

Ming Yang
Sachi Nandan Mohanty
Suneeta Satpathy
Shu Hu *Editors*

Demystifying AI and ML for Cyber–Threat Intelligence

Information Systems Engineering and Management

Volume 43

Series Editor

Álvaro Rocha, ISEG, University of Lisbon, Lisbon, Portugal

Editorial Board

Abdelkader Hameurlain, Université Toulouse III Paul Sabatier, Toulouse, France


Ali Idri, ENSIAS, Mohammed V University, Rabat, Morocco

Ashok Vaseashta, International Clean Water Institute, Manassas, VA, USA


Ashwani Kumar Dubey , Amity University, Noida, India

Carlos Montenegro, Francisco José de Caldas District University, Bogota, Colombia

Claude Laporte, University of Quebec, Québec, QC, Canada

Fernando Moreira , Portucalense University, Berlin, Germany

Francisco Peñalvo, University of Salamanca, Salamanca, Spain

Gintautas Dzemyda , Vilnius University, Vilnius, Lithuania


Jezreel Mejia-Miranda, CIMAT - Center for Mathematical Research, Zacatecas, Mexico

Jon Hall, The Open University, Milton Keynes, UK

Mário Piattini , University of Castilla-La Mancha, Albacete, Spain

Maristela Holanda, University of Brasilia, Brasilia, Brazil

Mincong Tang, Beijing Jiaotong University, Beijing, China

Mirjana Ivanović , Department of Mathematics and Informatics, University of Novi Sad, Novi Sad, Serbia

Mirna Muñoz, CIMAT Center for Mathematical Research, Progreso, Mexico

Rajeev Kanth, University of Turku, Turku, Finland

Sajid Anwar, Institute of Management Sciences, Peshawar, Pakistan

Tutut Herawan, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

Valentina Colla, TeCIP Institute, Scuola Superiore Sant'Anna, Pisa, Italy

Vladan Devedzic, University of Belgrade, Belgrade, Serbia

The book series “Information Systems Engineering and Management” (ISEM) publishes innovative and original works in the various areas of planning, development, implementation, and management of information systems and technologies by enterprises, citizens, and society for the improvement of the socio-economic environment.

The series is multidisciplinary, focusing on technological, organizational, and social domains of information systems engineering and management. Manuscripts published in this book series focus on relevant problems and research in the planning, analysis, design, implementation, exploration, and management of all types of information systems and technologies. The series contains monographs, lecture notes, edited volumes, pedagogical and technical books as well as proceedings volumes.

Some topics/keywords to be considered in the ISEM book series are, but not limited to: Information Systems Planning; Information Systems Development; Exploration of Information Systems; Management of Information Systems; Blockchain Technology; Cloud Computing; Artificial Intelligence (AI) and Machine Learning; Big Data Analytics; Multimedia Systems; Computer Networks, Mobility and Pervasive Systems; IT Security, Ethics and Privacy; Cybersecurity; Digital Platforms and Services; Requirements Engineering; Software Engineering; Process and Knowledge Engineering; Security and Privacy Engineering, Autonomous Robotics; Human-Computer Interaction; Marketing and Information; Tourism and Information; Finance and Value; Decisions and Risk; Innovation and Projects; Strategy and People.

Indexed by Google Scholar. All books published in the series are submitted for consideration in the Web of Science.

For book or proceedings proposals please contact Alvaro Rocha (amrrocha@gmail.com).

Ming Yang · Sachi Nandan Mohanty ·
Suneeta Satpathy · Shu Hu
Editors

Demystifying AI and ML for Cyber–Threat Intelligence

Editors

Ming Yang
Department of Information Technology
College of Computing and Software
Engineering
Kennesaw State University
Kennesaw, GA, USA

Suneeta Satpathy
Center For Cyber Security
SoA (Deemed to be University)
Bhubaneswar, Odisha, India

Sachi Nandan Mohanty
School of Computer Science
and Engineering
VIT-AP University
Vijayawada, Andhra Pradesh, India

Shu Hu
School of Applied and Creative Computing
Purdue University
West Lafayette, IN, USA

ISSN 3004-958X ISSN 3004-9598 (electronic)
Information Systems Engineering and Management
ISBN 978-3-031-90722-7 ISBN 978-3-031-90723-4 (eBook)
<https://doi.org/10.1007/978-3-031-90723-4>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) has revolutionized numerous industries, including cyber security. As cyber-threats grow in sophistication, AI-driven security solutions are emerging as powerful tools to predict, detect, and mitigate cyber risks. This book, *Demystifying AI and ML for Cyber-Threat Intelligence*, aims to bridge the gap between theoretical AI concepts and their practical applications in cyber security.

This volume presents a comprehensive collection of cutting-edge research, addressing various aspects of AI-powered security, privacy protection, blockchain innovations, fraud prevention, cryptography, and secure communications. The book is structured into ten parts, each dedicated to a critical area where AI plays a transformative role in cyber security.

- *AI-Powered Cyber Security and Threat Mitigation* explores AI-driven intrusion detection, deep learning-based defense mechanisms and adaptive learning in cyber security.
- *Privacy, Data Protection and Secure AI Systems* highlights innovative techniques to enhance data integrity, privacy-preserving biometric authentication, and AI-driven data protection frameworks.
- *AI for Fraud Prevention and Threat Intelligence* delves into AI-powered fraud detection, phishing defense mechanisms and hybrid ensemble learning models.
- *Blockchain Innovations for Cyber Security and Digital Trust* presents blockchain-driven security frameworks, secure digital transactions, and applications of blockchain in academic and financial domains.
- *AI in Social Media and Misinformation Detection* discusses AI-based rumor detection, fake profile identification, and misinformation control in online ecosystems.
- *AI in IoT, Smart Cities and Autonomous Systems* examines AI's role in securing IoT networks, enhancing smart city infrastructure and cyber-physical security for autonomous systems.

- *AI and Cryptography for Secure Communications* focuses on AI-enhanced cryptographic methods, secure data encryption techniques, and attack modeling in modern communication networks.
- *AI in Healthcare and Data Privacy* highlights AI applications in healthcare security, privacy-preserving AI models, and generative AI for synthetic data protection.
- *AI for Secure Digital Infrastructure* covers AI-driven secure cloud computing, document privacy techniques, and adaptive steganography frameworks.
- *Ethical Considerations and Future Perspectives in AI Security* discuss the ethical implications of AI in cyber security, deep learning in financial security, and predictive analytics for cyber risk management.

By bringing together the expertise of researchers and industry professionals, this book provides an in-depth analysis of AI-driven cyber security methodologies and their impact on global digital security. It is designed for cyber security professionals, AI researchers, students, and policymakers seeking to understand the intersection of AI and cyber security.

As AI continues to evolve, it is imperative to stay ahead of emerging threats. We hope this book serves as a valuable resource, fostering further research and innovation in AI-driven cyber security solutions.

Kennesaw, USA
Vijayawada, India
Bhubaneswar, India
West Lafayette, USA

Ming Yang
Sachi Nandan Mohanty
Suneeta Satpathy
Shu Hu

Contents

AI-Powered Cyber Security and Threat Mitigation	
A Comprehensive Review on the Detection Capabilities of IDS Using Deep Learning Techniques	3
Harinath Ankarboina, Jasmini Kumari, and Amit Kumar Singh	
Next-Generation Intrusion Detection Framework with Active Learning-Driven Neural Networks for DDoS Defense	13
Saurav Raj, Shweta Sharma, Sweeti Sah, Kamaldeep, Manisha Malik, Rasmita Lenka, Sachi Nandan Mohanty, and V. Balaji	
Ensemble Learning Based Intrusion Detection System for RPL-Based IoT Networks	27
Archana Chougule, Rohit Mane, Krishnanjan Bhattacharjee, and Swati Mehta	
Advancing Detection of Man-in-the-Middle Attacks Through Possibilistic C-Means Clustering	39
Saswati Chatterjee, Lalmohan Pattnaik, Suneeta Satpathy, and Deepthi Godavarthi	
CNN-Based IDS for Internet of Vehicles Using Transfer Learning	55
Samarjeet Singh Rathore, Shaurya Yadav, Nitin Singh, and Biswajit Brahma	
Real-Time Network Intrusion Detection System Using Machine Learning	67
Sam Peter, J. L. Aravind, Feba Mariyam Jacob, Johann Varghese George, and Tessy Mathew	

**OpIDS-DL: Optimizing Intrusion Detection in IoT Networks:
A Deep Learning Approach with Regularization and Dropout
for Enhanced Cybersecurity** 89
A. Pappurajan, Vinothkumar Kolluru, Y. Sunil Raj, Sudeep Mungara,
Advitha Naidu Chintakunta, and Charan Sundar Telaganeni

Privacy, Data Protection, and Secure AI Systems

**ML-Powered Sensitive Data Loss Prevention Firewall
for Generative AI Applications** 105
Soumya R. Saju, C. S. Sajeesh, Gigi Joseph, and N. Jaisankar

**Enhancing Data Integrity: Unveiling the Potential of Reversible
Logic for Error Detection and Correction** 117
Premanand Kadbe, Shriram Markande, and Manisha Waje

Enhancing Cyber Security Through Reversible Logic 139
Premanand Kadbe, Shriram Markande, and Manisha Waje

**Beyond Passwords: Enhancing Security with Continuous
Behavioral Biometrics and Passive Authentication** 161
Pankaj Chandre, Suvarna Joshi, Rahul Rathod, Jyoti Nandimath,
Bhagyashree Shendkar, and Yuvraj Nikam

AI for Fraud Prevention and Threat Intelligence

**A Greedy Hybrid Ensemble Approach for Security Applications:
Fraud, Intrusion, and Malware Detection** 177
Monika Mangla, Nonita Sharma, Madhuchhanda Tripathy,
Vaishali Mehta, and Manik Rakhra

**Optimizing Ensemble Models for Security Applications:
A Comparative Study of Greedy and Dynamic Approaches** 189
Monika Mangla, Nonita Sharma, Saumyaranjan Acharya,
Vaishali Mehta, and Manik Rakhra

**Strategic Deployment of Machine Learning in Combating Email
Spam and Cyber Threats** 203
Sarita Mohanty and Anupa Sinha

AI-Powered Multi-layered Phishing Defense Framework (AIPDF) 215
Pallavi Bhujbal, Jayashree Pasalkar, Madhura Eknath Sanap,
Bhagyashree Shendkar, Rajkumar Patil, and Moushmee Kuri

Blockchain Innovations for Cyber Security and Digital Trust

Ethereum Blockchain-Based Decentralized Voting Platform 227
Natasha Wanjari, Pratiksha Chafle, and Rahul Moriwale

BLESS: Blockchain-Enhanced Intelligent Security System Using BPSO and AVOA for Smart Home Network	239
Amrutanshu Panigrahi, Nilachakra Dash, Abhilash Pati, Bibhuprasad Sahu, Bidya Bhusan Panda, and Ghanashyam Sahoo	
Navigating Security and Privacy in Blockchain: Challenges and Future Directions	253
Navjot Kaur and Ramandeep Kaur	
Revolutionizing Academic Record Management: A Blockchain-Driven Solution for Secure and Transparent Student Documents	269
Swapnil S. Chaudhari, Vaishnavi Vikhe, Dipanjali Bhujbal, Shreyash Pangarkar, and Gokul Arya	
A Block-Chain-Based Security Framework for Protecting Patient Electronic Health Records	287
Natasha Wanjari, Pratiksha Chafle, and Rahul Moriwai	
Blockchain-Powered Secure EHR Exchange in Mobile Cloud E-Health Systems	303
S. P. Santhoshkumar, V. R. Navinkumar, Rekhasree Manthu, S. Hariharasudhan, N. Ramajayam, and S. Gajalakshmi	
AI in Social Media and Misinformation Detection	
Rumor Veracity Detection in Social Networks: A Brief Survey	319
Shruti Bajpai and Shashank Kumar Singh	
Techniques for Detecting False Information on Social Media to Strengthen Cybersecurity	331
Prabhat Kumar Sahu, Smita Rath, Alakananda Tripathy, Rashmi Rani Patro, and Sangam Malla	
Machine Learning for Fake Profile Users Detection in Social Network Systems: A Review and Implementation Phase	345
Aishwarya Waghmare, Vinothkumar Kolluru, Yagnesh Challagundla, I. V. S. Aditya Bhrugumalla, Advaita Naidu Chintakunta, and Sagar Pande	
Leveraging Advanced Technologies to Enhance Public Awareness and Mitigate Risks of Cryptocurrency Scams: A Qualitative Analysis	359
Vinay Kumar Kasula and Abdullah Alshboul	

AI in IoT, Smart Cities, and Autonomous Systems

Intelligent Sensor Placement in WSN for Maximum Coverage Using Simulated Annealing 373
V. Yathavraj, M. Saravanakumar, S. Rajkumar, P. Priyadharshini, Mani Deepak Choudhry, M. Sundarajan, and J. Akshya

Cyber-Physical Intrusion Detection System for UAVs 387
Sushant Mane, Jai Bhortake, Vidhi Wankhade, and Faruk Kazi

AI-Powered Video Analytics: Enhancing Real-Time Threat Detection and Public Safety 401
Kushal Walia, Namita Dandawate, and Bhumik Thakkar

Enhancing Women’s Safety and Security Through IoT Systems 415
Jyoti Yogesh Deshmukh, Mayura Vishal Shelke, Anuja Jadhav, and Saleha Saudagar

Secure Home Automation Using AI & IoT 431
Uzair Ahmad Ansari, Rahul Narendra Chunarkar, Shrishail Mungse, Rahul Agrawal, Nekita Chavhan Morris, Chetan Dhule, and Girish Bhavekar

