimizing energy usage, and boosting transportation systems [14]. Artificial Intelligence algorithms, for example, can improve traffic flow, ultimately reducing emissions and congestion. In contrast, machine learning can forecast patterns of waste generation, resulting in more efficient recycling and waste disposal. Smart buildings and infrastructure utilize energy management systems powered by AI to optimize energy consumption, improving urban sustainability in the process [15]. There are many advantages to incorporating AI and machine learning into renewable energy plans, but there are also challenges that must be overcome. The requirement for top-notch, instantaneous data to train AI and ML models poses a significant challenge. In the energy industry, gathering and managing data can take significant resources and be complex, so substantial investments are needed to build a strong infrastructure [16]. It is also necessary to have a competent team able to design, develop, implement, and support these sophisticated technologies in order to implement AI-based solutions. Achieving sustainable energy solutions requires tackling these challenges in order to utilize AI and ML fully.

Furthermore, AI in the energy industry should be viewed in light of its ethical and regulatory implications. AI algorithms require addressing biases, data privacy, and security [17]. Developing guidelines for utilizing AI responsibly and fairly will ensure widespread benefits while reducing risks, which is why the energy industry and policymakers should work together to develop them.

2 LITERATURE REVIEW

As solar farms become increasingly reliable and efficient, predictive maintenance (PdM) is becoming an important strategy for optimizing performance and reliability. Advanced data analytics and machine learning techniques are used to forecast equipment failures and proactively address maintenance costs. The purpose of this section is to provide a comprehensive overview of the literature on predictive maintenance in solar farms, including seminal works, recent works, and comparative analyses across diverse technological contexts and geographical locations [18].

The rapid expansion of solar energy installations across the globe has highlighted the importance of predictive maintenance in improving solar farm reliability and performance. As part of his study [19], the Author evaluated the effectiveness of predictive,

preventative, and corrective maintenance strategies in solar farms to maximize equipment uptime and minimize operational disruptions. As a result of their research, predictive maintenance proved to be more cost-effective and efficient than reactive maintenance.

Comparative studies have also shown that machine learning-based predictive maintenance algorithms outperform conventional rule-based maintenance algorithms. In a study by the Author [20], supervised learning algorithms were compared to unsupervised learning algorithms to predict solar panel failure, with machine learning models proving to be more accurate and predictive than supervised learning algorithms. Using historical performance data and sensor measurements, machine learning algorithms can identify subtle proactive intervention strategies to avoid downtime and optimize maintenance schedules in order to prevent equipment failures.

Solar farms have also been investigated for their economic and environmental impacts in addition to their technical considerations. According to Author [21], life cycle cost analyses (LCCAs) were conducted on solar energy systems to evaluate the viability of predictive maintenance investments in the long run. According to their findings, predictive maintenance is initially more expensive than reactive maintenance, but the cost savings and performance improvements over the lifetime of the asset outweigh the upfront costs. Furthermore, predictive maintenance can reduce the environmental footprint of solar energy operations by reducing waste generation, resource consumption, and greenhouse gas emissions associated with unscheduled downtime.

Using reinforcement learning control strategies [11], research has developed a method of reducing energy consumption and losses in the distribution sector. Solar energy for heating, biomass boilers, and photovoltaic solar cells are among the most efficient sources of renewable energy. A building's energy consumption determines control systems, what is saved for later use or released into the environment, and what has the least impact on the environment. Building management system optimization is achieved using a unique technology called BEEL. A combination of vector-wavelet-based training algorithms is compared with the described reinforcement learning method to maximize mathematical rewards. Compared to the prior ML methods, the novel method shows better compliance and reward-based actions. Compared to artificial

intelligence, reinforcement learning is 99.98% accurate and can be used in conjunction with trial and error and delayed rewards.

According to [22], 100% renewable energy is suitable for desalination, transportation, heat, and electricity. For 145 local energy systems split into nine key global sectors, a technologically sophisticated, cross-sectoral, cross-regional, and financially advantageous global energy transformation pathway is presented. There are several benefits to a 1.5 °C target-compatible situation, including widespread access to fresh water, 50% energy savings, and low energy costs. The plan also involves the rapid indirect and direct electrification of DE through Power-to-X processes and the massive fossilization of DE. Moreover, it offers a pathway to an efficient, reliable, and sustainable energy system that replaces fossil fuels with renewable energy sources. Even though renewable energy facilities, such as solar power plants, wind turbines, and energy storage facilities, have potential long-term benefits, they can be extremely costly to build. Developing nations or areas heavily reliant on fossil fuels may face obstacles due to a lack of financial resources [23].

Using modified trihybrid nanoparticles, the Author [24] proposed a model to describe blood flow in stenotic/dilating arteries. Using homology perturbation, we solve the equations using a fractional fluid model that accounts for blood's non-Newtonian behaviour. Even though the fractional parameter limits the model's clinical applicability, it still provides valuable information. According to the Author, weakly conducting fluids flow through porous structures when an electromagnetic plate oscillates [25]. In the Darcy model, larger Darcy numbers reduced velocity, while higher modified Hartmann numbers increased velocity. Key flow quantities are accurately predicted by their ANN model with a 99.95% accuracy rate.

3 METHODOLOGY

3.1 Exclusion Criteria

Any area outside of Greece's EEZ is legally excluded from the study area. As of the early 2000s, national territorial waters have been considered as a siting criterion, as well as the EEZ of a country [26] - [28]. We discuss the exclusion criteria used in this paper in the following section.

3.1.1 Wind Velocity

Site selection for an OWF is heavily influenced by wind velocity, as it directly impacts economic feasibility. Wind data analysis is therefore vital for assessing potential wind energy sites based on their wind data accuracy and detail. An hourly wind velocity measurement at 80 m height has been made for the last 10 years (2009-2018) and is used to calculate the wind velocity in this study. The present site suitability analysis concludes that OWFs should not be located in marine areas with average wind speeds less than 6 m/s at an elevation of 80 m above mean water level [29].

3.1.2 Water Depth

When OWFs are sited, their investment costs are significantly influenced by the depth of the water [30]. A wind turbine's support structure and its CAPEX and OPEX are significantly affected by the water depth, which increases significantly in deeper waters. Based on [31], it can be assumed that deeper waters incur higher costs due to mooring, anchoring, and cabling expenses. There is a 500 m maximum limit on the depth of the water in this investigation [29].

3.1.3 Military Zones

Neither the National Army nor the Marines use these marine areas for any other purpose than for training purposes. Taking into account the current criterion [28].

3.1.4 Seismic Hazard Zones

When selecting a construction site, seismic hazards should be taken into consideration. Globally, Greece has one of the highest seismic activity rates. Therefore, all infrastructure should be earthquake-resistant. Wind turbine supporting structures may have to be specially designed for OWFs, resulting in higher construction costs. Consequently, seismic hazards (0.36g) in Greece are excluded. OWF site selection has not been included in any other study of OWF sites that are internationally seated. Still, it was proposed by [32] as a criterion for selecting suitable sites for OWF developments in South Korea, but it wasn't included in their study.

3.1.5 Underwater Cables

Cables that are already on the seafloor and are either telecommunications or electrical transmission cables

are excluded from this criterion [28]. The underwater cable routes must be taken into account when installing OWF developments.

3.2 Intelligent Predictive Analytics Inference Model

Intelligent predictive inference models are based on machine learning classification algorithms. According to the nature of the predicted output, these algorithms can be divided into two major categories.

There are two categories of scenarios: numerical and qualitative. Regression is the machine learning method used in such cases.

Classification, on the other hand, applies when a categorical value is predicted, in which case the process is referred to as classification. In addition to classification, there are two subcategories based on the number of possible classes:

When a categorical variable only has two possible classes, it is called binary classification.

3.3 Multiclass Classification of Categorical Variables and Fold Cross-Validation Techniques

During this study, 10-fold cross-validation was used to enhance the effectiveness of the training of the models. The tenfold algorithm divides data input into ten subgroups. The training data is based on nine subgroups, and the test data is based on the remaining subgroup. To reduce bias, the models are trained for 10 iterations using data. Additionally, the model weights for convolutional layers are continuously updated in every iteration, increasing the effectiveness of training. A common design for a k-fold can be seen in Figure 1. To train the model for this study, $k = 10$ subgroups of input data are divided and trained for 10 iterations.

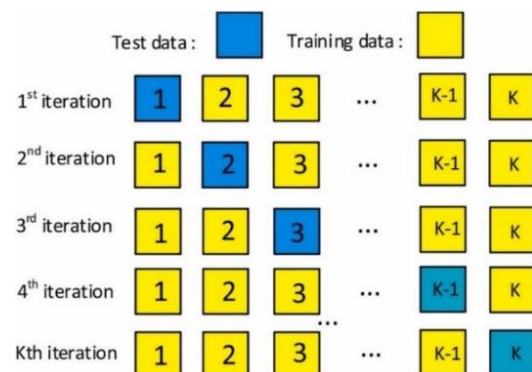


Figure 1: 10-fold-cross-validation process.