Intelligent Intrusion Detection: A Deep BiLSTM Approach Empowered by Hybrid Spider-Coyote Optimization for IIOT Security 447
Sushama L. Pawar and Mandar S. Karyakarte

AI and Cryptography for Secure Communications

Enhanced Elliptic Curve Cryptography with MHOTP Key Generation and Visual Cryptography Based Image Authentication 467
Sachin Madhukar Kolekar and Ram Kumar Solanki

Threat Analysis and Attack Modeling in Data-Centric Communication for Named Data Networking 483
Riddhi Mirajkar, Gitanjali Shinde, Parikshit Mahalle, and Nilesh Sable

Energy-Efficient Machine Learning-Based Data Encryption Techniques for Information Blocks: A Comprehensive Analysis 499
Mrunal S. Jagtap and D. Sangeetha

AI in Healthcare and Data Privacy

Synthetic Data Usage for Healthcare Privacy Using GENERATIVE AI 515
Pratyush Ranjan Sahu, Alakananda Tripathy, and Alok Ranjan Tripathy

The Future of Healthcare Chatbots: Balancing AI Innovation with Robust Cybersecurity Practices	527
Bhagyashree Shendkar, Pankaj Chandre, Ganesh Pathak, and Madhukar Nimbalkar	
AI for Secure Digital Infrastructure	
Optimized Outsourced Decryption for Attribute-Based Encryption: Cost-Efficiency for Users and Cloud Servers in Green Cloud Computing	545
Subhash G. Rathod, Meghana R. Yashwante, Sunita Nikam, Sushama Laxman Pawar, and Nilesh J. Uke	
Document Privacy Preservation Using Information Security Methods and Create Awareness About Privacy Policies	561
Vidhya Gavali, Aastha Shinde, Shweta Gumaste, Vaishnavi Akul, and Ameya Kunte	
An Enhanced Adaptive Steganography Framework Using Inverted LSB and Optimal Patterns (EASIOP)	569
Sheetal Agrawal and Kshiramani Naik	
Ethical Considerations and Future Perspectives in AI Security	
AI-Powered Grief Technology: The Ethical Implications of AI in Privacy	585
Sonal Mahapatra	
A Comparison of Interdependent Deep Learning Models and Exponential Smoothing Method for Predicting Bitcoin Price	601
Nrusingha Tripathy, Sarbeswara Hota, Debahuti Mishra, Meera Dash, Soumyarashmi Panigrahi, and Subrat Kumar Nayak	
Recognition and Classification of ARP Spoofing and DDoS Attack Using Machine Learning Approach	615
Saswati Chatterjee, Suneeta Satpathy, and Deepthi Godavarthi	

AI-Powered Cyber Security and Threat Mitigation

A Comprehensive Review on the Detection Capabilities of IDS Using Deep Learning Techniques



Harinath Ankarboina, Jasmini Kumari, and Amit Kumar Singh

Abstract Intrusion Detection Systems (IDS) are essential for enhancing cybersecurity in modern vehicular networks, especially as they become increasingly interconnected. However, traditional IDS approaches often face limitations in handling complex attack patterns, evolving threat landscapes, and the high volume of network data generated. This paper examines the incorporation of DL methodologies, including LSTM networks, CNNs, autoencoders, and DRL, within IDS frameworks. These methods offer enhanced detection accuracy, real-time anomaly identification, and adaptability, addressing key challenges in IDS deployment for vehicular networks. This review study highlights the improvements in IDS effectiveness and the future directions for DL-driven cybersecurity in connected and autonomous vehicles.

Keywords IDS · CNN · LSTM · DRL · CAN · Autoencoders · IoV · VANETs

1 Introduction

IDS is essential for augmenting the security of vehicular networks, especially given the growing interconnectivity of contemporary vehicles. Intrusion Detection Systems (IDS) have a central role to play in vehicular network security improvement, especially against the backdrop of rising interconnectivity of modern cars. IDS systems they are specifically geared to observe network traffic for anomalies and prospective intrusions, maintaining communication integrity between the vehicle's Electronic Control Units (ECUs) [1]. The introduction of Vehicular Ad Hoc Networks

H. Ankarboina · J. Kumari · A. K. Singh (✉)

Department of Computer Science and Engineering, SRM University AP, Amaravati, India

e-mail: amitkumar.s@srmap.edu.in

H. Ankarboina

e-mail: harinath_ankarboina@srmap.edu.in

J. Kumari

e-mail: jasmini_kumari@srmap.edu.in

(VANETs) further called for the application of efficient IDS since VANETs support vehicular-to-vehicle communication and vehicle-to-infrastructure communication, hence enhancing safety and service performance but subjecting them to all forms of cyber threats [2]. Recent breakthroughs in deep learning (DL) techniques have further boosted the performance and efficiency of IDS. By applying convolutional neural networks (CNN) and other DL models, researchers have made systems that can detect anomalies in real-time, which is essential for timely response to prospective intrusions [3]. Anomaly detection methods are most appropriate for identifying abnormal patterns that differ from regular vehicular operational patterns, indicating prospective vulnerability to security attacks [4]. More importantly, implementing specific intrusion detection algorithms tailored explicitly for vehicular networks is inevitable. The algorithms must deal with the challenges of limited computation capabilities and the need for timely responsiveness and standard features in the automobile environment [5]. Also included in the threat landscape are various types of cyber attacks, including Denial of Service (DoS) attacks that hinder normal network operations by overwhelming the system with excessive traffic [6]. A good IDS should be capable of identifying such attacks to assure service availability in vehicular networks. Vehicle network cybersecurity platforms are increasingly incorporating machine learning-based methods. These add another layer of protection against cyberattacks [7]. This integration is crucial in developing autonomous cars, which need strong cybersecurity controls to communicate and operate safely [8]. With the automotive sector moving towards IP-based protocols such as SOME/IP, which are intended for service-oriented communication in automotive Ethernet networks, the demand for efficient anomaly detection systems is even higher [9]. In summary, manufacturing advanced IDS for vehicular networks is fundamental for defense against a broad range of cyber attacks. With DL, anomaly detection, and customized algorithms, these systems can offer real-time monitoring and response capabilities, thereby improving the overall security of connected vehicles.

In this review paper, we will concentrate on using DL approaches to improve the detection features of IDS. DL has a significant contribution to the detection features of IDS in intra-vehicular networks by utilizing complex algorithms to study complex data patterns and yield precise detection of anomalies and attacks. These systems use deep learning architectures, including LSTM networks, CNN, and autoencoders, to study and process the enormous volumes of data in vehicular networks. Integrating DL into IDS frameworks allows for more precise detection of malicious activities, thereby bolstering the security of intra-vehicular networks.

The subsequent structure of the article is outlined as follows. Section 2 discusses the pertinent literature. Motivation is addressed in Sect. 3, DL techniques utilizing IDS are discussed in Sect. 4, and the conclusion is presented in Sect. 5.

2 Related Work

The increasing complexity and volume of network traffic in vehicular networks necessitate sophisticated intrusion detection capabilities. Conventional IDS, including rule-based and anomaly-based approaches, are constrained in identifying advanced and evolving threats because they depend on predetermined signatures and statistical frameworks. As vehicular networks expand, researchers have increasingly explored DL approaches to enhance IDS effectiveness, leveraging the strengths of various architectures to improve detection accuracy, scalability, and adaptability. The paper [10] presented an efficient IDS for network traffic patterns using the NSLKDD dataset. The technique, combining the Genetic Optimization Algorithm (GOA) and Naive Bayesian technique, achieved a detection accuracy of 95.0%, outperforming the recommended 53.0%. However, the study highlights potential attack surfaces and suggests further research to explore noise levels and patterns in machine-learning models. The paper [11] presented a hybrid classifier approach for intrusion detection in general network security, using the Beetle Swarm Optimization and K-RMS clustering algorithm. It achieves higher accuracy, precision, specificity, and recall rates than existing models. However, limitations include reliance on the CICIDS2017 dataset and the need for computational resources. The paper [12] presented a hybrid model for intrusion detection using ML and DL techniques, specifically CNN and LSTM. The model demonstrates high detection rates, good accuracy, and low false acceptance rates, addressing security limitations in network systems. The model is crucial due to increasing data transfer and attacker efforts. The paper [13] proposed a multistage framework using DL to enhance IDS in network traffic. The framework uses three sequential DNN architectures, including one for classifiers and two for autoencoders. The transfer learning technique enhances robustness against evolving cyber threats, achieving an average detection accuracy of 98.5%.

3 Motivation

IDS are critical for safeguarding networks against evolving cyber threats. Conventional IDS methodologies, typically dependent on rule-based or anomaly-based detection, find it challenging to adjust to the swift evolution of attack strategies and the growing intricacy of network data. The ever-increasing complexity of cyber attacks highlights the potential of utilizing DL techniques for IDS, making it a compelling area for investigation. DL models, with their ability to learn complex patterns and features in unsupervised datasets, offer significant benefits over traditional methods concerning accuracy, flexibility, and scalability. This review analyzes and assesses recent developments in combining DL methods with IDS. Based on recent studies and real-world applications, we try to emphasize not only the accuracy in detection offered by DL but also the capability of DL to overcome the common pitfalls of IDS.

4 Various Deep-Learning Techniques Implemented in IDS

DL dramatically improves the detection capacity of IDS in intra-vehicular networks. Through the use of sophisticated neural network architectures to interpret sophisticated data patterns. This will enhance accuracy in the detection of anomalies and attacks.

4.1 *RNN and LSTM-Based IDS*

LSTM networks have shown improved performance in false message attack detection in VANETs over conventional approaches. LSTMs are effective since they can learn temporal patterns and time series dependencies, essential for anomaly detection in the dynamic VANET environment. This enables LSTMs to outperform conventional machine learning approaches in accuracy and reliability in IDS. The following discussion explains how LSTM networks improve false message detection in VANETs. Wang et al., in the paper [14], introduced a new IDS based on time series classification and DL to enhance false emergency message detection in VANETs. The system uses LSTM to detect patterns in traffic parameters related to time to strengthen the detection of false messages from internal and collusive attackers. Extensive simulations confirm the effectiveness of the approach in real-world deployment. Wang et al., in the paper [15], introduced a new IDS based on time series classification and deep learning to enhance false emergency message detection in VANETs. The system uses LSTM networks to detect patterns in traffic parameters related to time to strengthen the detection of false messages from internal and collusion attackers. The LSTM-based IDS is more precise in detecting false messages than existing machine-learning methods. This paper introduces a prediction-driven IDS approach for detecting anomalies and attacks in a Controller Area Network (CAN) bus by examining the temporal relationships of message content. The proposed IDS framework outperforms top classifiers, showing near 100% detection accuracy and F-scores. In the paper [16], the CNN-LSTM with Attention model (CLAM) presented a novel intrusion detection method specifically for automotive networks with a special focus on the CAN protocol, which is prone to attack because it does not have in-built security features. The model uses one-dimensional convolution for signal value feature extraction and bidirectional LSTM to effectively detect temporal relationships in the data. The approach also uses attention mechanisms to identify significant time steps, improving the rate of convergence and the accuracy of the predictive results. The CLAM model attains an average F1 score of 0.951 and an error rate of 2.16%, with a 2.5% improvement in accuracy in attack detection compared to previous work. This model improves convergence speed and prediction accuracy, eliminating the necessity to parse the CAN communication matrix. The paper [17] introduced a DL-centered bidirectional LSTM architecture for intrusion detection in Autonomous Vehicles. The framework uses temporal patterns in communication

data to identify intrusions in real-time, reducing false alarms. It achieves a high accuracy rate, and the framework detects zero-day attacks in IoV networks. The paper [18] presented a DL-based IDS designed for ITS, focusing on detecting suspicious network activity in In-Vehicle Networks, Vehicle-to-Vehicle communications, and Vehicle-to-Infrastructure networks. The system outperforms eight other intrusion detection techniques.

4.2 Convolutional Neural Network (CNN) Based IDS

CNN-based IDS have significant potential in enhancing the security and reliability of vehicular communication networks. These systems leverage the advanced pattern recognition capabilities of CNNs to detect and mitigate various cyber threats in vehicular networks, which are increasingly becoming targets due to their interconnected nature. The paper [19] analyzed secrecy performance in mobile vehicular networks, focusing on using a dense-inception convolution neural network (DI-CNN) for predicting secrecy performance. The DI-CNN model demonstrates a 48.8% better prediction accuracy than the Transformer method, enhancing real-time prediction capabilities for secrecy performance in IoV communication systems. The paper emphasizes the importance of physical layer security modeling for secure data transmission. Wang et al. [20] proposed CNN-based IDS can improve security in vehicular communication networks by accurately detecting and locating malicious data frames from external nodes and compromised electronic control units. This system efficiently minimizes bandwidth usage, thus making it suitable in resource-limited environments, especially in intelligent connected vehicles (ICVs). The FeatureBagging-CNN combined model efficiently detects and locates malicious data frames without requiring developer documentation, thus making it more feasible to protect in-vehicle networks. The FeatureBagging-CNN combined model provides detection capabilities without using CAN bus bandwidth, thus making it suitable in resource-limited environments. The paper [21] proposed a flow-based intrusion detection system for Vehicular Ad Hoc Networks (VANETs) using CNN and Context-Aware Feature Extraction-Based CNN (CAFECNN). This system improves security by detecting and countering potential threats in real time, thus improving the overall security architecture. The study emphasizes the need for context-aware mechanisms in the detection process. The author of the paper [22] proposed an Intelligent IDS (IIDS), which utilizes a modified CNN to improve intrusion detection in Connected and Autonomous Vehicles (CAVs). It utilizes hyperparameter optimization to detect and classify malicious autonomous vehicles (AVs), thus preventing accidents and pandemonium. The IIDS is implemented in a 5G Vehicle-to-Everything (V2X) environment, achieving a high 98% accuracy rate in cyberattack detection.