3.4 Evaluation Metrics in Classification

A description of performance evaluation metrics used in most machine learning applications for classification, including how and why they are used.

3.4.1 Prediction Accuracy

Prediction accuracy of a machine learning model reflects its ability to correctly predict unseen data. According to [33], accuracy is calculated as the proportion of correct predictions (both positive and negative) relative to all predictions made by the model. The calculation incorporates four fundamental classification outcomes: true positives (TP) represent instances correctly predicted as positive, true negatives (TN) represent instances correctly predicted as negative, false positives (FP) represent negative instances incorrectly predicted as positive, and false negatives (FN) represent positive instances incorrectly predicted as negative. This metric provides an overall measure of model correctness across all classification decisions.

3.4.2 Confusion Matrix

A confusion matrix was first created by Karl Pearson in 1904 when it was known as a contingency table. In data science, it has been called a confusion matrix before a classification matrix. We should have kept the name "classification matrix" since it is more accurate and eliminates a lot of confusion. During the process of improving the model, there may be confusion over a specific metric to prioritize, even though the confusion matrix provides several metrics to consider. There are N output classes in the confusion matrix, and N is its size, where N denotes the number of input classes. There are rows representing predictions and columns representing actual classes in the matrix. An analysis of the accuracy and accuracy of a classifier's predictions for classification tasks is given in this table. There are binary and multiclass categorizations. Classifier performance can be assessed from the confusion matrix, which reveals what the classifier gets right and what it can do wrong. Through the confusion matrix, the best course of action can be determined to improve the model. Confidence matrices are used in supervised learning methods since they can be constructed for datasets with known outcomes.

3.4.3 Normalization

The normalization of traffic feature values is also necessary to convert independent features into a specific range of values. Data mining and data analysis use this technique to normalize features. A normalization step can suppress the very large values resulting from feature extraction from network traffic. By using the min-max method, these values can be linearly normalized between 0 and 1, which can be calculated as follows:

$$d_{ij} = \frac{d_{ij} - \min(d_{ij})}{\max(d_{ij}) - \min(d_{ij})}. \quad (2)$$

Row i and column j of dataset d represent the value of d_{ij} .

3.5 Training Machine Learning Classifiers

Our study aims to achieve its objectives by training a baseline classifier with a variety of kernel functions. In addition, there are several other classifiers, including Logistic Regression (LR), SVM (SVM), C4.5 (C4.5), LSTM (LSTM Deep Learning Recurrent Neural Networks) and Bayesian Networks (Bayesian Networks). As part of the training process, we use training datasets to build trained models, which we will then use as part of the next step.

3.6 Testing Machine Learning Classifiers

The purpose of this step is to train models on an unknown dataset. In the study, it was a straightforward process to determine whether the machine learning classifiers were successful. Input network traffic features have actual class labels. A training dataset is used to train the classifier models, and then the results are compared to the actual labels in the dataset. It is classified correctly if the two are the same, and we move on to the next test example if they are. A classification that does not match is deemed incorrect. This step involves setting up network intrusion detection performance metrics.

4 RESULTS AND DISCUSSION

Figure 2 presents a comparison of the predicted accuracy of six different machine learning classifiers, including Logistic Regression (LR), Support Vector Machine (SVM), decision trees based on C4.5, K-Nearest Neighbors (KNN), Long Short-Term Memory (LSTM), and Bayesian Networks (BN). Using a percentage (%) as a measure of prediction accuracy, classifiers are measured according to their performance. There is a significant variation in accuracy among the five models, with BN achieving the highest accuracy at 79.42%, followed closely by LSTM (77.36%) and KNN (76.54), and LR recording the lowest accuracy at 70.36%. Even though all classifiers perform reasonably well, advanced models like LSTM and BN provide superior predictive capabilities compared to traditional methods like LR and SVM. This highlights their potential effectiveness in complex predictive analytics.

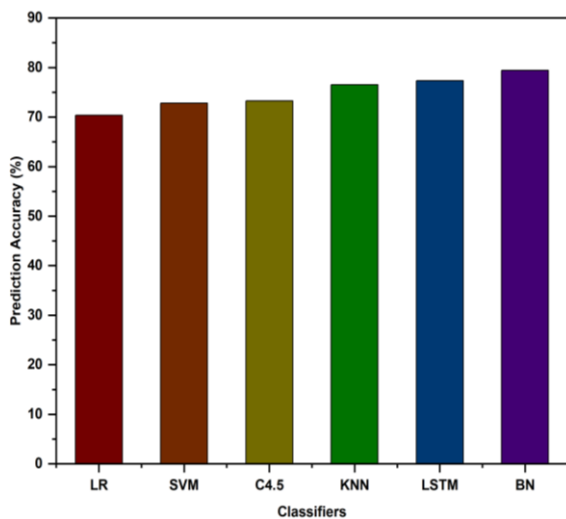


Figure 2: Accuracy of classifier predictions.

Figure 3 depicts the distribution of predictions across two classes, 0 and Class 1-for six different classifiers (LR, SVM, C4.5, KNN, LSTM, and BN) in relation to investments in renewable energy (RES). There is a significant dominance of class 0, represented in brown, across all classifiers, indicating that the majority of predictions fall into this category. All models predict Class 1 relatively few times, with SVM showing slightly higher numbers than other models. It may be an underlying class imbalance in the dataset or that the models are not able to identify investment opportunities (Class 1) in renewable energy as a result of the imbalance between Class 0 and Class 1.

According to Figure 4, six machine learning classifiers were used to classify the "Not Invest in Renewable Energy Sources (RES)" category: LR, SVM, C4.5, KNN, LSTM, and BN. For most model classifications, Class 0 (brown) consistently predicts non-investment decisions more accurately than Class 1 (green). As one of the most prominent features of the C4.5 classifier, almost no predictions are made for Class 1 as a result of the extreme skew towards Class 0. SVMs and LSTMs demonstrate somewhat better balance, though Class 0 continues to dominate. As a result, classifiers predict non-investment outcomes more accurately, even when class imbalances and subtle factors influencing investment decisions are present.

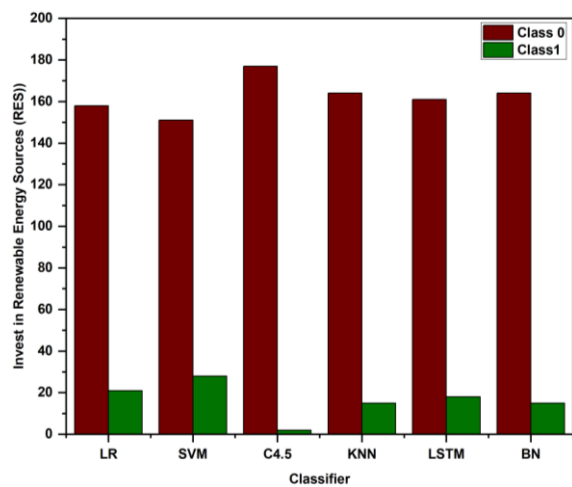


Figure 3: Class-wise prediction results for renewable energy investments.

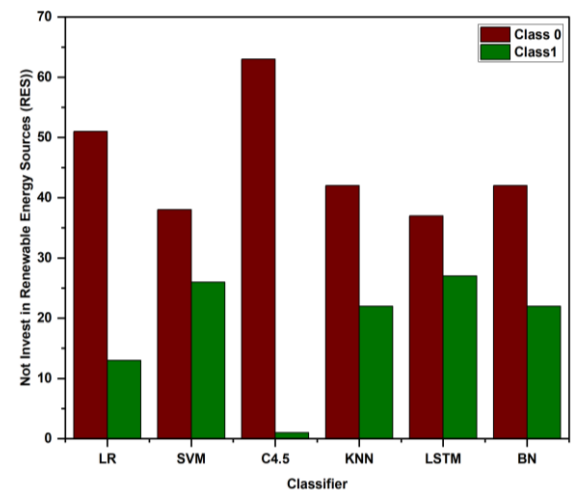


Figure 4: Prediction distribution for non-investment in renewable energy.

5 CONCLUSIONS

Using machine learning to predict renewable energy investment decisions can significantly improve investment decisions. Bayesian networks and LSTM models performed the best among the six classifiers evaluated, demonstrating their superior abilities to model complex relationships in renewable energy data. Using the class distribution analysis, it was revealed that non-investment predictions were dominant, suggesting that dataset imbalance and the complexity of identifying promising investment opportunities are challenges. Machine learning, however, can significantly enhance the accuracy of forecasts, optimize resource management strategies, and improve financial planning associated with sustainable energy projects by effectively capturing complex patterns and trends. Achieving a greener, more sustainable energy future will necessitate machine learning techniques capable of integrating broader, diverse datasets, addressing class imbalances, reducing predictive uncertainty, and enhancing decision-making reliability.