4.3 *Auto Encoder-Based IDS*

As a neural network, autoencoders possess tremendous potential to improve vehicular communication network security and reliability. The networks, being an integral part of the IoV is experiencing issues in data security, network congestion, and real-time data processing. Autoencoders can solve these problems with better data representation, anomaly detection, and effective data transmission. The paper [23] introduced a semi-supervised learning-based convolutional adversarial autoencoder model for in-vehicle intrusion detection. The approach improves security by identifying anomalies in communication patterns, enhancing data compression, and enabling effective feature extraction for threat detection. The model is better for intrusion detection than traditional approaches, decreasing false positives and improving detection rates, which is promising for real-world applications in vehicle security. The paper [24] presented a VeNet hybrid learning system, which employs a stacked autoencoder to predict future vehicle locations and network traffic. The system improves data transmission efficiency and decreases congestion, leading to more secure and reliable communication. The VeNet model decreases required signaling network traffic and prediction error by 75%, decreasing vehicle energy consumption and learning delays. The paper [25] proposed a deep Q-learning-based strategy to secure cellular V2X communications using Autoencoders. The plan aims to maximize secrecy rates while managing interference levels. It addresses the vulnerability of V2X links to eavesdropping attacks and ensures the required Signal-to-Interference-plus-Noise Ratio (SINR) for both V2X and I2V communications. The simulation results of the study prove its efficiency. The method solves data scarcity issues and enhances the model's generalization capability in real-world settings, improving security in vehicular communication systems. The paper [26] introduced a Secure and Intelligent System for the Internet of Vehicles (SISIV), which employs DL structures, graph convolutional networks, and attention mechanisms to enhance traffic forecasting and secure data transmission via blockchain technology. The system surpasses the current forecasting rate, F-measure, and attack detection solutions and is an efficient and trustworthy solution for traffic flow prediction in the IoV.

4.4 *Deep Reinforcement Learning (DRL) Based IDS*

Deep Reinforcement Learning (DRL) offers great promise to improve the reliability and safety of vehicular communication networks. When employing DRL, the networks can address problems like ultra-reliable low-latency communication (URLLC), resource allocation, and trust management, which are essential to intelligent transportation systems development. The article [27] proposed a novel DRL approach for joint resource allocation for ultra-reliable, low-latency vehicle-to-everything (V2X) communications. The approach optimally addresses the limitations

of traditional optimization-based algorithms when deployed in dynamic environments. It introduces an efficient event-triggered DRL algorithm, reducing execution frequency by up to 24% while achieving 95% of traditional performance. This approach improves overall performance and security in V2X communication systems. The paper [28] introduced a DRL-assisted hybrid precoding method for vehicle-to-infrastructure communication in the 5G new radio frequency range 2. The method balances complexity, reliability, and data rate while considering Doppler shift and delay spread. The study uses a downlink transmission model and demonstrates that incorporating RNN significantly enhances training efficiency, with the twin delayed deep deterministic policy gradient model showing superior spectral performance. The paper [29] introduced a distributed trust-sharing mechanism utilizing Reservoir Computing within the context of the IoT. This mechanism integrates the Echo State Network with Reinforcement Learning to enhance vehicle trust and communication. The model is structured as a Partially Observable Markov Decision Process, considering vehicle storage and computational capacity constraints. Simulation results show the algorithm outperforms traditional methods, demonstrating its feasibility and effectiveness in vehicle trust sharing. The paper [30] discussed the use of DRL in optimizing resource allocation strategies in vehicular edge computing systems. It introduces a duopolistic edge service market model for vehicles, where edge servers announce pricing strategies and vehicles generate reviews. The paper proposes a DRL framework to maximize vehicle utility, even when they prefer not to disclose their requests. The paper [31] proposed a double deep Q-network framework based on reinforcement learning for fiber-wireless vehicular communication networks. It presents a priority-driven V2V data offloading strategy, categorizing data packets according to urgency and significance. This method improves data transmission efficiency, decreases latency, and enhances network performance in dynamic settings, showcasing its potential for future vehicular communication systems.

5 Conclusion

In conclusion, DL techniques offer transformative potential in strengthening IDS for vehicular networks, addressing the limitations of traditional methods in a highly dynamic and vulnerable environment. By employing models like LSTM, CNN, autoencoders, and DRL, IDS can achieve higher accuracy in anomaly detection, reduce false positives, and improve adaptability to new attack types. These advancements are critical for the security of vehicular networks, where real-time detection and response are paramount. Future research should refine these DL models to be more resource-efficient, interpretable, and resilient to adversarial attacks, further enhancing their applicability in real-world vehicular environments. DL-based IDS encounters considerable obstacles concerning computational resources and model interpretability. The challenges arise from the intricate nature of DL models, which necessitate significant computational resources and frequently function as “black boxes,” complicating the comprehension of their decision-making

mechanisms. Resolving these issues is essential for the efficient implementation of IDS in cybersecurity.

References

1. Khan, J., Lim, D.-W., Kim, Y.-S.: Intrusion detection system can-bus in-vehicle networks based on the statistical characteristics of attacks. *Sensors* **23**(7), 3554 (2023)
2. Lavanya, R., Kannan, S.: Intrusion detection system for energy efficient cluster-based vehicular ad-hoc networks. *Intell. Autom. & Soft Comput.* **32**(1) (2022)
3. Alfardus, A., Rawat, D.B.: Machine learning-based anomaly detection for securing in-vehicle networks. *Electronics* **13**(10), 1962 (2024)
4. Neupane, S., Fernandez, I.A., Patterson, W., Mittal, S., Rahimi, S.: A temporal anomaly detection system for vehicles utilizing functional working groups and sensor channels. In: 2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC), pp. 99–108. IEEE (2022)
5. Bi, Z., Xu, G., Xu, G., Tian, M., Jiang, R., Zhang, S.: Intrusion detection method for in-vehicle can bus based on message and time transfer matrix. *Secur. Commun. Netw.* **2022**(1), 2554280 (2022)
6. Attar, A.E., Wehby, A., Chbib, F., Mehrez, H.A., Fadlallah, A., Hachem, J., Khatoun, R.: Analysis of machine learning algorithms for DDoS attack detection in connected cars environment. In: 2023 Eighth International Conference on Mobile And Secure Services (MobiSecServ), pp. 1–7. IEEE (2023)
7. Shihab, M.A.: Intrusion detection using machine learning-hardened domain generation algorithms. *Period. Eng. Nat. Sci.* **8**(4), 2539–2546 (2020)
8. Onur, F., Gönen, S., Barışkan, M.A., Kubat, C., Tunay, M., Yılmaz, E.N.: Machine learning-based identification of cybersecurity threats affecting autonomous vehicle systems. *Comput. & Ind. Eng.* **190**, 110088 (2024)
9. Casparsen, A., Sørensen, D.G., Andersen, J.N., Christensen, J.I., Antoniou, P., Krøyer, R., Madsen, T., Gjoerup, K.: Closing the security gaps in some/ip through the implementation of a host-based intrusion detection system. In: 2022 25th International Symposium on Wireless Personal Multimedia Communications (WPMC), pp. 436–441. IEEE (2022)
10. Umukoro, I., Eke, B., Edward, O.: An efficient intrusion detection technique for traffic pattern learning. *Sci. Afr.* **23**(2), 26–41 (2024)
11. Pran, S.G., Raja, S., Jeyasudha, S.: Intrusion detection system based on the beetle swarm optimization and K-RMS clustering algorithm. *Int. J. Adapt. Control Signal Process.* **38**(5), 1675–1689 (2024)
12. Sajid, M., Malik, K.R., Almogren, A., Malik, T.S., Khan, A.H., Tanveer, J., Rehman, A.U.: Enhancing intrusion detection: a hybrid machine and deep learning approach. *J. Cloud Comput.* **13**(1), 123 (2024)
13. Hore, S., Ghadermazi, J., Shah, A., Bastian, N.D.: A sequential deep learning framework for a robust and resilient network intrusion detection system. *Comput. & Secur.* 103928 (2024)
14. Yu, Y., Zeng, X., Xue, X., Ma, J.: LSTM-based intrusion detection system for VENTs: a time series classification approach to false message detection. *IEEE Trans. Intell. Transp. Syst.* **23**(12), 23906–23918 (2022)
15. Mansourian, P., Zhang, N., Jaekel, A., Kneppers, M.: Deep learning-based anomaly detection for connected autonomous vehicles using spatiotemporal information. *IEEE Trans. Intell. Transp. Syst.* **24**(12), 16006–16017 (2023)
16. Sun, H., Chen, M., Weng, J., Liu, Z., Geng, G.: Anomaly detection for an in-vehicle network using CNN-LSTM with an attention mechanism. *IEEE Trans. Veh. Technol.* **70**(10), 10880–10893 (2021)

17. Khan, I.A., Moustafa, N., Pi, D., Haider, W., Li, B., Jolfaei, A.: An enhanced multi-stage deep learning framework for detecting malicious activities from autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.* **23**(12), 25469–25478 (2021)
18. Ashraf, J., Bakhshi, A.D., Moustafa, N., Khurshid, H., Javed, A., Beheshti, A.: Novel deep learning-enabled lstm autoencoder architecture for discovering anomalous events from intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.* **22**(7), 4507–4518 (2020)
19. Xu, L., Tang, H., Li, H., Li, X., Gulliver, T.A., Le, K.N.: Secrecy performance intelligent prediction for mobile vehicular networks: an DI-CNN approach. *IEEE Trans. Intell. Transp. Syst.* (2024)
20. Xun, Y., Deng, Z., Liu, J., Zhao, Y.: Side channel analysis: a novel intrusion detection system based on vehicle voltage signals. *IEEE Trans. Veh. Technol.* **72**(6), 7240–7250 (2023)
21. Shams, E.A., Rizaner, A., Ulusoy, A.H.: Flow-based intrusion detection system in Vehicular Ad hoc Network using context-aware feature extraction. *Veh. Commun.* **41**, 100585 (2023)
22. Anbalagan, S., Raja, G., Gurumoorthy, S., Suresh, R.D., Dev, K.: IIDS: intelligent intrusion detection system for sustainable development in autonomous vehicles. *IEEE Trans. Intell. Transp. Syst.* **24**(12), 15866–15875 (2023)
23. Hoang, T.-N., Kim, D.: Detecting in-vehicle intrusion via semi-supervised learning-based convolutional adversarial autoencoders. *Veh. Commun.* **38**, 100520 (2022)
24. Balasubramanian, V., Otoum, S., Reisslein, M.: VeNet: hybrid stacked autoencoder learning for cooperative edge intelligence in IoV. *IEEE Trans. Intell. Transp. Syst.* **23**(9), 16643–16653 (2022)
25. Jameel, F., Javed, M.A., Zeadally, S., Jäntti, R.: Secure transmission in cellular V2X communications using deep Q-learning. *IEEE Trans. Intell. Transp. Syst.* **23**(10), 17167–17176 (2022)
26. Djenouri, Y., Belhadi, A., Djenouri, D., Srivastava, G., Lin, J.C.-W.: A secure, intelligent system for internet of vehicles: a case study on traffic forecasting. *IEEE Trans. Intell. Transp. Syst.* **24**(11), 13218–13227 (2023)
27. Khan, N., Coleri, S.: Event-triggered reinforcement learning based joint resource allocation for ultra-reliable low-latency V2X communications. *IEEE Trans. Veh. Technol.* (2024)
28. Ye, J., Jiang, Y., Ge, X.: Deep reinforcement learning assisted hybrid precoding for V2I communications with Doppler shift and delay spread. *IEEE Trans. Veh. Technol.* (2024)
29. Jing, T., Liu, Y., Wang, X., Gao, Q.: Deep echo state Q-network aided trust sharing provisioning for internet of vehicle. *IEEE Trans. Veh. Technol.* (2023)
30. Zhang, H., Liang, H., Hong, X., Yao, Y., Lin, B., Zhao, D.: DRL-based resource allocation game with the influence of review information for vehicular edge computing systems. *IEEE Trans. Veh. Technol.* (2024)
31. Gupta, A., Jaiswal, S., Bohara, V.A., Srivastava, A.: Priority-based V2V data offloading scheme for FiWi based vehicular network using reinforcement learning. *Veh. Commun.* **42**, 100629 (2023)

Next-Generation Intrusion Detection Framework with Active Learning-Driven Neural Networks for DDoS Defense



Saurav Raj, Shweta Sharma, Sweeti Sah, Kamaldeep, Manisha Malik, Rasmita Lenka, Sachi Nandan Mohanty, and V. Balaji

Abstract Next-generation intrusion Detection performs an essential task in modern cyberspace to differentiate between normal and abnormal network traffic in incoming and outgoing network packets. This is one of industrial control systems' most important security solutions to detect potential attacks like ransomware, DDoS, etc. Financial institutions are persistent targets of DDoS attacks that disrupt the services, and IDS can detect these attacks by monitoring abnormal traffic. Therefore, this research

S. Raj

Department of Chemical Engineering, Birsa Institute of Technology Sindri, Sindri, India
e-mail: rsaurav087@gmail.com

S. Sharma · S. Sah

Department of Computer Engineering, National Institute of Technology Kurukshetra, Kurukshetra, India
e-mail: shweta.sharma@nitkkr.ac.in

S. Sah

e-mail: sweetisah3@nitkkr.ac.in

Kamaldeep

Department of Media Engineering, National Institute of Technical Teachers Training and Research, Chandigarh, India
e-mail: kamal.katyal@yahoo.com

M. Malik

Department of Computer Science and Engineering, Thapar Institute of Engineering and Technology, Patiala, India
e-mail: manishamalik53@gmail.com

R. Lenka (✉)

School of Electronics Engineering, KIIT Deemed to Be University, Bhubaneswar, India
e-mail: rasmitafet@kiit.ac.in

S. N. Mohanty

School of Computer Science and Engineering, VIT-AP University, Amaravati, India
e-mail: sachinandan.m@vitap.ac.in

V. Balaji

Department of CSE (AI and ML), Vardhaman College of Engineering, Hyderabad, India
e-mail: vuppalabalaji802@gmail.com

focuses on enhancing IDS performance by combining Active Learning with Artificial Neural Networks (ANN). We have implemented and compared two models, ANN with Active Learning and ANN without Active Learning. The experimental results show that Active learning with ANN consumes fewer resources and performs better than ANN without Active learning with an accuracy of 99%.