REFERENCES

- [1] N. Kumar, P. Rani, V. Kumar, P. K. Verma, and D. Koundal, "Teech: Three-tier extended energy efficient clustering hierarchy protocol for heterogeneous wireless sensor network," *Expert Syst. Appl.*, vol. 216, p. 119448, 2023.
- [2] A. Entezari, A. Aslani, R. Zahedi, and Y. Noorollahi, "Artificial intelligence and machine learning in energy systems: A bibliographic perspective," *Energy Strategy Rev.*, vol. 45, p. 101017, Jan. 2023, doi: 10.1016/j.esr.2022.101017.
- [3] D. Rangel-Martinez, K. D. P. Nigam, and L. A. Ricardez-Sandoval, "Machine learning on sustainable energy: A review and outlook on renewable energy systems, catalysis, smart grid and energy storage," *Chem. Eng. Res. Des.*, vol. 174, pp. 414–441, Oct. 2021, doi: 10.1016/j.cherd.2021.08.013.
- [4] M. Sharifzadeh, A. Sikinioti-Lock, and N. Shah, "Machine-learning methods for integrated renewable power generation: A comparative study of artificial neural networks, support vector regression, and Gaussian Process Regression," *Renew. Sustain. Energy Rev.*, vol. 108, pp. 513–538, Jul. 2019, doi: 10.1016/j.rser.2019.03.040.
- [5] J.-P. Lai, Y.-M. Chang, C.-H. Chen, and P.-F. Pai, "A Survey of Machine Learning Models in Renewable Energy Predictions," *Appl. Sci.*, vol. 10, no. 17, p. 5975, Aug. 2020, doi: 10.3390/app10175975.
- [6] K. Kumar, R. S. Rao, O. Kaiwartya, S. Kaiser, and S. Padmanaban, *Sustainable developments by artificial intelligence and machine learning for renewable energies*. Academic Press, 2022.
- [7] G. H. Gu, J. Noh, I. Kim, and Y. Jung, "Machine learning for renewable energy materials," *J. Mater. Chem. A*, vol. 7, no. 29, pp. 17096–17117, 2019.
- [8] N. Kumar, P. Rani, V. Kumar, S. V. Athawale, and D. Koundal, "THWSN: Enhanced energy-efficient clustering approach for three-tier heterogeneous wireless sensor networks," *IEEE Sens. J.*, vol. 22, no. 20, pp. 20053–20062, 2022.
- [9] I. Antonopoulos et al., "Artificial intelligence and machine learning approaches to energy demand-side response: A systematic review," *Renew. Sustain. Energy Rev.*, vol. 130, p. 109899, Sep. 2020, doi: 10.1016/j.rser.2020.109899.
- [10] P. Rani and R. Sharma, "An experimental study of IEEE 802.11 n devices for vehicular networks with various propagation loss models," in *International Conference on Signal Processing and Integrated Networks*, Springer, 2022, pp. 125–135.
- [11] L. Wang, G. Zhang, X. Yin, H. Zhang, and M. Ghalandari, "Optimal control of renewable energy in buildings using the machine learning method," *Sustain. Energy Technol. Assess.*, vol. 53, p. 102534, Oct. 2022, doi: 10.1016/j.seta.2022.102534.
- [12] P. Rani, S. P. Yadav, P. N. Singh, and M. Almusawi, "Real-World Case Studies: Transforming Mental Healthcare With Natural Language Processing," in *Demystifying the Role of Natural Language Processing (NLP) in Mental Health*, A. Mishra et al., Eds., IGI Global, 2025, pp. 303–324. doi: 10.4018/979-8-3693-4203-9.ch016.
- [13] L. Zhang, J. Ling, and M. Lin, "Artificial intelligence in renewable energy: A comprehensive bibliometric analysis," *Energy Rep.*, vol. 8, pp. 14072–14088, Nov. 2022, doi: 10.1016/j.egyr.2022.10.347.
- [14] P. Sharma et al., "Recent Advances in Machine Learning Research for Nanofluid-Based Heat Transfer in Renewable Energy System," *Energy Fuels*, vol. 36, no. 13, pp. 6626–6658, Jul. 2022, doi: 10.1021/acs.energyfuels.2c01006.
- [15] R. R. Batcha and M. K. Geetha, "A Survey on IOT Based on Renewable Energy for Efficient Energy Conservation Using Machine Learning Approaches," in *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, Jaipur, India: IEEE, Feb. 2020, pp. 123–128. doi: 10.1109/ICETCE48199.2020.9091737.
- [16] Z. Yao et al., "Machine learning for a sustainable energy future," *Nat. Rev. Mater.*, vol. 8, no. 3, pp. 202–215, Oct. 2022, doi: 10.1038/s41578-022-00490-5.
- [17] A. Raihan, "A comprehensive review of artificial intelligence and machine learning applications in energy sector," *J. Technol. Innov. Energy*, vol. 2, no. 4, pp. 1–26, 2023.
- [18] N. Hussain, P. Rani, N. Kumar, and M. G. Chaudhary, "A deep comprehensive research architecture, characteristics, challenges, issues, and benefits of routing protocol for vehicular ad-hoc networks," *Int. J. Distrib. Syst. Technol. IJDST*, vol. 13, no. 8, pp. 1–23, 2022.
- [19] S. Han, N. Zhang, and P. Zhang, "Optimal operation and scheduling of electric vehicle charging stations integrated with renewable energy resources," *Energy Procedia*, vol. 158, pp. 4470–4473, 2019.

- [20] H. Li and Y. H. Song, "Coordinated charging strategy of electric vehicles for renewable energy integration," *IEEE Trans. Sustain. Energy*, vol. 10, no. 2, pp. 649–660, 2019.
- [21] B. Wang, J. Wu, and X. Zhou, "Optimal planning for charging infrastructure considering renewable energy and vehicle-to-grid integration," *Appl. Energy*, vol. 228, pp. 127–137, 2018.
- [22] D. Bogdanov et al., "Low-cost renewable electricity as the key driver of the global energy transition towards sustainability," *Energy*, vol. 227, p. 120467, Jul. 2021, doi: 10.1016/j.energy.2021.120467.
- [23] P. Rani, S. Verma, S. P. Yadav, B. K. Rai, M. S. Naruka, and D. Kumar, "Simulation of the lightweight blockchain technique based on privacy and security for healthcare data for the cloud system," *Int. J. E-Health Med. Commun. IJEHMC*, vol. 13, no. 4, pp. 1–15, 2022.
- [24] A. Ali and S. Das, "Applications of neuro-computing and fractional calculus to blood streaming conveying modified trihybrid nanoparticles with interfacial nanolayer aspect inside a diseased ciliated artery under electroosmotic and Lorentz forces," *Int. Commun. Heat Mass Transf.*, vol. 152, p. 107313, Mar. 2024, doi: 10.1016/j.icheatmasstransfer.2024.107313.
- [25] P. Karmakar and S. Das, "A neural network approach to explore bioelectromagnetics aspects of blood circulation conveying tetra-hybrid nanoparticles and microbes in a ciliary artery with an endoscopy span," *Eng. Appl. Artif. Intell.*, vol. 133, p. 108298, Jul. 2024, doi: 10.1016/j.engappai.2024.108298.
- [26] J. Schallenberg-Rodríguez and N. García Montesdeoca, "Spatial planning to estimate the offshore wind energy potential in coastal regions and islands. Practical case: The Canary Islands," *Energy*, vol. 143, pp. 91–103, Jan. 2018, doi: 10.1016/j.energy.2017.10.084.
- [27] A. Chaouachi, C. F. Covrig, and M. Ardelean, "Multi-criteria selection of offshore wind farms: Case study for the Baltic States," *Energy Policy*, vol. 103, pp. 179–192, Apr. 2017, doi: 10.1016/j.enpol.2017.01.018.
- [28] L. Ou, W. Xu, Q. Yue, C. L. Ma, X. Teng, and Y. E. Dong, "Offshore wind zoning in China: Method and experience," *Ocean Coast. Manag.*, vol. 151, pp. 99–108, Jan. 2018, doi: 10.1016/j.ocecoaman.2017.10.016.
- [29] C.-D. Yue and M.-H. Yang, "Exploring the potential of wind energy for a coastal state," *Energy Policy*, vol. 37, no. 10, pp. 3925–3940, Oct. 2009, doi: 10.1016/j.enpol.2009.04.055.
- [30] S.-P. Breton and G. Moe, "Status, plans and technologies for offshore wind turbines in Europe and North America," *Renew. Energy*, vol. 34, no. 3, pp. 646–654, 2009.
- [31] H. Díaz and C. G. Soares, "An integrated GIS approach for site selection of floating offshore wind farms in the Atlantic continental European coastline," *Renew. Sustain. Energy Rev.*, vol. 134, p. 110328, 2020.
- [32] J.-Y. Kim, K.-S. Kang, K.-Y. Oh, J.-S. Lee, and M.-S. Ryu, "Assessment of possible resources and selection of preparatory sites for offshore wind farm around Korean peninsula," *New Renew. Energy*, vol. 5, no. 2, pp. 39–48, 2009.
- [33] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, pp. 37–63, 2011, doi: 10.9735/2229-3981.

Using the Program Evaluation and Review Technique and the Critical Path Method to Help Project Managers

Esraa Tayyeh Ehaimed and Mushtaq Abd Al Ghany

Department of Mathematics, College of Education for Pure Sciences, University of Babylon, 51002 Hila, Iraq
edu158.asra.tayeh@student.uobabylon.edu.iq, pure.mushtaq@uobabylon.edu.iq

Keywords: Critical Path Method (CPM), Program Evaluation and Review Technique (PERT), Critical Activity.

Abstract: This study examines two fundamental project management techniques: the Critical Path Method (CPM) and the Program Evaluation and Review Technique (PERT). CPM is a deterministic approach designed for repetitive projects with well-defined task durations, enabling the identification of critical activities that directly determine project completion time. PERT, in contrast, is a probabilistic method suited for projects with uncertain task durations, incorporating optimistic, most likely, and pessimistic time estimates to account for variability and risk. Through numerical applications and comparative analysis, the study highlights the methodological differences, computational procedures, and practical implications of each technique. The findings demonstrate that CPM is effective for structured, predictable projects, while PERT provides superior flexibility under conditions of uncertainty. By evaluating the strengths and limitations of both methods, this work offers practical guidance for project managers in selecting appropriate scheduling tools based on project complexity, uncertainty levels, and resource availability, ultimately supporting efficient project planning and timely completion.

1 INTRODUCTION

Networks can be used to describe and assess the success of most project kinds, as well as to identify specific outcomes like cutting down on project completion time and determining which tasks should be expedited to shorten the project's overall duration. As a result, the project's longest path from start to finish determines the minimum overall project time. In the 1950s [1], a helpful tool called the critical path technique emerged to assist managers in organizing and scheduling major projects. This technique complements the Gantt chart and can be used in place of it. In 1959 [2], Kelly and Walker drew attention to the application of (PERT) and (CPM) in project management in a range of industries. PERT and CPM differ in that PERT is mainly utilized in situations where the time needed to complete each task is uncertain, but the CPM approach is based on project execution on the understanding and expertise of previous projects [3], [4]. The primary market for the CPM was the construction sector. The use of CPM is a part of total quality management in health services, and it was tested in a pilot program [5]. A deterministic term analysis of a duration project can be performed with

(CPM) [6]. Total quality management aims to improve an organization's culture, practices, products, and services to satisfy consumers. In project management studies, planning is a crucial and challenging topic [7]. The complexity and dynamic nature of the undertaking both contribute to its difficulties. Plans are constantly being revised before and during their execution [8]. Project management methods, such as CPM, analyze project durations deterministically, finding the minimum intended duration [9]. Some techniques, like PERT, use data values for specific tasks as arbitrary variables, such as hopeful (shortest), most probable (middle), and pessimistic (longest) estimation times [10]. These methods help manage and oversee projects based on network dependencies and changes in project conditions [11]. The construction industry has been the primary user of CPM in the past [12]. CPM is among the different methods for carrying out project planning with PERT. CPM is utilized in projects consisting of multiple discrete tasks [13]. In (2022) [14], Bashiru and Douglas showed that CPM allows project managers to schedule work, track project objectives, and make necessary adjustments; it serves as a mirror for the project, enabling resource

planning and schedule preparation, and ensuring the project is executed correctly [15].

By determining the critical path a set of tasks that must be finished promptly in order for the project to be finished on time the critical path approach is a strategy for planning and organizing project schedules [16]. When the oldest and most recent times of an action are the same, it is said to be essential [17]. This work is concerned with presenting two useful techniques for project managers in scheduling, planning and determining the maximum possible time to complete projects, which is defined by the critical path, and from it determining the critical activities in the project, as well as some numerical applications of each technique and the most prominent differences between the two techniques.

2 SOME IMPORTANT DEFINITIONS

Network [18]: An interconnected geometric representation that includes two types of data: events and activities. It begins with one initial event called the beginning and ends with one final event called the end. Each activity contains a time that represents the time required to complete the activity

Activity [19]: is any straight or curved arrow that is not broken on the network that connects two events and has a beginning and an end. The activity takes time and resources, and no activity can be started until all activities are completed. The previous one.

Event [20]: In the network, an event is represented by a numbered circle, or it can be a numbered square or triangle that represents the beginning and end of the activity in a sequential manner and does not take time or resources. There is a head and a tail of the event, and there are types of events, merged events and burst events. Generally, the head events have a higher value than the tail events.

Predecessor event and successor event [7]: An event is predecessor if it comes directly before an event. An event is successor event if it follows an event directly. So, any event may be a predecessor event for an event and successor event for another events.

Critical path [21]: The bigger path connecting the network of project from beginning to end is made up of the network's critical activities and summation of all critical activities.

3 CRITICAL PATH METHOD (CPM)

It is a commonly recognized tool utilized to schedule and manage tasks in complex construction networks. CPM enables project managers to identify the sequence of tasks that determine the project's overall duration, known as the critical path [22]. After representing the project in the form of a network, the next step is calculating the total time to complete the project by CPM.

3.1 Forward Pass (Earliest Time)

The computations start with Activity 1 and advance recursively to Activity n , including the earliest start time (ES) and earliest finished time (EF).

In the first step:

$$E_1 = 0. \quad (1)$$

Second step: let i, j are two Activities, A, B There are two activities and D indicator to the duration between two Activities then

$$ESB = EFA + Di, j. \quad (2)$$

General Step j . Given that events $A, B, \dots$ and K are connected directly to the event j by incoming activities $(A, j), (B, j), \dots$ and (K, j) and that the earliest occurrence times of events $A, B, \dots$ and K have already been computed, then the earliest occurrence time of Activity j is

$$E_j = \max \{EA + DA, j, EB, j + DB, j, \dots, EK + DK, j\}. \quad (3)$$

The forward pass [23] is complete when En at event n has been finished, Ej is the longest path (duration) to the event j .

3.2 Backward Pass (Latest Times)

This computations start at node n and ends at node 1 and including latest start time (LS) and latest finished time (LF).

Firstly, for any network then:

$$L = E, \quad (4)$$

To indicate that latest time of the last event the same value of the total duration of the project.

Second step: Let i, j are two events and A, B are two activities between i, j then:

$$LSB = LFA - Di, j, \quad (5)$$

General Step j : Given that events $A, B, \dots$ and K are connected directly to the event j by outgoing activities $(j, A), (j, B) \dots$ and (j, K) and that the latest events $A, B \dots$ and K have computed the latest time is:

$$L_j = \min\{LA - Dj, A, LB - Dj, B \dots LK - Dj, K\} \quad (6)$$

The backward pass (Latest Times, L) [24], [25] ends at event 1 with:

$$L_1 = 0. \quad (7)$$

By above calculations, an activity $A = (i, j)$ Then the critical if it satisfies three conditions.

$$\begin{aligned} *-Ei &= Li, \quad **-Ej = Lj, \\ ***-Ej - Li &= Di, j. \end{aligned} \quad (8)$$

Any activity that does not meet all three requirements is considered noncritical [14].

4 THE PROGRAM EVALUATION AND REVIEW TECHNIQUE

PERT is a statistical technique used in project management that was created to describe and analyze the activities required to finish a project, particularly the amount of time needed to do each task and determine the minimal amount of time needed to finish the project as a whole [26]. The three estimates are plotted on a graph, with the probability distribution closely associated with the Beta distribution curve, which is unimodal and not symmetric [27]. Three primary sorts of estimations can be produced using probabilistic time estimation [28], [29]:

- 1) A pessimistic time estimate (T_p).
- 2) A hopeful time estimate (T_h).
- 3) A most probable time estimate (T_m).

The hopeful time estimate assumes a task can be completed quickly, while the pessimistic estimate anticipates unexpected issues, leading to increased time consumption. The most likely estimate falls between these two as calculated in Tables 1 and 2.

$$Tp \geq Tm \geq Th. \quad (11)$$

The most probable time estimate should be given a weight of 4, and the hopeful and pessimistic time predictions should be given a weight of 1. A time estimate (Te) is obtained as the weighted (mean) The following is the average of these estimations:

$$Te = \frac{Tp + 4Tm + Th}{6}, \quad (12)$$

and Variance (σ^2)

$$\sigma^2 = \left(\frac{Tp - Th}{6}\right)^2, \quad (13)$$

where the root of variance is known as the Standard deviation (σ) [28]

$$\sigma = \left(\frac{Tp - Th}{6}\right), \quad (14)$$

$$\sum T_{critical} = \text{Calculation of the total sum of anticipated critical values}, \quad (15)$$

$$V_{critical} = \text{Calculation of all variances of critical activities}. \quad (16)$$

PERT uses simplifying plan updates during network modifications.

Problem 1. To find project time by using CPM for the following activity network as in Figure 1.