Keywords Next-generation intrusion detection system · DDoS · Active learning · Artificial Neural Networks

1 Introduction

Cyber-security is one of the top 10 global dangers for the present and the future. Distributed Denial of Service (DDoS) attacks that repress systems with the flood of network traffic are among the most common and devastating cyber-attack types [1]. A DDoS is an attack that floods the victim server/system with malformed network traffic, potentially overloading and slowing down the system, causing financial and reputational loss [2]. There are different ways to flood a system: ICMP echo request flood, SYN flood, HTTP-GET request flood, etc. An attacker can choose traffic according to the victim's network traffic; if the victim uses a web server, the adversary can flood the victim with HTTP request traffic.

Intrusion Detection System (IDS) can help reduce these attacks by detecting abnormal patterns and features [3]. Moreover, other types of attacks, such as phishing that spoof individuals into disclosing sensitive information, can cause significant financial loss and reputational damage. Some methods like spear and email phishing complicate the conventional detection methods, so IDS is very vital in detecting and reducing attempts by continuously monitoring and analyzing emails and network traffic [4–6]. Imposing powerful IDS is essential for ensuring the systems work smoothly, even under a cyber-attack, by maintaining service availability [7]. Nowadays, machine learning techniques are implemented for building IDS models from network audit data [8]. Moreover, deep learning is employed in intrusion detection tasks and is a vital area of research in cyber-security.

As the cyber threat increases, robust IDS are becoming a necessity. Therefore, signature-based, behavior-based, and machine learning-based techniques have been used for intrusion detection in next-generation IDS. The signature-based method works by matching the data points with their signature. These methods accurately detect known attacks but fail at detecting zero-day attacks. The anomaly detection method learns normal behavior in the data and flags deviations from this behavior as attacks. This method is successful in detecting unseen attacks [9]. Researchers have discovered many machine learning methods, such as support vector machines (SVM) and decision trees, to analyze network traffic data to enhance detection rates and reduce false predictions [10]. Besides this, other ensemble learning techniques, such as Random Forest (RF) and gradient boosting, have shown promising performance in IDS by integrating multiple classifiers [11].

However, these techniques are supervised, and the output depends entirely on the labeled data. On the other hand, deep learning also shows better results than machine learning if the dataset is large. Therefore, in this research work, we apply active learning. This semi-supervised training method helps reduce dependency on the labeled data by choosing very informative or uncertain event samples for training or reducing the need to train a model with a large dataset. We integrate active learning with deep learning, namely, artificial neural networks (ANN), in IDS to detect DDoS attacks.

Active learning is a technique that works by asking the user to choose the most valuable data points from a big training set in machine learning. This helps cut down on how many resources the model uses. The goal is to use training data smartly by picking out the most informative bits. The ANN model acts as a base learner, which helps active learners pick up complex patterns from the training data. The ANN acts as a base learner in intrusion detection to help with the generalization since it can fit the training data well.

The UNSW-NB15 dataset is used for intrusion classification tasks. Other Datasets such as KDD98, KDDCUP99, and NSLKDD don't capture modern network traffic like zero-footprint attacks. On the bright side, the UNSW-NB15 dataset gives current regular network traffic alongside newly created attack patterns [12]. Moreover, the existing literature shows better detection accuracy with the UNSW-NB15 dataset while evaluating the model compared to the rest [13].

1.1 Contributions

- **Integration of Active Learning with ANN:** Our research introduces novelty by integrating active learning with the ANN model. The active learning integrated with ANN detects DDoS attacks with an accuracy of 99.00% compared to ANN without active learning at 98.96%.
- **Semi-supervised Learning:** The existing research significantly depends on the labeled data and selects the most uncertain or informative data sample for labeling. Therefore, we apply active learning, a semi-supervised learning approach, which helps reduce dependency on the labeled data by choosing very uncertain data samples for training. Moreover, it takes fewer resources by reducing the proportion of labeled data needed for learning compared to traditional training methods.
- **Balancing Technique:** The UNSW-NB15 dataset has an imbalanced class of benign and malicious samples, leading to biased results. Therefore, we apply the SMOTE to prevent data imbalance problems.

2 Related Work

The literature review of Active learning in IDS shows very expressive advancements through various techniques, which are discussed as follows.

Almgren and Jonsson [14] used the KDD CUP99 dataset and integrated active learning in IDS using the SVM as a base learning model to detect Denial of Service (DoS) attacks. They discovered that the active learning integrated algorithm performed better than supervised learning by using almost 80% less labeled data and showed a slightly greater accuracy of 96.71% from the base model. Seliya and Khoshgoftaar [8] utilized neural networks combined with Active learning for Intrusion detection with the DARPA KDD-1999 dataset to detect DoS attacks. A comparison between an actively learned neural network and the C4.5 decision tree shows an active learning-based neural network outperforms a C4.5 decision tree with an accuracy of 95.94%, and the C4.5 decision tree shows an accuracy of 90%.

McElwee [15] used the KDD-CUP99 dataset and integrated active learning with k-means clustering and RF in IDS to detect DoS attacks. They found that active learning integrated with RF shows an accuracy of 90%, with only 0.13% of data points needing manual labeling by human experts. Kumari and Varma [16] used the NSL-KDD dataset and integrated active learning with SVM and Fuzzy C-Means (FCM) clustering to detect DoS attacks. They found that active learning with SVM and FCM shows an accuracy of 99.6% compared to SVM and FCM without active learning, which shows an accuracy of 99.40%.

Li and Gui [17] used the KDDCUP99 dataset and integrated active learning with the Transductive Confidence Machines for the K-Nearest Neighbor (TCM-KNN) algorithm to discover DOS attacks. It was found that active learning with TCM-KNN shows an accuracy of 99.7% with just selecting 40 instances, in comparison to TCM-KNN without active learning, which takes 2000 instances to reach a similar result as active learning. Zakariah & Abdulaziz [18] used the UNSW-NB15 dataset and integrated active learning with the RF algorithm to detect DoS attacks. It is found that active learning with RF gives an accuracy of 99.75% in comparison to the models, RF, SVM, and VLSTM without active learning using the UNSW-NB15 dataset, showing an accuracy from 90.50% to 98.67% using labeled data.

Aouedi [19] used the EDGE-IIOTSET dataset and integrated semi-supervised federated learning (FL) with active learning to detect DDoS attacks. It is found that federated learning with active learning performs better, with an accuracy of 94.67%, slightly greater than federated learning without active learning. It also takes less data to be labeled.

Table 1 summarizes the existing techniques for next-generation IDS using active learning. In contrast to the existing models, we have proposed a novel framework by integrating active learning with ANN using the UNSW-NB15 dataset for next-generation IDS. This dataset contains modern network traffic samples alongside newly created attack patterns.

Table 1 Summary and comparison of existing techniques with the proposed framework for next-generation IDS

Author	Dataset	Year	Active learning driven model	Attacks
Almgren and Jonsson [14]	KDD Cup 1999	2004	SVM	DoS
Seliya and Khoshgoftaar [8]	DARPA KDD 1999	2010	ANN	DoS
McElwee [15]	KDD Cup 1999	2017	RF	DoS
Kumari and Varma [16]	NSL-KDD	2017	SVM and FCM	DoS
Li and Gui [17]	KDD Cup 1999	2007	TCM-KNN	DoS
Zahariah and Abdulaziz [18]	UNSW-NB15	2023	RF	DoS
Aouedi [19]	EDGE-IIOTSET	2024	FL	DDoS
Proposed framework	UNSW-NB15	2024	ANN	DDoS

2.1 Research Gaps

In the existing literature, there are several gaps, particularly in Almgren and Jonsson [14], Seliya [8], and McElwee [15]. The authors used the KDD Cup 1999 and DARPA KDD 1999 datasets, which are old and lacked modern and unusual attack types. These are imbalanced data, i.e., it has more benign and fewer attack instances and cannot detect complicated cyber-attacks [8]. Moreover, the detection accuracy is low in Aouedi [19].

3 Proposed Framework for Detection of DDoS Attacks

The proposed framework is shown in Fig. 1, which has the following components.

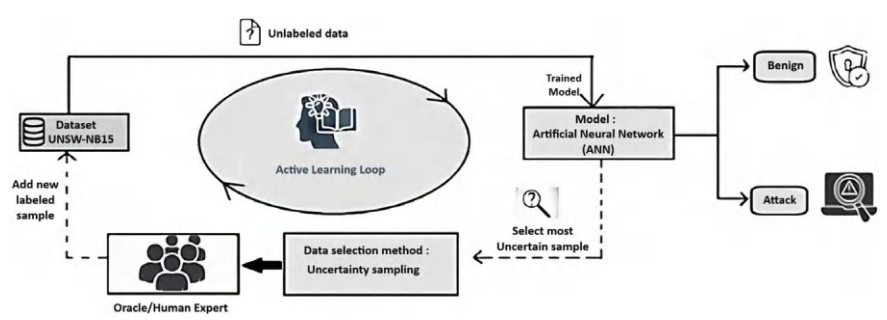


Fig. 1 Proposed framework with active learning for next-generation IDS

Table 2 Summary of DDoS attacks features of the UNSW-NB15 dataset

Features	Count
IPV4_SRC_ADDRESS	40
IPV4_DST_ADDRESS	40
L4_SRC_PORT	64586
L4_DST_PORT	64610
PROTOCOL	255
L7_PROTO	265
IN_BYTES	12256
OUT_BYTES	16047
IN_PKTS	892
OUT_PKTS	1207
TCP_FLAGS	15
FLOW_DURATION_MILLISECONDS	17377

3.1 Dataset Description

The UNSW-NB15 dataset [12] has been used to detect DDoS attacks in the IDS. This dataset has around 1,623,118 rows and 14 columns filled with features and labels. The labels include 1,550,712 benign cases and 72,406 DDoS attack instances. Table 2 summarizes the features of the UNSW-NB15 dataset.

3.2 Data Balancing Technique

The dataset used in our study initially suffers from a significant imbalance. This imbalance is a generic issue in intrusion detection and fraud detection tasks. When a dataset is imbalanced, models tend to be biased toward the majority class, achieving high accuracy for that class while performing poorly on the minority class. To tackle this problem, we employed SMOTE, a popular technique for handling data imbalance. SMOTE works by identifying samples from the minority class, selecting one or more of their nearest neighbors, and creating synthetic samples by interpolating the features of these samples and their neighbors.

In our case, the original dataset contains 1,550,712 instances of benign behavior, the majority, and only 72,406 DDoS attacks, leading to bias in the intrusion detection model. We generated 1,478,306 synthetic data points by applying SMOTE, resulting in a balanced dataset with 1,550,712 instances each for both benign and attack behaviors. This balanced dataset enables the model to detect benign and attack instances accurately, improving overall performance and prediction accuracy.

3.3 Data Pre-processing

The dataset contains categorical features, such as `ipv4 src addr` and `ipv4 dst addr`, that must be converted into a numerical format before training the machine learning model. We achieve this through label encoding, a method that assigns a unique integer to each category. This transforms the categorical data into a numerical format suitable for the ANN model, which can only process numerical data.

Additionally, the numerical features in the dataset require normalization to ensure that each parameter contributes uniformly to the model's training process. Normalization scales these features within a specific limit, typically from 0 to 1 or -1 to 1, which improves the model's convergence and performance. We applied Min-Max scaling, a common normalization technique using Eq. (1), to rescale the data.

$$X(\text{scaled}) = \frac{X - (X) \min}{(X) \max - (X) \min} \quad (1)$$

3.4 Active Learning-Based ANN for Detection of DDoS Attacks in Next-Generation IDS

We developed an Active Learning-based ANN model for detecting DDoS attacks in a next-generation IDS. Although our dataset is fully labeled, we simulate an active learning scenario by retaining a small subset of data labeled (x_{train} , y_{train}) while treating the rest as unlabeled (x_{pool} , y_{pool}). After pre-processing the data, we constructed an ANN model integrated with active learning techniques as outlined in Algorithm 1.

The ANN architecture has an input layer with 12 neurons, followed by two hidden layers containing 50 neurons using the ReLU function. Later on, one neuron in the output layer uses a sigmoid function. Our model is optimized with the Adam optimizer and binary cross-entropy for the cost function.