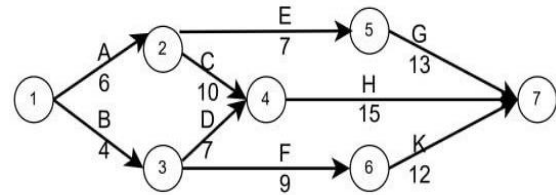


Figure 1: Network diagram.

Table 1: CPM application.

Activity	Duration (weeks)	ES	EF	LS	LF
A	6	0	6	6	0
B	4	0	4	9	0
C	10	6	16	16	6
D	7	4	16	16	9
E	7	6	13	18	6
F	9	4	13	19	9
G	13	13	31	31	18
H	15	16	31	31	16
K	12	13	31	31	19

The critical path is $A \rightarrow C \rightarrow H$. Therefore, the project time complete = Critical path length = 31 Weeks

Problem 2: We will use PERT to find the time complete to the following data.

From the three time estimates T_p, T_m and T_h calculate Te for each activity, show Table 3.

The four paths are as follows:

- 1) Path 1: $A \rightarrow C \rightarrow H$, Time path = $6 + 10 + 11 = 27$ weeks, Path 3: $A \rightarrow E \rightarrow G$, Time path = $6 + 7 + 7 = 20$ weeks

- 2) Path 2: $B \rightarrow D \rightarrow H$,Time path = $4 + 7 + 10 = 21$ weeks, Path 4: $B \rightarrow F \rightarrow K$,Time path = $4+7+5=16$ weeks

From the four paths. Maximum of {27,21,20,16} = 27. Path 1 is the critical path, i.e., $B \rightarrow M \rightarrow N \rightarrow K$. And the critical activities are B, M, N and K . Project time Complete = Critical path length =34 ,
i.e $\sum T_{critical} = 6 + 10 + 11 = 27 \text{ weeks}$ and the total variance of the critical path = Calculation of

all variances of essential activities i.e. $V_{critical} = 0.444 + 1.777+5.444=7.665$

The above problems show that the difference between PERT and CPM techniques is 4 weeks. Differences between PERT and CPM techniques in scheduling and planning projects which listed in Table 4.

Table 2: PERT application.

Activity	Predecessor Activity	Duration (weeks)	Pessimistic time estimate (T_p)	Most probable time estimate (T_m)	Hopeful time estimate (T_h)
A	-	6	9	6	3
B	-	4	6	4	2
C	A	10	7	5	3
D	B	7	18	5	4
E	A	7	9	8	3
F	B	9	16	8	6
G	E	13	8	7	5
H	D,C	15	3	2	1
K	F	12	8	5	2

Table 3: PERT Calculation results and variance (σ^2).

Activity	Duration (weeks)	T_p	$4T_m$	T_h	$T_e = \frac{T_p + 4T_m + T_h}{6}$	$\sigma^2 = \left(\frac{T_p - T_h}{6}\right)^2$
A	6	9	24	3	6	1
B	4	6	16	2	4	0.444
C	10	14	40	6	10	1.777
D	8	18	20	4	7	5.444
E	7	28	12	2	7	18.777
F	9	16	32	6	9	2.777
G	8	10	28	4	7	1
H	12	20	40	6	11	5.444
K	5	8	20	2	5	1

Table 4: Comparison between CPM and PERT.

PERT	CPM
1. PERT is used to manage uncertain project activities; it uses a probabilistic model. 2. PERT is used where the time required to complete the activities is unknown. 3. The network is constructed based on the events. 4. Because jobs are non-repetitive, PERT has been given attention in areas like research and development programs. 5. Critical and non-critical activities are not differentiated 6. Management of Unpredictable Activities. 7. While notations like arrow diagram, node, task, float, etc. are always used in CPM, PERT uses notations like network, event activity, slack, etc.	1. CPM is a tool used to manage statistics well. It uses a deterministic model. 2. In CPM, there is just one estimate 3. CPM is used mainly for construction programs. 4. Critical and non-critical activities are differentiated 5. Management of predictable Activities 6. CPM notations like arrow diagram, event, function, slack, etc., are invariably used.

5 CONCLUSIONS

This work explores the use of PERT and CPM in project management. CPM is accurate for specific activities, while PERT works on uncertain projects by estimating three different periods. Combining both methods can improve planning and scheduling, reducing potential obstacles, which helps project managers to achieve the best goals within the project time frame. When using the CPM on a project containing nine activities, the total time was 31 weeks as a maximum to complete the project without delaying the delivery date, while when conducting the PERT method on the same project after entering the three most likely ominous and hopeful estimates, the total time was 27 weeks, a difference of 4 weeks from CPM. CPM calculates total time sequentially based on activities, while PERT prioritizes activities without strictly sequencing them. Both methods are successful in planning, scheduling, and managing large projects, and selecting the appropriate method depends on the project's complexity, level of uncertainty, available resources, and specific management goals to ensure optimal efficiency and timely completion.

REFERENCES

- [1] A. M. Aliyu, "Project management using critical path method (CPM): A pragmatic study," *Global Journal of Pure and Applied Sciences*, vol. 18, no. 3-4, pp. 197–206, 2013.
- [2] J. J. Kelly and M. Walker, "Critical path planning and scheduling," in *Proceedings of the Eastern Joint Computer Conference*, 1959, pp. 160–173.
- [3] A. H. Abdullah et al., "A modern approach of solution of assignment problems," in *8th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, Turkiye, 2024, pp. 1–4.
- [4] M. S. M. Zabiba et al., "A new technique to solve the maximization of the transportation problems," in *AIP Conference Proceedings*, vol. 2414, p. 040042, 2023.
- [5] P. A. Hofmann, "Critical path method: an important tool for coordinating clinical care," *Jt Comm J Qual Improv*, vol. 19, no. 7, pp. 235–246, 1993.
- [6] Y. A. Hussein and M. A. K. Shiker, "Using the largest difference method to find the initial basic feasible solution to the transportation problem," *Journal of Interdisciplinary Mathematics*, vol. 25, no. 8, pp. 2511–2517, 2022.
- [7] H. J. Kadhim et al., "A new technique for finding the optimal solution to assignment problems with maximization objective function," *J. Phys.: Conf. Ser.*, vol. 1963, p. 012104, 2021.
- [8] M. M. Mahdi et al., "Hybrid spectral algorithm under a convex constrained to solve nonlinear equations," *Journal of Interdisciplinary Mathematics*, vol. 25, no. 5, pp. 1333–1340, 2022.
- [9] L. H. Hashim et al., "An application comparison of two negative binomial models on rainfall count data," *J. Phys.: Conf. Ser.*, vol. 1818, p. 012100, 2021.
- [10] N. K. Dreeb et al., "Using a new projection approach to find the optimal solution for nonlinear systems of monotone equation," *J. Phys.: Conf. Ser.*, vol. 1818, p. 012101, 2021.
- [11] K. Amini et al., "A line search trust-region algorithm with nonmonotone adaptive radius for a system of nonlinear equations," *4OR*, vol. 14, no. 2, pp. 133–152, 2016.
- [12] D. H. Allawi and M. A. K. Shiker, "A modified technique of spectral gradient projection method for solving non-linear equations systems," *Journal of Interdisciplinary Mathematics*, vol. 27, no. 4, pp. 655–665, 2024.
- [13] M. Bashuru and K. B. Douglas, "Optimal project management schedule for PhD studies with focus on selected Northern sector universities in Ghana using CPM and PERT," *American Journal of Operations Research*, vol. 12, no. 1, pp. 1–7, 2022.
- [14] S. M. Kadham and M. A. Mustafa, "Medical applications of the new-transform," *Journal of Interdisciplinary Mathematics*, vol. 26, no. 6, pp. 1341–1353, 2023, doi: 10.47974/JIM-1632.
- [15] H. H. Dwail et al., "CG method with modifying β_k for solving unconstrained optimization problems," *Journal of Interdisciplinary Mathematics*, vol. 25, no. 5, pp. 1347–1355, 2022.
- [16] Z. S. Mahdi et al., "Solving transportation problems by using modification to Vogel's approximation method," in *AIP Conference Proceedings*, vol. 2834, p. 080110, 2023.
- [17] H. A. Taha, *Operations Research: An Introduction*, 10th ed., Global Edition. Fayetteville, AR, USA: Pearson, 2017.
- [18] Z. K. Hashim et al., "An easy technique to reach the optimal solution to the assignment problems," in *AIP Conference Proceedings*, vol. 2414, p. 040045, 2023.
- [19] H. S. Habib and M. A. K. Shiker, "A modified (CG) method for solving nonlinear systems of monotone equations," *Journal of Interdisciplinary Mathematics*, vol. 27, no. 4, pp. 787–792, 2024.
- [20] H. J. Kadhim et al., "New technique for finding the maximization to transportation problems," *J. Phys.: Conf. Ser.*, vol. 1963, p. 012070, 2021.
- [21] M. A. Saeed, H. Kh. Alkhayyat, and S. M. Kadham, "Enhance MRI images of lung cancer using hybrid transform," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 26, no. 7, pp. 1897–1902, 2023, doi: 10.47974/JDMSC-1682.
- [22] P. R. Murthy, *Operations Research*, 2nd ed. New Delhi, India: New Age International (P) Limited Publishers, 2007.
- [23] U. C. Orumie, "Implementation of Project Evaluation and Review Technique (PERT) and Critical Path Method (CPM): A Comparative Study," *International Journal of Industrial and Operations Research*, 2020.
- [24] R. A. Fadhil and Z. A. H. Hassan, "A hybrid The Dwarf Mongoose Optimization Algorithm with Nelder-Mead method and its application for allocation reliability," *Iraqi Journal of Science*, vol. 65, no. 7, pp. 3850–3859, 2024.