We utilized the `modAL` framework for Python 3 for the active learning component, initializing the `ActiveLearner` with uncertainty sampling as the query strategy. This strategy enables the model to label the uncertain instances from the unlabeled dataset (x_{pool}). The ANN, the base classifier, is trained on the labeled data (x_{train} , y_{train}). During each iteration, `ActiveLearner` selects the most uncertain samples from the x_{pool} , which an oracle or human expert then labels. These newly labeled samples are added to the training set (x_{train} , y_{train}) and removed from the unlabeled dataset (x_{pool}). This iterative process is carried out for 30 iterations, with 100 instances selected in each iteration, allowing the model to achieve high performance with minimal labeled data.

Ultimately, the model efficiently classifies network instances as either benign or DDoS attack, demonstrating the effectiveness of active learning in optimizing performance with fewer labeled examples.

Algorithm 1 IDS with Active Learning-based ANN

- 1: **Input:** Labeled and unlabeled data from the UNSW-NB15 dataset
- 2: **Output:** Classify instances as Benign or Attack
- 3: **Data:** Labeled dataset ($train, train$), Unlabeled dataset ($pool, pool$)
- 4: **Data Preprocessing:** Applied encoding on categorical data, SMOTE to balance the dataset, and MinMaxScaler() to normalize the dataset.
- 5: Encode categorical data:
- 6: **Initialize** the ANN model
- 7: **Initialize** the Active Learning framework:
 - a. Use *uncertainty sampling* as the query strategy
 - b. Define the *ActiveLearner* with the ANN model as the base estimator
 - c. **Train** the *ActiveLearner* on ($train, train$)
 - d. **Query** the most uncertain samples from (x_{pool}) using the *ActiveLearner*
 - e. Add the new samples to (x_{train}, y_{train})
 - f. Remove the newly labeled samples from (x_{pool})
 - g. The selected samples are labeled from oracle/human(in this experiment it is y pool)
- 8: **Initialize** n iterations = 30
- 9: **for** idx in range(n iterations) **do**
- 10: query idx,query sample = learner.query(x_{pool} , n instances=100, verbose=1)
- 11: learner.teach(x_{pool}, y_{pool} , verbose = 1)
- 12: $x_{pool} = np.delete(x_{pool}, query\ idx, axis = 0)$
- 13: **end for**
- 14: **Evaluate** the model on the testing set to classify instances as Benign or DDoS Attacks.

4 Experimental Setup and Results

The experimental setup contains the following components.

4.1 Training and Testing

The data set is divided into 90% training and 10% testing data for baseline ANN (i.e., without Active learning). For active learning, we use 10% as training data, which is labeled, and 90% as testing data, which is unlabeled. After training active learning with fewer labeled data, the model is tested on unlabeled data where it employs uncertainty sampling as a query strategy to choose the uncertain sample from the unlabeled dataset iteratively till the stopping criteria are not met and update the newly labeled data to the training data and remove it from an unlabeled data set.

After completing the active learning loop, the remaining unlabeled datasets, x_{pool} and y_{pool} , are used to estimate the performance of an Active learning framework. For baseline ANN, test datasets are used to assess the framework's performance.

Table 3 Hyperparameters details

Hyperparameters	Detail
Initial labeled data	x_train, y_train
Unlabeled data pool	x_pool, y_pool
Models	Active learning with ANN, Baseline ANN
Input layer (no. of features)	12 neurons
Hidden layers	50 neurons each in 2 hidden layers
Output layer	1 neuron
Activation function in hidden layers	ReLU
Activation function in the output layer	Sigmoid
ANN without active learning	30 epochs, 100 batch size per epoch
ANN with active learning	30 iterations, 100 instances per iteration
Query strategy	Uncertainty Sampling
Oracle	Y_pool
Optimizer	ADAM
Loss function	Binary cross-entropy

Table 3 demonstrates the dataset attributes and hyperparameters employed in the active learning-based ANN model to identify DDoS attacks. The build of ANN consists of 2 hidden layers with 50 neurons in each hidden layer, and ReLU is used for the hidden layer and Sigmoid for the output layer. For the Active learning loop, 30 iterations were implemented with 100 instances in each iteration. For Active Learning, we chose modAL, a modular framework that works with Python3. To pick out the most uncertain data samples, we applied the uncertainty sampling query strategy that modAL offers [20].

Uncertainty sampling has three types: least confidence, max-margin, and max-entropy. For our experiment, we used the least confidence sampling method. The Active learning loop stops at 30 iterations with 100 instances (100 most uncertain or informative instances) in each iteration, i.e., after the 30th iteration for the most uncertain unlabeled data sample, the Active learning loop stops. For the ANN baseline model, 30 epochs with 100 batch sizes are used for each epoch.

4.2 Experimental Results

In this segment, we compare the practical results of proposed models, ANN with Active learning and ANN without Active learning. The performance of the models is estimated using the metrics mentioned in Eqs. (2), (3), (4), and (5), where TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

Table 4 Performance comparison of traditional ANN and Active Learning-Driven ANN

Model	Accuracy	Precision	Recall	F1-Score
ANN without active learning	0.9896	0.9898	0.9896	0.9896
ANN with active learning	0.9900	0.9902	0.9900	0.9900

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$F1-Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5)$$

Table 4 shows the results of ANN with and without active learning for detecting DDoS attacks. The ANN that uses Active learning does better than the one that doesn't. It hits an accuracy of 99%, with precision at 99.02%, recall at 99%, and a fantastic f1-score of 99%. Conversely, the ANN without Active learning gets an accuracy of 98.96%, precision of 98.98%, recall of 98.96%, & an f1-score of 98.96%. Active learning boosts our model's performance, reduces false predictions, and improves its ability to distinguish benign traffic from DDoS attack traffic.

5 Comparison with the Existing Literature

The proposed research on Active Learning-Driven Neural Networks for next-generation IDS has been compared with the existing literature. The existing methods typically rely on extensive labeled data and traditional machine learning algorithms, which can take enormous resources and are less adaptable to rapidly changing cyber attacks. Our research focuses directly on Active learning integrated with ANN for next-generation IDS. We have reached a promising result with an accuracy of 99% with SMOTE, and uncertainty sampling for active learning enabled ANN to detect DDoS attacks.

Figure 2 demonstrates that our proposed model outperforms other active learning-based approaches, including Active Learning-driven RF [15], Active Learning-driven SVM [14], and Active Learning-driven federated learning [19]. The superiority of our model can be attributed to the limitations of the datasets used in these existing studies, namely KDD Cup 1999 and DARPA KDD 1999 Chemical Engineering. These datasets are considered traditional and contain numerous duplicate samples, making it challenging for models to learn and detect novel or unusual attack patterns effectively.

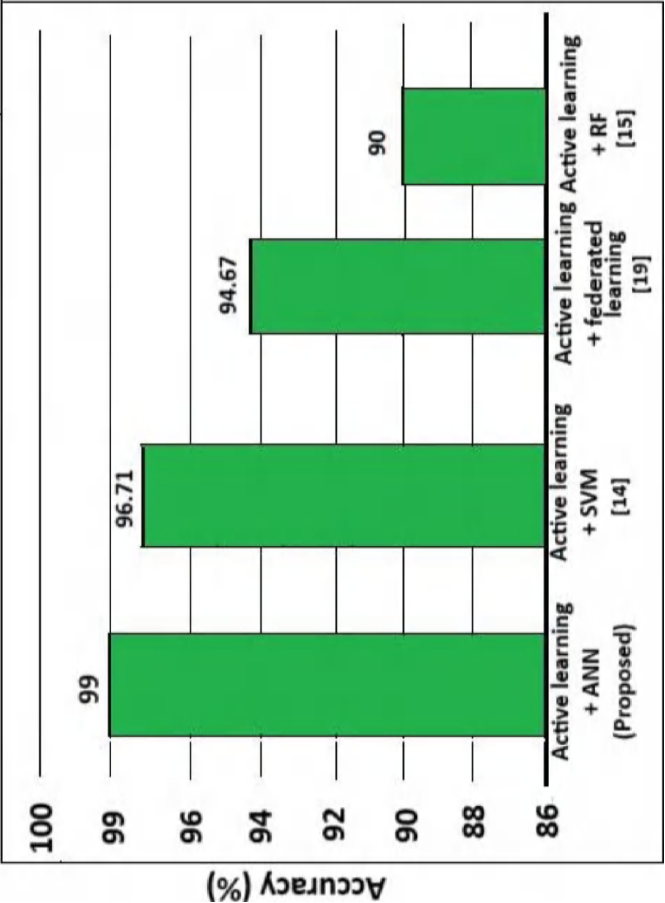


Fig. 2 Comparison of accuracy of proposed Active Learning-Driven Neural Networks for Next-Generation Intrusion Detection with the existing methods for IDS

Furthermore, they suffer from the significant class imbalance between benign and attack instances. They do not include modern cyber-attack types, reducing their relevance and effectiveness in contemporary DDoS attack scenarios. Additionally, it is evident from the literature that using imbalanced datasets for active learning can introduce bias into the results, further impacting the models' performance and generalizability.

6 Conclusion and Future Work

Our research demonstrates the effectiveness of integrating active learning with an ANN to enhance the performance of next-generation IDS in detecting DDoS attacks. By leveraging the UNSW-NB15 dataset, that captured recent network flow, our approach addresses several limitations of traditional intrusion detection data sets such as NSL-KDD, KDD CUP, and DARPA KDD, as they often contain outdated and redundant data samples. Additionally, the class imbalance issue is mitigated using the SMOTE technique. The practical outcomes show that the ANN framework combined with active learning achieves higher results, with higher accuracy, precision, recall, and F1-Score compared to the ANN model without Active Learning, reaching an accuracy of 99.00% while reducing the number of labeled data needed for training.

Given these promising results, future work could focus on further enhancing IDS model performance through various strategies. Potential directions include experimenting with alternative query strategies such as hierarchical sampling and query-by-committee and integrating other machine learning or deep learning models with active learning.

References

1. Mirkovic, J., Reiher, P.: A taxonomy of DDoS attack and DDoS defense mechanisms. *ACM SIGCOMM Comput. Commun. Rev.* **34**, 05 (2004)
2. Nazario, J.: DDoS attack evolution. *Netw. Secur.* **2008**(7), 7–10 (2008)
3. Bhuyan, M.H., Bhattacharyya, D.K., Kalita, J.K.: An empirical evaluation of information metrics for low-rate and high-rate DDoS attack detection. *Pattern Recognit. Lett.* **51**, 1–7 (2015)
4. Hong, J.: The state of phishing attacks. *Commun. ACM* **55**(1), 74–81 (2012)
5. Jagatic, T.N., Johnson, N.A., Jakobsson, M., Menczer, F.: Social phishing. *Commun. ACM* **50**(10), 94–100 (2007)
6. Mirkovic, J., Prier, G., Reiher, P.: Attacking DDoS at the source. In: 10th IEEE International Conference on Network Protocols, pp. 312–321. IEEE (2002)
7. Zargar, S.T., Joshi, J., Tipper, D.: A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks. *IEEE Commun. Surv. & Tutor.* **15**(4), 2046–2069 (2013)
8. Seliya, N., Khoshgoftaar, T.M.: Active learning with neural networks for intrusion detection. In: 2010 IEEE International Conference on Information Reuse & Integration, pp. 49–54. IEEE (2010)

9. Gamage, S., Samarabandu, J.: Deep learning methods in network intrusion detection: a survey and an objective comparison. *J. Netw. Comput. Appl.* **169**, 102767 (2020)
10. Wang, K., Stolfo, S. J.: Anomalous payload-based network intrusion detection. In: *International Workshop on Recent Advances in Intrusion Detection*, pp. 203–222. Springer (2004)
11. Folino, G., Pizzuti, C., Spezzano, G.: GP ensembles for large-scale data classification. *IEEE Trans. Evol. Comput.* **10**(5), 604–616 (2006)
12. Moustafa, N., Slay, J.: UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: *2015 Military Communications and Information Systems Conference (MilCIS)*, pp. 1–6. IEEE (2015)
13. Kumar, V., Das, A.K., Sinha, D.: Statistical analysis of the UNSW-NB15 dataset for intrusion detection. In: *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019*, pp. 279–294. Springer (2020)
14. Almgren, M., Jonsson, E.: Using active learning in intrusion detection. In: *Proceedings 17th IEEE Computer Security Foundations Workshop, 2004*, pp. 88–98. IEEE (2004)
15. McElwee, S.: Active learning intrusion detection using k-means clustering selection. In: *SoutheastCon 2017*, pp. 1–7. IEEE (2017)
16. Kumari, V.V., Varma, P.R.K.: A semi-supervised intrusion detection system using active learning SVM and fuzzy c-means clustering. In: *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics, and Cloud) (I-SMAC)*, pp. 481–485. IEEE (2017)
17. Li, Y., Guo, L.: An active learning-based TCM-KNN algorithm for supervised network intrusion detection. *Comput. & Secur.* **26**(7–8), 459–467 (2007)
18. Zakariah, M., Almazyad, A.S.: Anomaly detection for IoT systems using active learning. *Appl. Sci.* **13**(21) (2023)
19. Aouedi, O., Jajoo, G., Piamrat, K.: METALS: seMi-Supervised fEderaTed Active Learning for Intrusion Detection Systems (2024)
20. Danka, T.: Modal Python Documentation. <https://modal-python.readthedocs.io/en/latest/>. Last accessed 14 Dec. 2024

Ensemble Learning Based Intrusion Detection System for RPL-Based IoT Networks



Archana Chougule , Rohit Mane, Krishnanjan Bhattacharjee, and Swati Mehta

Abstract RPL (Routing Protocol for Low-Power and Lossy Networks) is a widely adopted routing protocol for 6LoWPAN-based IoT networks. Yet, it is vulnerable to several attacks compromising network security and reliability. This paper presents a machine learning-based approach for detecting four major RPL-specific attacks: Blackhole Attack, Flooding Attack, Decreased Rank Attack, and DODAG Version Number Attack. Using an existing IoT-RPL dataset generated from the Cooja simulator on the Mendeley platform, we implement ensemble learning techniques, including Random Forest, Gradient Boosting, AdaBoost, and Stacking Ensemble, to enhance attack detection accuracy. Feature selection techniques, such as Recursive Feature Elimination (RFE) and filter-based methods, are employed to identify key RPL-specific metrics, including packet delivery ratio, rank, and DODAG version number, which are critical for detecting attack patterns. The Stacking Ensemble model demonstrates the highest accuracy at 99.1%, outperforming other models in detecting the four types of attacks while maintaining a low false positive rate. To ensure the interpretability of the models, we apply SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations). SHAP values reveal the most influential features contributing to attack detection, such as packet delivery ratio and rank, while LIME provides local interpretability for individual predictions. These explainability methods confirm the reliability of the models, making them suitable for real-world IoT deployments. Future work will enhance model efficiency in real-time and under dynamic network conditions.

A. Chougule (✉) · R. Mane
Dr. J. J. Magdum College of Engineering, Jaysingpur, India
e-mail: chouguleab@gmail.com

R. Mane
e-mail: rohit.mane@jjmcoe.ac.in

K. Bhattacharjee · S. Mehta
Centre for Development of Advanced Computing, Pune, India
e-mail: krishnanjanb@cdac.in

S. Mehta
e-mail: swatim@cdac.in

Keywords RPL · IoT · Security · Ensemble Learning

1 Introduction

The Internet of Things (IoT) has enabled significant advancements in smart cities, healthcare, and industrial automation by connecting resource-constrained devices through low-power and lossy networks. These networks are often supported by 6LoWPAN (IPv6 over Low-Power Wireless Personal Area Networks), which facilitates IPv6 communication over constrained wireless links, and the Routing Protocol for Low-power and Lossy Networks (RPL), which provides efficient and scalable routing for these devices. However, despite the benefits of 6LoWPAN and RPL, such networks are vulnerable to various security threats that can severely disrupt communication, cause data loss, and lead to energy depletion.

RPL-based networks, in particular, are susceptible to several types of routing attacks, including Blackhole Attacks, Flooding Attacks, Decreased Rank Attacks, and DODAG Version Number Attacks. The Blackhole Attack occurs when a malicious node falsely advertises an optimal route to attract traffic, only to drop the packets, causing significant data loss. The Flooding Attack involves the malicious node generating excessive control messages, such as DIO (DODAG Information Object) or DAO (Destination Advertisement Object), which overwhelms the network with unnecessary traffic, leading to congestion and energy depletion. The Decreased Rank Attack is another serious threat where an attacker reduces its rank to deceive neighboring nodes into selecting it as the preferred parent, resulting in inefficient routing paths and communication delays. In the DODAG Version Number Attack, a malicious node artificially increases the version number of the DODAG (Destination Oriented Directed Acyclic Graph), forcing unnecessary re-routing and causing instability in the network.

Detecting and mitigating these attacks in RPL-based IoT networks is a complex challenge due to the resource-constrained nature of the devices involved. Traditional security mechanisms that require significant computational resources are not viable in this context. To address these limitations, machine learning (ML), mainly supervised learning techniques, has emerged as a powerful tool for attack detection. Supervised learning models can analyze labeled network traffic to detect anomalous patterns that indicate attacks. Furthermore, using feature selection techniques allows for identifying the most relevant features from the data, improving the model's efficiency while reducing computational overhead and making it more suitable for deployment in resource-constrained IoT environments.

In this paper, we leverage the IoT-RPL dataset obtained from the Cooja simulator to investigate the detection of Blackhole, Flooding, Decreased Rank, and DODAG Version Number attacks. The Cooja simulator is widely used to simulate 6LoWPAN and RPL-based networks, generating realistic network traffic data under every day and attack scenarios. The IoT-RPL dataset provides a comprehensive set of labeled traffic patterns, which allows us to train and evaluate various supervised machine

learning models. Feature selection techniques are employed to reduce the dimensionality of the dataset, ensuring that the resulting detection models are accurate, lightweight, and suitable for IoT devices with limited resources.

The remainder of this paper is organized as follows: Sect. 2 reviews related work on RPL security and machine learning-based attack detection. Section 3 introduces the IoT-RPL dataset generated using the Cooja simulator, describes the feature selection process, and outlines the supervised learning models used in the proposed framework. Section 4 presents the experimental results, evaluating the performance of the detection framework based on the IoT-RPL dataset. Finally, Sect. 5 concludes the paper and suggests directions for future research.

2 Related Work

This section briefs about work done by various researchers on RPL-based network intrusion detection systems. Jayaprakash and Lalitha [1] proposed a novel Network Intrusion Detection System (NIDS) for RPL-based IoT networks to combat routing attacks. It utilizes a bio-inspired voting ensemble classifier and feature selection technique (SA-ISSA). The method achieves 96.4% accuracy, 97.7% attack detection rate, and 3.6% false alarm rate using the RPL-NIDDS17 dataset. The paper combines multiple classifiers (SVM, KNN, LR, DT, Bi-LSTM) and employs SMOTE for dataset balancing. The proposed system effectively detects routing attacks in resource-constrained IoT environments [2].

Touzen et al. [3] proposed a hybrid deep learning-based intrusion detection system for RPL-based IoT networks, combining supervised and semi-supervised learning. Their system achieved an accuracy of 98% for known attacks like DIS, Rank, and Wormhole. The IoTR-DS dataset was used, and they reported high effectiveness even for untrained attacks with an average accuracy of 95%. Osman et al. [4] developed an artificial neural network for detecting decreased rank attacks in RPL networks. Their approach utilized IoT data from the Cooja simulator and achieved an accuracy of 93% for reduced rank attack detection. The model's effectiveness was evaluated in resource-constrained IoT environments. Verma and Ranga [5] implemented an ensemble learning-based intrusion detection system (ELNIDS) for RPL-based 6LoWPAN networks. Their work addressed DIS flooding attacks using a combination of decision trees and random forests, achieving an accuracy of 94.7%. Cakir et al. [6] introduced a GRU-based deep learning model to detect and prevent RPL-specific attacks. Their approach targeted DIS flooding attacks in IoT networks and demonstrated robust performance with an accuracy of 96.5%. Momand et al. [7] employed multiple machine-learning algorithms for detecting various attacks, including black hole and version number attacks in RPL networks. Using random forest and SVM models, they achieved 91 and 96% accuracy across different attack types. Agiollo et al. [8] developed DETONAR, a system for detecting routing attacks in RPL-based IoT. This system applied a hybrid approach of decision trees and ensemble methods, resulting in accuracy rates of over 95% for detecting multiple attack vectors. Shafiq

et al. [9] evaluated the effectiveness of machine learning algorithms like random forest and SVM in detecting botnet-based attacks in IoT smart city environments. Their method achieved an accuracy of 98.5%, showing excellent performance in identifying complex attack patterns.

Seth et al. [10] focused on detecting decreased rank attacks in RPL-based 6LoWPAN networks using round-trip time measurements. Their machine learning approach using SVM classifiers achieved an accuracy of 89%. Reshi et al. [11] developed a defense algorithm against blackhole attacks in IoT networks, utilizing a trust-based system. Their work achieved 92% detection accuracy in real-time scenarios. Almusaylim et al. [12] proposed SRPL-RP, a machine learning-based framework designed to detect and mitigate rank and version number attacks in IoT environments. Their system achieved accuracy rates of over 90%, demonstrating the robustness of the approach.

3 Methodology

This section describes the detection of four types of security attacks using ensemble learning techniques trained on an existing IoT-RPL dataset. This dataset, publicly available on the Mendeley platform, contains realistic traffic data for various standard and attack scenarios within a simulated RPL-based network. Ensemble learning models, which combine the strengths of multiple base models, are employed to provide robust and accurate attack detection. Feature selection techniques are also applied to reduce the complexity of the detection model and improve performance for resource-constrained IoT environments. Figure 1 shows the steps used to develop the attack detection system.

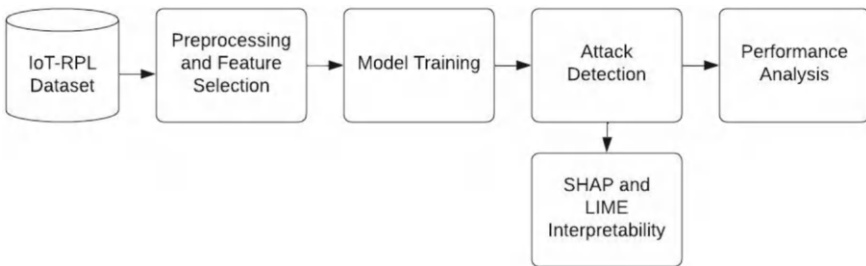


Fig. 1 Methodology adapted to develop IoT-RPL security attack detection system

3.1 IoT-RPL Dataset

The IoT-RPL dataset, which serves as the foundation for this research, is obtained from the Mendeley platform. This dataset contains traffic data generated using the Cooja simulator, a widely used simulation tool for 6LoWPAN networks. The dataset includes various regular network traffic instances and attack scenarios, explicitly focusing on the four types of attacks: Blackhole, Flooding, Decreased Rank, and DODAG Version Number Attacks. Each attack introduces distinct anomalies in RPL control messages, rank values, and other network metrics, making the dataset suitable for training machine learning models.

The dataset consists of labeled instances with various features, including the following network-related attributes:

- RPL control messages: The rate and type of control messages, such as DIO (DODAG Information Object), DAO (Destination Advertisement Object), and DIS (DODAG Information Solicitation).
- Rank values: The rank of each node within the RPL DODAG topology indicates the node's distance to the root.
- Version number: The DODAG version number is manipulated during DODAG Version Number Attacks.
- Energy consumption metrics are essential for detecting resource exhaustion caused by attacks like flooding.
- Packet delivery ratio (PDR): Measures the success rate of data packets delivered to their intended destinations, a key feature impacted by blackhole and decreased rank attacks.

The dataset has been pre-processed to label instances as either standard or attack. For this research, the dataset is further divided into training and testing sets to evaluate the performance of the machine learning models.

3.2 Feature Selection

Given the resource-constrained nature of IoT devices, the efficiency of machine learning models is paramount. Therefore, feature selection plays a critical role in this methodology. Feature selection helps reduce the dimensionality of the dataset, retaining only the most relevant features and reducing the computational complexity of the model while maintaining high detection accuracy.

The following feature selection techniques are applied:

- Filter-based methods: Techniques such as the Pearson correlation coefficient and chi-square test are used to evaluate the importance of each feature by measuring its correlation with the attack labels. Features that have strong correlations are prioritized for inclusion in the model.

- **Wrapper methods:** These methods evaluate the performance of different subsets of features using machine learning models. Using a base model (e.g., Decision Trees), subsets of features are systematically assessed, and those that contribute most to performance are selected.
- **Recursive feature elimination (RFE):** This method recursively removes the least essential features based on model performance, helping narrow down the final set of features that contribute most to attack detection.

The final set of selected features includes RPL-specific metrics, such as the rate of control messages (DIO, DAO, and DIS), rank, version number, hop count, energy consumption, and packet delivery ratio. These features highly indicate the various attack scenarios and are used to train the machine learning models.

3.3 *Ensemble Learning Techniques*

Ensemble learning techniques are employed to enhance the robustness and accuracy of the attack detection model. Ensemble methods combine multiple base learners to create a more accurate and reliable model. This study's main ensemble learning techniques are Bagging, Boosting, and Stacking, each providing unique benefits for attack detection.

3.3.1 **Bagging (Bootstrap Aggregating)**

Bagging works by training multiple classifiers on different subsets of the data generated through bootstrapping (random sampling with replacement). The final prediction is made by aggregating the predictions of all classifiers, typically through majority voting (for classification tasks). Bagging reduces model variance and improves stability.

The Random Forest algorithm, a widely used bagging method, is applied for this research. Random Forest trains multiple decision trees on different subsets of the IoT-RPL dataset. Each tree in the forest provides a classification, and the majority vote of all trees determines the final output. This method is particularly effective in detecting network anomalies, as Random Forest can handle noisy data and remains robust against overfitting.

3.3.2 **Boosting**

Boosting is another ensemble technique that sequentially trains weak classifiers, where each new classifier attempts to correct the errors made by the previous ones. Boosting is designed to reduce bias and variance, making it suitable for learning complex patterns in the data.

In this work, two boosting algorithms are used:

- **Gradient Boosting Machine (GBM):** This method builds models sequentially, with each new model attempting to correct the errors of its predecessor. GBM is robust in handling imbalanced data, which is crucial when specific attacks (such as Decreased Rank Attacks) are rarer in the dataset.
- **AdaBoost:** This is another boosting algorithm that assigns weights to misclassified instances in each iteration, encouraging subsequent models to focus on those complicated cases. AdaBoost is efficient for detecting subtle attacks, such as the DODAG Version Number Attack, which can be difficult to detect through simple thresholding techniques.

Both GBM and AdaBoost are trained on the IoT-RPL dataset and evaluated for their ability to detect the four types of attacks. These algorithms provide high accuracy if the attack patterns are subtle, as they continuously improve the classification boundary.