- [25] R. Fadhil and Z. Hassan, "Improvement of network reliability by hybridization of the penalty technique based on metaheuristic algorithms," *Iraqi Journal for Computer Science and Mathematics*, vol. 5, no. 1, pp. 99–111, 2024.
- [26] A. M. Mohammed and Z. Al-Khafaji, "A new technique to find the importance of reliability for a mixed-series system," in *AIP Conference Proceedings*, vol. 3097, no. 1, 2024.
- [27] N. S. Hassan and Z. Al-Khafaji, "Evaluation of the reliability and importance of the units in the minimal cut and minimal path for a complex network," in *AIP Conference Proceedings*, vol. 3097, no. 1, May 2024.
- [28] F. Alsharify and Z. Al-Khafaji, "A review of optimization techniques: Applications and comparative analysis," *Iraqi Journal for Computer Science and Mathematics*, vol. 5, no. 2, pp. 122–134, 2024.
- [29] R. A. Fadhil and Z. A. H. Hassan, "A hybrid honey-badger intelligence algorithm with Nelder-Mead method and its application for reliability optimization," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 4s, pp. 136–145, 2023.

House Price Prediction Using Diverse Machine Learning Techniques

Sabyasachi Pramanik¹, Abdulkhaleq Husham Yousif², Salah Jasim³, Raushan Raj¹,
Muskan Kumari¹, Atanu Roy⁴ and Ahmed J. Obaid⁵

¹Haldia Institute of Technology, 721657 Haldia, West Bengal, India

²Al-Iraqia University, Research and Studies Center, 10021 Baghdad, Iraq

³Department of Civil Engineering, Dijlah University College, 10021 Baghdad, Iraq

⁴Department of Marine Engineering, The Neotia University, 700160 Kolkata, India

⁵Faculty of Computer Science and Mathematics, University of Kufa, 54003 Najaf, Iraq

sabyasachi.pramanik@hithaldia.ac.in, abdulkhaleq.h.yousif@aliraqia.edu.iq, Salah.jasim@duc.edu.iq,
raushanraj9950@gmail.com, muskanp1803@gmail.com, atanuroy@tnu.in, abdulkhaleq.h.yousif@aliraqia.edu.iq

Keywords: House Price Prediction, Lasso Regression, Decision Tree Regressor, Real Estate Market Trends, Feature Engineering, Predictive Modeling in Real Estate.

Abstract: With the rapid growth of society and evolving market demands, understanding market trends has become increasingly crucial. Accurate prediction of house prices based on current trends is vital for informed decision-making. It enables individuals to plan their financial needs effectively and align them with their goals. As a continually expanding industry, the real estate sector plays a significant role in this context. For investors, identifying market patterns is essential for making strategic investments that can maximize returns. However, the lack of transparency in real estate pricing, often influenced by inflated rates set by intermediaries, poses challenges for clients. The availability of extensive datasets has opened new possibilities for researchers to create predictive models with improved accuracy. Traditional models often face lower precision and overfitting issues, which reduce their effectiveness. In contrast, the proposed system addresses these challenges, offering a robust and efficient model complemented by an intuitive ui. The primary goal of this research is to create an all-encompassing solution that benefits both businesses and individuals, reducing manual efforts while saving time and money. This system utilizes several ML techniques, such as Linear Regression, Lasso Regression, and Decision Tree. These algorithms are integrated using the stacking technique to enhance performance and accuracy. The proposed approach aims to deliver a user-friendly and reliable tool that simplifies real estate decision-making while ensuring precise predictions.

1. INTRODUCTION

Predicting the sale price of houses involves analyzing various factors that influence property values. These factors include the property's size, location, construction materials, and age, the number of bedrooms and garages, and additional amenities. By leveraging machine learning algorithms, this study aims to develop an accurate predictive model for house prices. The algorithms used in this research include Linear Regression, Lasso Regression, and Decision Tree Regressor, which are ideal for managing both linear and non-linear data relationships. Linear Regression provides a straightforward approach to model the relationship between input features and the target variable, while

Lasso Regression adds regularization to reduce overfitting by penalizing less relevant features. The Decision Tree Regressor, on the other hand, offers flexibility in capturing non-linear dependencies and interactions between features, making it a powerful tool for this predictive task. The real estate sector is a dynamic and highly lucrative investment domain, not just within individual countries but globally. Property pricing is influenced by numerous criteria, such as square footage, number of rooms, balconies, and proximity to essential services. Additionally, factors like community reviews and the surrounding infrastructure play a pivotal role in determining property values. These elements often change over time, contributing to fluctuations in pricing trends.

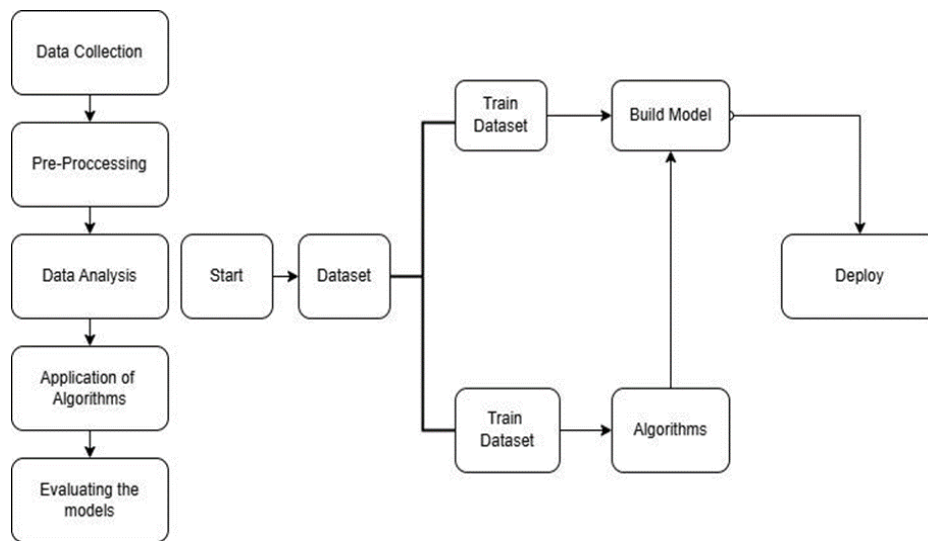


Figure 1: Block diagram for prediction of house prices.

For various stakeholders – house buyers, property owners, brokers, and investors – predicting house prices accurately are of significant importance. The volatile nature of real estate markets makes it essential to identify and analyze the features that drive property prices. The main aim of this study is to integrate these features into a machine learning framework that delivers reliable and accurate predictions. To achieve this, the process begins with collecting comprehensive datasets, followed by meticulous data cleaning to ensure accuracy. This involves addressing missing values, eliminating outliers, and converting raw data into an appropriate format. By integrating multiple predictive features into a robust machine learning pipeline, this study aims to model the complex relationships that influence property prices, resulting in reliable and actionable insights for stakeholders. The Property prediction factors will depend upon some of the nearby features which are shown in Figure 1.

2 LITERATURE REVIEW

The field of house price prediction [1] has advanced significantly in recent years, with the adoption of machine learning methods playing a crucial role. More robust and sophisticated approaches have progressively replaced traditional models that relied solely on linear relationships between features and prices. This evolution is driven by the increasing availability of real-time [2] datasets, algorithm advancements, and the need for actionable insights in

a volatile real estate market. Recent research has highlighted the value of integrating various machine learning models to enhance prediction precision. Algorithms like Linear Regression [3], Lasso Regression [4], and Decision Tree Regressor [5] remain fundamental, but there has been a growing preference for ensemble techniques like Random Forest and Gradient Boosting, which better capture nonlinear patterns. Additionally, techniques like Neural Networks [6] and Boosted Regression Models have proven effective in identifying intricate trends within real estate data. For example, hybrid models that integrate boosted algorithms with neural networks have been proposed to refine predictions by taking feedback from multiple regression algorithms and iteratively enhancing model performance. Real estate prices are influenced by an array of features, including location, infrastructure, property size, and amenities. Emerging factors, such as proximity to green spaces, access to smart city features, and energy efficiency, are becoming increasingly relevant in prediction models. Furthermore, external variables like economic trends, interest rates, and global events (e.g., the COVID-19 pandemic) have been found to exert significant influence on housing markets, underscoring the need for dynamic models that adapt to changing conditions.