3.3.3 Stacking

Stacking is a more advanced ensemble technique that combines the predictions of several base models using a meta-model (also known as a meta-learner). The base models provide their forecasts as input to the meta-model, which learns to make a final prediction based on these inputs. In this research, a Stacking ensemble is built using a combination of different classifiers, including Random Forest, Gradient Boosting, and Support Vector Machines (SVM), as base models. The meta-learner chosen is Logistic Regression, which aggregates the predictions from the base models. The idea behind stacking is that different classifiers capture different aspects of the data. For instance, Random Forest may excel in detecting Blackhole and Flooding Attacks due to their distinct network traffic signatures, while Gradient Boosting may be better at detecting Decreased Rank Attacks. By combining these models through a meta-learner, stacking can achieve superior performance over individual classifiers.

3.4 Model Training and Evaluation

The IoT-RPL dataset is split into training and testing sets, with 70% of the data used for training and 30% for testing. The ensemble learning models are trained using the training set, and cross-validation ensures that the models generalize well to unseen data.

- **Model fitting:** Each ensemble model is trained using the selected features from the dataset. For bagging (Random Forest), each decision tree is trained on a bootstrapped training set sample. In the case of boosting (GBM, AdaBoost), each weak learner is sequentially trained to improve performance on misclassified

instances from the previous learners. For stacking, the base models are trained first, and their predictions are fed into the meta-learner (Logistic Regression) for final prediction.

- **Hyperparameter tuning:** Hyperparameter tuning is conducted using grid search and random search techniques to optimize the performance of each ensemble model.

The trained models are then evaluated on the test set using various performance metrics:

- **Accuracy:** The proportion of correctly classified instances out of the total instances.
- **Precision:** The proportion of actual positive attack instances out of all predicted attacks.
- **Recall (Sensitivity):** The proportion of actual attack instances the model correctly identifies.
- **F1-score:** The harmonic mean of precision and recall, providing a balanced measure of model performance.
- **False Positive Rate (FPR):** The proportion of regular traffic instances incorrectly classified as attacks.

4 Results and Discussion

The ensemble models are evaluated for their ability to detect the four types of attacks. The performance metrics for the models are summarized in Table 1. The results show that the Stacking Ensemble model outperforms the individual classifiers in all metrics, achieving the highest accuracy of 99%. The Random Forest model closely follows with an accuracy of 98.5%, while Gradient Boosting and AdaBoost also demonstrate commendable performance but fall short of the ensemble methods.

The ensemble models perform exceptionally well in detecting black holes and flooding attacks, as indicated by the high recall and precision values. However, the Decreased Rank Attack is more challenging, with the AdaBoost model achieving a lower recall rate. Nonetheless, all models maintain a low false positive rate, indicating they are reliable for practical deployment in RPL-based IoT networks.

Table 1 Performance metrics for the models on the IOT-RPL dataset

Model	Accuracy (%)	Precision (%)	Recall (%)	FPR (%)	F1-score (%)
Random forest	98.30	98.30	98.00	98.10	1.50
Gradient boosting	97.60	97.50	97.20	97.30	2.00
AdaBoost	95.60	95.70	95.30	95.50	3.90
Stacking ensemble	99.00	99.10	99.00	99.00	1.10

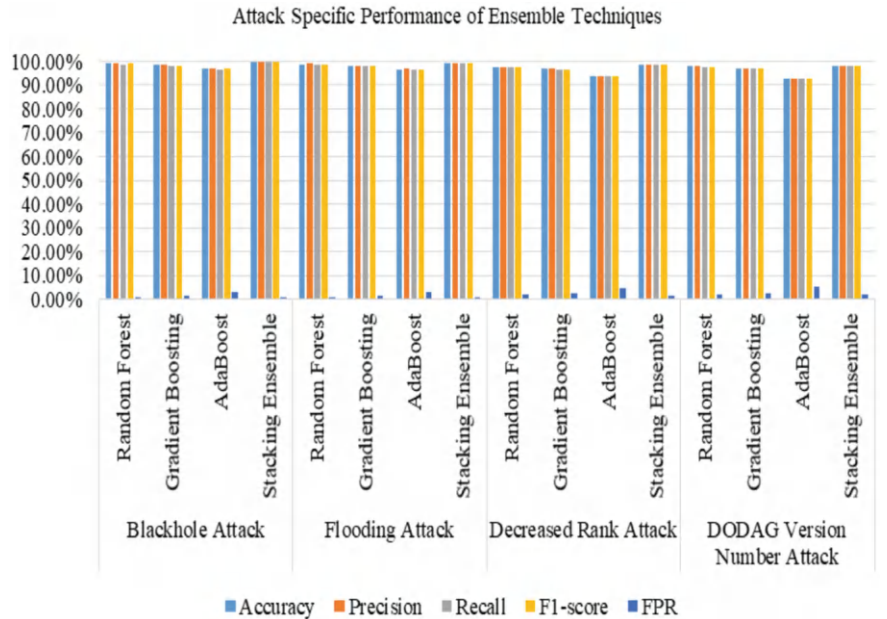


Fig. 2 Attack specific performance of ensemble techniques

4.1 Attack-Specific Performance

For a detailed understanding of each model’s performance in detecting specific attacks, the graph in Fig. 2 summarizes the performance metrics.

The Stacking Ensemble model shows the highest accuracy of 99.4%, with a precision of 99.6% and a recall of 99.4%, making it the best at detecting packet drops caused by blackhole Attacks. All models perform well for Flooding Attacks, but the Stacking ensemble achieves 99.2% accuracy with a 0.7% FPR, making it the most reliable in detecting this attack type. The decreased rank attack proves challenging, especially for the AdaBoost model, which shows a lower recall of 93.5%. The Stacking Ensemble improves detection rates with a recall of 98.6%. DODAG version number attack is subtle, and AdaBoost struggles with a recall of 92.5%, but the Stacking Ensemble still achieves 97.9% recall, ensuring reliable detection.

4.2 Explainability of Models

SHAP and LIME are used to explain the model predictions, offering insights into feature importance and the rationale behind each decision. The following are the results from both explainability techniques.

Table 2 SHAP values for the most critical features across all attack types

Feature	SHAP value
Packet Delivery Ratio (PDR)	0.45
Rank value	0.32
Control messages (DIO, DAO, DIS)	0.29
DODAG version number	0.27
Energy consumption	0.21

SHAP provides a unified measure of feature importance based on cooperative game theory. The SHAP values quantify the average contribution of each feature to the model’s predictions. Table 2 presents SHAP values for the most critical features across all attack types.

The SHAP analysis indicates that PDR (Packet Delivery Ratio) is the most influential feature for Blackhole Attack detection, with a SHAP value of 0.45. A sharp drop in PDR is a strong indicator of this attack. For Flooding Attacks, the SHAP value of 0.29 for Control Messages reflects how excessive DIO or DIS messages significantly contribute to detecting this attack. The Rank Value plays a crucial role, with a SHAP value of 0.32, highlighting how abnormal rank decreases are key indicators of Decreased Rank Attacks. The SHAP value for version number is 0.27, indicating that frequent changes in DODAG version number are the primary features used to detect this attack.

LIME provides local interpretations by generating explanations for individual model predictions. Here are key insights from LIME analysis: In 90% of cases, low PDR values dominate the local explanation, making them the primary factor in correctly identifying Blackhole Attacks. For 85% of correctly detected Flooding Attacks, the high rate of DIO control messages was recognized as the leading cause of classification. LIME shows that sudden drops in Rank Values were responsible for correct predictions in 80% of instances. The frequent increments in Version Number were the primary explanation for 95% of correctly detected DODAG Version Number Attacks.

5 Conclusion

The experimental results demonstrate that ensemble learning techniques, particularly the stacking model, are highly effective for detecting the four types of attacks—Blackhole, Flooding, Decreased Rank, and DODAG Version Number Attacks—in RPL-based IoT networks. By combining multiple base classifiers (Random Forest, Gradient Boosting, and SVM) and using Logistic Regression as a meta-learner, the stacking model achieved superior performance across all attack types, with a notably low False Positive Rate.

Furthermore, incorporating SHAP and LIME provides essential insights into the model’s decision-making process, ensuring interpretability and trustworthiness. The

results indicate that machine learning techniques can effectively safeguard RPL-based IoT networks against prevalent attacks, paving the way for more secure deployments in real-world applications. Future work will focus on real-time implementation and further model refinement based on dynamic network conditions.

References

1. Jayaprakash, P., Lalitha, B.: A Novel Intrusion Detection System for RPL Based IoT Networks with Bio-Inspired Feature Selection and Ensemble Classifier, pp. 1–19. Pokala2021ANI, <https://orcid.org/0000-0002-2131-9454>
2. IoT-RPL 2021: Cyber Attack Dataset Based on RPL Routing for IoT (2024). <https://data.mendeley.com/datasets/4rcbbry2sc/1>, <https://doi.org/10.17632/4rcbbry2sc.1>
3. Touzen, A., et al.: Hybrid deep learning-based intrusion detection system for RPL IoT networks. *J. Sens. Actuator Netw.* 25–31 2023. <https://doi.org/10.3390/jsan12020021>
4. Osman, M., et al.: Artificial neural network for decreased rank attack detection in RPL networks. *Int. J. Netw. Secur.* 120–125 (2021)
5. Verma, A., Ranga, V.: Ensemble learning based IDS for RPL networks. In: *Proceedings of IoT-SIU*, pp. 145–150 (2020)
6. Cakir, S., Toklu, S.: Detection of DIS flooding attacks using GRU. In: *IEEE Access*, pp. 170–176 (2020)
7. Momand, M.D., et al.: Machine learning-based multiple attack detection in RPL networks. In: *Proceedings of ICCCI*, pp. 35–40 (2021)
8. Agiollo, A., et al.: DETONAR: detection of routing attacks in RPL networks. In: *IEEE Transactions on Network and Service Management*, pp. 85–91 (2021)
9. Shafiq, M., et al.: Machine learning for Bot-IoT attack detection. *Future Gener. Comput. Syst.* 145–152 (2020)
10. Seth, A.D., Biswas, S., Dhar, A.K.: Detection and verification of decreased rank attack. In: *Proceedings of IEEE ANTS*, pp. 90–94 (2020)
11. Reshi, I.A., et al.: Mitigating blackhole attacks in IoT networks. *J. Eng. Res.* 210–215 (2024)
12. Almusaylim, Z.A., et al.: SRPL-RP: detection of rank and version number attacks in IoT. *Sensors* 55–61 (2020)

Advancing Detection of Man-in-the-Middle Attacks Through Possibilistic C-Means Clustering



Saswati Chatterjee, Lalmohan Pattnaik, Suneeta Satpathy,
and Deepthi Godavarthi

Abstract Man-in-the-middle attacks are among the most dangerous types of cyber threats, which implies unauthorized interception of information exchange between two or more users. Real-time Identification of these attacks has been deemed particularly difficult because of the complexity of the data traffic and sometimes the overlap of the attack classes. In this work, we aim to improve the detection of these attacks based on the Machine Learning algorithm using the NSL-KDD dataset, a well-known dataset for applications in network intrusion detection. We use Possibilistic C-Means (PCM) clustering as the primary detection method. PCM clustering effectively handles uncertainty and overlapping clusters, making it well-suited for distinguishing Man-in-the-middle attacks from regular traffic. Thus, from this dataset, Chi-square and Information Gain feature selection methods are used to extract the attack features with the most distinguishing attributes. State experiments were performed with the help of an open-source software KNIME (Konstanz Information Miner), and several machine learning algorithms such as Naive Bayes Gaussian, SVM, SVM-SOM K-Means, Random Forest, and PCM clustering were tested. This paper proves that the proposed method, PCM clustering, outperforms other techniques in the actual positive rate and accuracy of identifying the attacks of its high detection rates and its enhanced handling of ambivalent data. This approach shows the strength of the PCM clustering for practical Man-in-the-middle attack detection and confirms the advantage of the proposed method over generic approaches.