2.1 Challenges and Opportunities

Despite advancements, challenges persist. Real estate markets are subject to fluctuations caused by macroeconomic variables, regional policies, and

unpredictable events, which can introduce noise into datasets. Moreover, inconsistencies in online listings, such as outdated information or deliberately undervalued properties, complicate prediction efforts. A recent study on real estate in Madrid's Salamanca district addressed these issues by developing a real-time machine learning application that identifies undervalued properties. This use case highlighted the potential for leveraging regression models to estimate market prices and uncover investment opportunities.

2.2 Economic and Social Implications

The real estate sector is a cornerstone of economic stability, representing both a basic need and a significant wealth indicator. In India, which ranks second globally in household numbers and continues to experience rapid economic growth (7% in 2024), predicting property values is a critical economic index. Accurate price prediction models can help not only buyers and investors but also policymakers and financial institutions by providing insights into market trends, mitigating risks, and promoting transparency.

3 RESEARCH METHODOLOGY

3.1 Linear Regression

Linear Regression is one of the most fundamental techniques in supervised machine learning [7], where the goal is to establish a linear relationship between the dependent and predictor variables. In the context of house price prediction [8], this technique estimates a house's value using features like total area, room count, location, and more.

In Simple Linear Regression, the relationship between a single independent variable x (e.g., the square footage of a house) and the dependent variable y (e.g., the house price) is modeled using the (1):

$$y = mx + c \quad (1)$$

Where:

- y is the predicted house price;
- m is the coefficient. (Reflecting how much the price increases as the feature x changes);
- x is the feature (e.g., square footage);
- c is the intercept (the predicted price when $x=0$).

In Multiple Linear Regression, the equation extends to multiple predictor variables. Here, the

price of a house is influenced by multiple factors (e.g., square footage, number of rooms, location). The model is expressed as:

$$y = m_1x_1 + m_2x_2 + \dots + m_kx_k + c \quad (2)$$

Where:

- $x_1, x_2, \dots, x_k$ are the multiple features or input variables;
- $m_1, m_2, \dots, m_k$ are the corresponding coefficients (weights) for each feature.

The linear regression model seeks to reduce the sum of squared residuals which evaluates the deviation of predictions from actual values. The model's effectiveness can be assessed using the Mean Squared Error (MSE) [9], computed as:

$$MSE = \frac{1}{k} \sum_{i=1}^k (e_i - \hat{e}_i)^2 \quad (3)$$

Where:

- e_i is the actual value;
- $\hat{e}_i$ is the predicted value.

3.2 Lasso Regression

Lasso Regression is a modification of Linear Regression that adds a penalty term, specifically L1 regularization [10], to the model. The primary objective of Lasso is to shrink the coefficients of less important features to zero, thereby performing feature selection. This produces a streamlined model that reduces the risk of overfitting [11], particularly in datasets with numerous features.

The Lasso Regression model is expressed as:

$$L(\theta) = \sum_{i=0}^n (y_i - \hat{y}_i)^2 + \sum_{j=0}^m |\theta_j| \quad (4)$$

Where:

- λ represents the regularization parameter that determines the intensity of the penalty.
- θ_j represents the model coefficients.
- The first term represents the sum of squared errors, and the second adds a penalty based on the coefficients' absolute values.

The strength of the regularization is controlled by the parameter λ . A large λ leads to more aggressive shrinkage of the coefficients, while a smaller λ makes the model behave more like ordinary linear regression.

No dot should be included after the sub subsection title number.

Lasso is especially useful for datasets with numerous features, as it simplifies the model by

eliminating less significant ones thereby simplifying the interpretation and improving generalization.

3.3 Decision Tree Regressor

A Decision Tree Regressor is a non-linear approach that makes predictions through repeated splits of the input space determined by specific feature thresholds [12]. At each step, the model selects the feature and threshold that minimizes variance within each partition, ultimately forming terminal nodes (leaves) where predictions are made.

The Decision Tree algorithm works as follows:

- 1) The feature space is divided based on a feature and threshold that minimizes the Mean Squared Error (MSE) within the resulting partitions.
- 2) The process continues iteratively, splitting each subset of data until a specified stopping condition is met.
- 3) Once the tree is constructed, predictions for new data points are made by navigating from the root to a leaf, where the output is the mean of the target values within that leaf.

The model's performance can be evaluated using MSE or other metrics like Root Mean Squared Error, which penalizes larger errors more heavily:

$$\text{RMSE} = \sqrt{\frac{1}{d} \sum_{p=0}^d (y_p - \hat{y}_p)^2}. \quad (5)$$

Where:

- y_p is the actual value;
- $\hat{y}_p$ is the predicted value.

A major advantage of Decision Trees is their capacity to model intricate, non-linear relationships between features and the target variable. However, they are susceptible to overfitting, especially when the tree becomes excessively deep. This problem can be mitigated using methods like pruning or by applying ensemble techniques such as Random Forests.

4 RESULTS AND DISCUSSION

Various metrics are often utilized to measure the accuracy and effectiveness of predictive models [13]:

4.1 Mean Absolute Error (MAE)

MAE calculates the average size of errors in the model's predictions, ignoring whether they are overestimations or underestimations. It is represented as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|. \quad (6)$$

MAE provides a straightforward interpretation of how far off the predictions are, but it does not differentiate between small and large errors.

4.2 Root Mean Squared Error (RMSE)

RMSE quantifies the average error, with a higher penalty for larger discrepancies. It is widely used in regression tasks to assess the performance of the model. The formula is:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (7)$$

Where:

- y_i is the actual value;
- $\hat{y}_i$ is the predicted value.

RMSE provides a more sensitive measure of model performance, particularly in cases where larger errors are more problematic.

4.3 Mean Squared Error (MSE)

MSE is a widely used metric to assess model performance, calculated as the average of the squared differences between actual and predicted values:

$$\text{MSE} = \frac{1}{L} \sum_{j=1}^L (s_j - \hat{s}_j)^2. \quad (8)$$

MSE gives greater weight to larger errors compared to MAE [14], offering a more comprehensive evaluation of the model's accuracy.

4.4 R² (Variance Score)

The R² score, or coefficient of determination, measures how much of the variance in the target variable is explained by the independent variables.

$$R^2=1 - \frac{\sum(y_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}. \quad (9)$$

Where y_i represent the mean of the actual values. A larger R^2 value indicates that the model explains a significant share of the variance in the target variable, reflecting better performance

4.5 Dataset

Data Description - this project utilizes a dataset from Kaggle, centered on predicting house prices in Bengaluru. Key features include location, total square footage (total_sqft), and number of bathrooms (bath), house price, and the number of bedrooms (bhk), all of which significantly influence property value. Other important attributes, such as area type, availability, society, and balcony, offer valuable context that helps improve the precision of price predictions. By thoroughly analyzing these attributes, the model aims to deliver precise price estimations based on the distinct characteristics of each property (Fig. 2).

4.5.1 Data Cleansing

Data cleaning [15] is crucial for the effectiveness of machine learning models, especially in projects like house price prediction, where the quality of the data significantly affects the reliability of the predictions. The task is to ensure that the data is free from errors, inconsistencies, and irrelevant features, thereby

allowing the model to learn from clean, structured, and relevant input. In this study, data collected from various sources were preprocessed and cleaned to eliminate these issues. The data cleaning steps followed in this project were designed to remove missing values, handle outliers, transform categorical data, and engineer meaningful features. The dataset initially contained missing values in some of the columns, like "society" and "balcony." Since these features had minimal impact on the target variable (price), they were dropped. Missing values in crucial columns such as "total_sqft" and "price" were also removed to ensure the integrity of the dataset.

4.5.2 Data Visualization

Data visualization plays a crucial role in data analysis projects by aiding in the understanding of complex relationships within the data. It allows us to explore and interpret data through various graphical representations such as histograms, scatter plots, and bar charts. In this project, the following data visualizations were generated to better understand the dataset and the relationships between features, as well as to identify patterns that may influence house prices.

Histogram of Price per Square Foot. The histogram illustrates a skewed distribution, where most houses are priced lower per square foot, while a few high-priced outliers stand apart (Fig. 3).

	area_type	availability	location	size	society	total_sqft	bath	balcony	price
0	Super built-up Area	19-Dec	Electronic City Phase II	2 BHK	Coomee	1056	2.0	1.0	39.07
1	Plot Area	Ready To Move	Chikka Tirupathi	4 Bedroom	Theanmp	2600	5.0	3.0	120.00
2	Built-up Area	Ready To Move	Uttarahalli	3 BHK	NaN	1440	2.0	3.0	62.00
3	Super built-up Area	Ready To Move	Lingadheeranahalli	3 BHK	Soiewre	1521	3.0	1.0	95.00
4	Super built-up Area	Ready To Move	Kothanur	2 BHK	NaN	1200	2.0	1.0	51.00

Figure 2: The dataset.

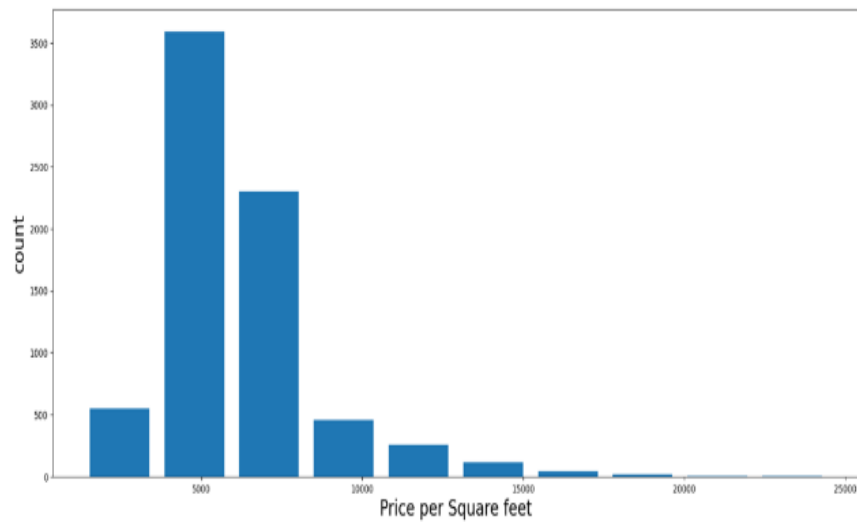


Figure 3: Histogram of price per square foot.

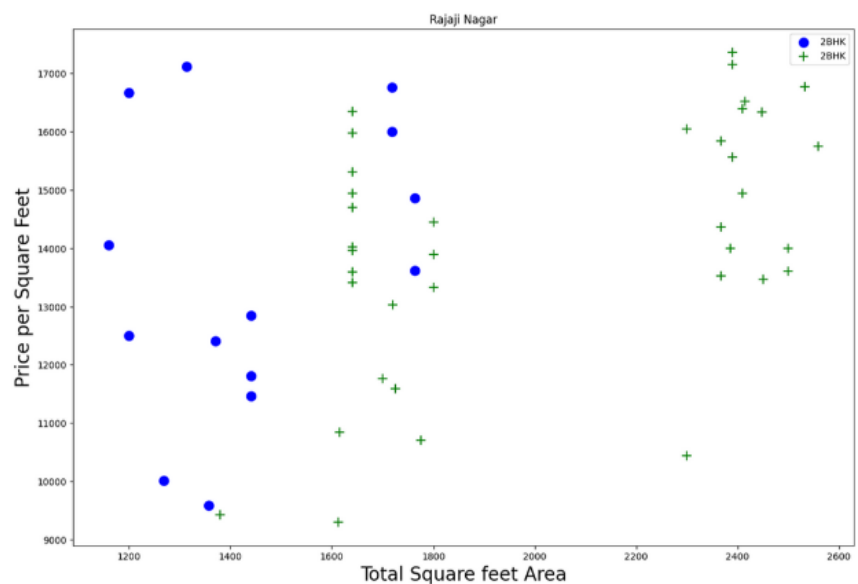


Figure 4: Scatter plot of price vs. total square feet.

Scatter Plot of Price vs. Total Square Feet. A scatter plot was used to visualize how the total square footage of a house correlates with its price (Fig. 4), helping to identify trends and patterns in pricing relative to house size. This helps in understanding how price increases with the size of the house. By plotting different colors for different "bhk" (bedroom count), we could also see how the price varies by house type.

4.5.3 Future Directions

Looking ahead, advancements in Artificial Intelligence (AI) and Big Data will likely transform real estate analytics. Predictive models are expected to integrate real-time data from Internet of Things (IoT) devices, blockchain-based property records,

and advanced geospatial analysis. Additionally, algorithms will need to account for climate change impacts, such as flood risks or rising temperatures, which could significantly influence property valuations. Another promising avenue involves the use of Explainable AI (XAI) to make machine learning models more transparent and interpretable for stakeholders like policymakers, realtors, and investors.

5 CONCLUSIONS

In conclusion, predicting property and residential prices using machine learning has proven to be valuable in understanding the ever-changing nature of the real estate market. The prices of properties are determined by multiple factors, including location, size, and amenities, all of which need to be precisely captured to make accurate predictions. As we continue to harness the power of machine learning algorithms, the insights derived from these predictions offer significant potential for both immediate decision-making and long-term strategic planning. The effectiveness of the model does not solely rely on the precision of the predictions but rather on how well these predictions are applied in real-world scenarios. By using the predicted values effectively, we can create models that provide predictions and actionable insights that help stakeholders—whether investors, homeowners, or policymakers—make informed decisions. The adaptability of the model to various use-cases and its potential to evolve with new data ensure its relevance in future applications. However, challenges remain, particularly in data collection. As the real estate market evolves, obtaining high-quality, up-to-date data continues to be a time-consuming and complex process. While open data sources provide some convenience, they may not always be directly applicable or sufficient for accurate predictions. Moreover, the volume of data is growing rapidly, and the risk of data becoming outdated is an ongoing concern. To address these challenges, continuous updates to the dataset and the refinement of prediction models are necessary to maintain their relevance. Looking forward, there are abundant opportunities for growth in this domain. The need for enhanced data collection, better feature engineering, and more sophisticated algorithms is paramount. With improved data retention strategies and advancements in machine learning, the potential for even more

accurate and efficient real estate price predictions is vast. As we refine the prediction process and explore new ways of utilizing these predictions, we can anticipate a future where machine learning models in real estate will drive innovation, creating smarter and more resilient housing markets.

REFERENCES

- [1] Y. Li and Q. Chen, "Predicting House Prices Using Machine Learning Algorithms: A Comparative Study," *International Journal of Data Science and Analytics*, vol. 9, no. 4, pp. 365-380, 2020.
- [2] S. Jain and P. Kumar, "An Overview of Machine Learning Algorithms for Real Estate Price Prediction," *Journal of Artificial Intelligence in Real Estate*, vol. 10, no. 2, pp. 112-130, 2021.
- [3] A. Kumar and N. Sharma, "Machine Learning in Real Estate: Addressing Challenges and Future Opportunities," *Journal of Smart Cities and Machine Learning*, vol. 15, no. 1, pp. 45-62, 2022.
- [4] R. Chaudhary and S. Patel, "Forecasting Housing Prices Using Ensemble Learning Techniques: A Case Study," *International Journal of Artificial Intelligence and Data Mining*, vol. 17, no. 3, pp. 201-220, 2023.
- [5] R. Sinha and A. Gupta, "Improving Real Estate Price Predictions through Hybrid Machine Learning Models," *Journal of Real Estate Analytics*, vol. 21, no. 4, pp. 325-340, 2024.
- [6] J. Zhang and H. Liu, "Dynamic Real Estate Price Forecasting Using Reinforcement Learning," *Journal of Machine Learning for Finance and Real Estate*, vol. 22, no. 1, pp. 78-95, 2024.
- [7] M. Patel and P. Rath, "Data-Driven Approach to Real Estate Price Predictions: Leveraging Big Data and AI," *Journal of Real Estate Technology and Innovation*, vol. 12, no. 2, pp. 150-169, 2023.
- [8] R. Sharma and P. Agarwal, "Predicting Property Prices Using Hybrid Deep Learning Models: A Comparative Study," *Journal of Computational Economics*, vol. 14, no. 1, pp. 88-103, 2022.
- [9] S. Singh and D. Verma, "Real-Time Real Estate Market Analysis Using Machine Learning and Neural Networks," *Journal of Urban Studies and Data Science*, vol. 19, no. 3, pp. 223-241, 2024.
- [10] V. Patel and P. Bansal, "A Comprehensive Review of Machine Learning Techniques for Predicting Property Prices," *Journal of Applied Artificial Intelligence*, vol. 18, no. 2, pp. 177-195, 2023.
- [11] A. Roy and S. Das, "Application of Gradient Boosting Models in Housing Price Predictions: A Comparative Analysis," *Journal of Advanced Data Analytics in Real Estate*, vol. 11, no. 3, pp. 120-139, 2023.
- [12] T. Wang and L. Zhao, "Leveraging Transfer Learning for Real Estate Valuation Across Geographies," *International Journal of Artificial Intelligence Applications in Urban Planning*, vol. 16, no. 4, pp. 289-310, 2024.
- [13] P. Mehta and N. Choudhary, "Deep Reinforcement Learning Models for Real-Time Property Price Predictions," *Journal of Computational Intelligence and Smart Systems*, vol. 13, no. 1, pp. 65-82, 2023.

- [14] F. Ahmed and S. Khan, "Real Estate Market Forecasting Using LSTM Networks: A Regional Study," *Journal of Machine Learning Applications in Real Estate*, vol. 20, no. 2, pp. 142-160, 2024.
- [15] X. Wei and R. Sun, "Exploring Feature Engineering Techniques for Housing Price Prediction with ML Models," *International Journal of Data Engineering and Artificial Intelligence*, vol. 9, no. 5, pp. 321-340.