Keywords Man-in-the-middle attack · Classifier · K-Means · Possibilistic C-Means clustering

S. Chatterjee
FET, Sri Sri University, Cuttack, India

L. Pattnaik · S. Satpathy
SoA Deemed to Be University, Bhubaneswar, India

D. Godavarthi (✉)
School of Computer Science and Engineering, VIT-AP University, Amaravati, India
e-mail: saideepthi531@gmail.com

1 Introduction

The development of computers and communication, including portable applications and greater networks, has internationally changed the nature of network security. Thus, open computer networks are subject to interruption since Internet attacks and misleading acts have been directed toward businesses and private systems. Today, there are many application areas for modern technologies of automation [1], and their availability is essential for the continuous provision of many important services. Organizations involved in regulation, standardization, and direction aim to improve dependability, safety, and quality/excellence of service. In diagnosing anomalous behavior in Industrial Automation Control Systems (IACS), tracking the control components frequently, especially the Programmable Logic Controllers (PLCs), would be important. It has also been noticed that PLCs facilitating Modbus/TCP [2] are mostly on the receiving end of repeated attacks. These might include trying to flood or amplify a network to perform a DoS [3] objective. Sophisticated attacks such as the Man in the Middle (MITM) assaults use the method of ARP poisoning to disrupt data flow and/or hide it. A direct and purposeful strike against the system of another person or organization is called a network attack. However, with the assistance of information, it is possible to help others, while at the same time, cyberattacks are often a real menace [4]. Several widespread hostile network attacks, for instance, contemporary cyberattacks typically include insider threats, injection of SQL queries, denial of service (DoS), hacking or spoofing, and ransomware, have called for the implementation of more complex remedies that embrace modern advancements such as the use of artificial intelligence and creation of security measures. Cloud has significantly improved scholarly and secure transportation systems since scholars have benefited from the cost and flexibility features. Various facilities provided by cloud computing that are in use today encounter challenges such as privileged access to the platform [5]. Cloud users and supported applications are usually accessed through insecure HTTP, which indicates tremendous vulnerability to external risks and threats. Therefore, it has to limit authority access and maintain the change efficiently in the structure, as cloud processing servers use similar working systems. Indirect construction processes of the weaknesses make them a severe threat to the whole distributed computing environment. In today's complex business environment, most organizations must have adequate computing skills to communicate through networks. Network users have to trust the privacy and security of their communications to learn effectively and share information across such environments. However, most attackers build techniques designed to compromise the network through eavesdropping, tampering with, or intercepting what should be private messages. Based on Najafabadi et al. [6, 7], it is estimated that 63% of specified data breaches can be linked to attacks that target credentials [8]. One technique in which such operations can be done is the Man-in-the-Middle (MITM) assault. The attacked computer [9, 10] is in the middle of a dialogue between two connected machines; MITM attacks are still a current threat [11] and aim at eavesdropping on the conversation. Network security and anomaly detection are priorities for ISPs because these organizations

have to manage increasing events that threaten the network's capability and availability. Since the data dimensionality that current network monitoring systems handle is vast, using machine learning to detect and classify anomalous events is possible and practical. This must be tricky because it is quite well established that no set solution exists for all problems in a single instance when choosing the right machine-learning approach to a given problem. Although many models may be incredibly valuable regarding a particular issue, it could be awkward when struggling to select a suitable set for various patterns in conjunction with analytical blends. The blending method allows combining one or many models to create a new model, which can (potentially) be better. In ensemble approaches, several learning algorithms are applied to achieve better prediction potential than using any individual learning method. In theory, the general concept of aggregation, or ensemble learning, is more complex than a single-based learning algorithm. However, this constraint is slightly addressed because most large data platforms and regular analysis power can quickly execute many algorithms simultaneously [12]. The current study incorporates artificial intelligence to enhance the protection of the organizational network by developing an MITM attack detection model. The information and communication are analyzed for the desired data using the machine learning algorithm and looking for MITM attacks. The experiment results show that classifiers are better than standardized learning. The authors employ machine learning in the current study to enhance the organizational network's security by developing an MITM attack model. The network traffic is monitored for specific data by machine learning technology and by searching MITM attacks. The results of the experiment presented here show that classifiers are superior to standard learning techniques.

2 Related Work

Efficient and secure communication over computer networks is essential for modern industries. Network users must trust that their communications are private and safe to exchange sensitive information. However, many attackers attempt to compromise network security through techniques that enable them to steal, modify, or monitor confidential communications. According to the Verizon Data Breach Investigations Report, 63% of confirmed data breaches are attributed to credential theft attacks. One prominent method for achieving this is the Man-in-the-Middle (MITM) attack, which remains a significant threat today. MITM attacks involve intercepting communications between two machines by positioning the attacker's system between them. The attacker can monitor the communication or modify the data before forwarding it to the intended recipient. Both passive and active MITM attacks pose serious security risks, with passive attacks being potentially more insidious. MITM attacks can also lead to malicious activities, such as Distributed Denial of Service (DDoS) or Domain Name System (DNS) spoofing. Communication through computer networks is now required in almost all branches of modern industry and commerce. There exists a mandatory need to ensure that the users of that network are assured that

their communication information is safe and secure. However, most attackers are interested in violating the security of computer networks in a way that allows them to intercept, alter, or eavesdrop on sensitive messaging. The Verizon Data Breach Investigations Report shows that about 63% of confirmed data breaches result from different attacks seeking to obtain users' credentials. It has been considered that both passive and active MITM attacks would seriously jeopardize security measures, but the potential for passive attacks is lurking more. This MITM attack can also result in other evil doings, such as Distributed Denial of Service (DDoS) or Domain Name System (DNS) spoofing. Due to the MITM attack's flexibility can be carried out using various techniques based on the network topology and/or attack objectives: passive and/or active. The most common forms of computer attacks include ARP spoofing, DHCP spoofing, and port stealing. These methods are easy to implement and use due to their availability; moreover, the variety of the methods covers all necessary aspects connected with MITM attack behaviors and traffic characteristics. The previous research on capturing MITM attack traffic is often used on closed networks and does not consider real environment characteristics. Most of these experiments don't consider the regular traffic experienced in live networks, such as web browsing, file transfers, server interactions, and multimedia streaming. However, most of such research centers on crude data download objectives regarding the FTP or file transfer protocol, which are somewhat constrained in their range.

In contrast, our approach involves capturing MITM attack traffic on a live, large-scale campus network with over 60 active users. The traffic includes a wide range of everyday network activities, and attacks are directed toward machines actively engaged in everyday tasks. Our goal is to explore the challenges of capturing MITM attack traffic in a live environment and identify patterns and behaviors characteristic of such attacks. We also analyze the collected traffic and describe our unique labeling procedure. We can distinguish between regular and attack traffic by focusing on specific attack behaviors. We observe distinct relationships between IP addresses and their associated Media Access Control (MAC) addresses for each MITM attack variant. These patterns enable us to label traffic as part of an attack accurately. Our analysis further discusses the behaviors and anomalies associated with MITM attacks. On the other hand, this paper is based on capturing the MITM attack traffic on a live large-scale campus network with over 60 active users. The traffic encompasses many ordinary network activities, and attacks are launched on busy machines dealing with regular business. This work aims to analyze the prospects of capturing MITM attack traffic in the real environment and the typical patterns and behavior related to these attacks. We also discuss the collected traffic analysis and give an overview of our peculiar labeling process. One can differentiate between regular and attack traffic by examining precise attack behaviors. In each MITM attack variant, the researcher notes different correlations between the IP addresses and the respective MAC addresses. These patterns help us correctly identify traffic as attack traffic. We then outlined the behaviors and anomalies related to MITM attacks in our further analysis section. Section 7 presents a case study to show how these attack behaviors affect the detection performance. The researcher has given [13] a comprehensive survey of the different ML techniques used in traffic identification. There are only a few

comprehensive ML-based algorithms for network anomaly identification presented in Chuang and Ye [14], Mittal et al. [15] and Lu et al. [16] and broad domain anomaly detection methodologies. The researcher has extended the utilization of the learning methodologies to address privacy. However, minimal literature has specifically used ensemble learning approaches for confidentiality and the identification of anomalies. However, it is often noted in the methodology that groups appear to produce better results if there is sufficient variation among the predictions about them [17, 18]. In Jingle and Rajsingh [19], the authors proposed dealing with this problem by implementing monitoring agents at different levels of the network. During experimentation, a global monitoring agent was used on the side of the gateway router.

Remote surveillance monitors across the network also kept a table regarding address information and timer values, which provided details. The timer value indicated the nodal presence and how long the node had been present. The primary focus of the defense against Man-in-the-middle attacks is to invent new protocols, sensors, or security techniques for various systems [20]. A new strategy to resolve the IP/MAC mapping problem is needed for the researcher's proposed modification of the present ARP to function against APR poison-based MIMT attacks. An unsupervised WLAN method was developed by researchers [21, 22]. The experts [23] proposed the MITM detector, Vesper. To illustrate the shape of the setting, it uses architecture as an idea analogous to Reflections in the Cave. It also captures subtle changes that enable accurate identification of MITM assault attacks. Internet-related applications for man-in-the-middle (MITM) assault requires a perpetrator to take over an internet connection between two different ends. The contact traffic of the victims included message exchanging and phone calls and was listened to, monitored, and manipulated by the MITM attacker [20]. The research study established that any information transmitted between endpoints is possibly vulnerable to MITM attacks in OSI model layers.

Another experiment was conducted regarding the Artificial Neural Network Solution. The experts explained a DNS spoofing type based on the MITM attack with the help of the ANN type [24]. Thus, the defensive strategy could be considered relatively successful by looking at the total discovery ratio of 98%. The present study used SOM-based visualization because data with high dimensionality could be mapped into low dimensions, exposing a large amount of information simultaneously. This was done because the occurrence of access to data, as well as the absence of effective forensic investigation procedures for digital crime scenarios, are making digital crimes a big issue. IoT devices are not developed with privacy in mind because resource limitations are the primary constraint in IoT implementation. Therefore, it becomes a challenging task to build a reliable IoT protection system that can identify security intrusions [25]; the MANET [26] usually assumes that every node in the network is reliable. One of the dangerous nodes damaging a MANET is a node that misinterprets the channel of communication between the originator and the recipient to launch what is referred to as the "man-in-the-middle (MITM)" attack [27, 28]. Other attacks referred to as "The-Middle" (MITM), in which the attacker intercepts the communication media across wireless communication networks, can easily compromise the security of wireless fidelity networks [29]. Phishing [30] today is

the most common online threat due to the World Wide Web's volume growth rate [31]. In addition to making a framework recommendation for the current study, the current work uses experimental analysis to fill any gap that earlier studies may have left.

The success of a predictor is defined by how closely it was able to estimate the intuitive forecast or the most stated perfect classifier based on the actual collection of data. The predictability of all techniques depends on the problems they have to solve, and due to training datasets, each algorithm learns different kinds of factual characteristics. There are various types of supervised learning, including states that simple learning models or single hypothesis algorithms may experience one of three different bottlenecks: When it comes to the amount of data for training, there are too many hypotheses; statistical problem: when most methods have similar accuracy and the risk of choosing a process that cannot accurately predict the following data sets; computational challenge Many algorithms may not find the world's optimum; A representation issue arises from the non-existence of a model in the set of hypothesis that matches the actual distribution. Ensemble approaches build some hypotheses and choose one using the combinatorial form instead of choosing the optimal framework for reporting the facts. This strategy seeks unity where the learning limitations of one predictor are balanced by the other. Therefore, different predictions are arrived at if several of these algorithms are executed simultaneously. Some articles have attempted to address techniques that exploit the existence of the above variety to enhance the general predictive performance indicated by the merged results of several algorithms [32, 33]. This is done using a process known as ensemble learning, where the result of the predictors is passed into a completely new algorithm referred to as the second-level or meta-learner.

3 Proposed Methodology

This study investigates the extraction of a relevant feature set for detecting Man-in-the-Middle (MITM) attacks using two distinct sources of Internet traffic datasets: The results are obtained based on the NSL-KDD Cup dataset available to the public and traffic data from the Smart and Secure Environment (SSE) network. MITM attack detection investigations: In this case, the ability to identify traffic parameters that indicate deviation in traffic is emphasized. Initially, twenty-three features were selected; Information gain and chi-square tests applied yielded eight relevant features at 1-s intervals. Due to the pleasant distinction between attack and regular traffic classes, the study used various machine learning algorithms such as SVM, KNN, Naïve Bayes, Decision Tree, K-means clustering, and possibilistic C-means clustering. Possibilistic C-means clustering is most helpful as this form of clustering is less strict about assigning the data point to the clusters since traffic analysis separates the regular traffic from the attack traffic. This method helps to analyze the nature of uncertainty in traffic classification by highlighting the possibilities of MITM attacks. Information Gain is a successful strategy for ranking attributes linked

to the target class. In this respect, we proposed how it is possible to choose the most important features based on rank. In more detail, the Chi-Square test determines the level of dependency of feature X on cluster Y. The significance of the feature and target relationship can be tested by comparing the result to a chi-square distribution with one degree of freedom.

$$A^2 = \sum_{i=1}^z \sum_{j=1}^m \frac{(Y_{i,j} - R_{i,j})}{G_{i,j}}$$

(1)

In this regard, Eq. (1) denotes the feature K, while m symbolizes the clusters. $R_{i,j}$ Is the count of feature values that occur with cluster j, and the expected count given the occurrence of feature value I and cluster j is represented by $G_{i,j}$ This indicates that the higher the A^2 , the higher the relevance of the feature to the cluster has been recognized. Using both values and information gain, eight features of significance emerged: The following table, Table 1, ranks these features using the chi-square and information gain methods.

Features Extraction of Data

The firm correlated details that need to be used to detect MITM attacks are highlighted by correlating all features. It is also necessary to include all the 14 attributes indicated for evaluation and forecasts of reference samples to identify the anomaly identification outcomes learning model. The lack of certain features in samples from the model will impact the results; therefore, features should be kept to a minimum to develop a sound model. The Heat Map correlation matrix is valuable in determining

Table 1 Dataset parameters and its description

Feature name	Outline
IPV4_SRC_ADDR	IPv4 source address
L4_SRC_PORT	IPv4 source port number
IPV4_DST_ADDR	IPv4 destination address
L4_DST_PORT	IPv4 destination port number
PROTOCOL	IP protocol identifier byte
L7_PROTO	Layer 7 protocol (numeric)
IN_BYTES	Incoming number of bytes
OUT_BYTES	Outgoing number of bytes
IN_PKTS	Incoming number of packets
OUT_PKTS	The outgoing number of packets
TCP_FLAGS	Cumulative of all TCP flags
FLOW_DURATION_MILLISECONDS	Flow duration in milliseconds
Label	Anomaly or typical conduct
Attack	Type of Attack

related variables, making it easier to choose features. Heat maps help interpret the data received more simply than other visualizations.

Based on the Chi-Square test and Information Gain shown in Table 2, the following eight features were identified as the most significant in detecting Man-in-the-Middle (MITM) attacks: **IPV4_SRC_ADDR**: The source IP address plays a central role in differentiating between regular and attack traffic. While using a ‘reliable’ network source address is impossible, it would be easier to flag unknown ones as suspicious or at least strange. **L4_SRC_PORT**: Source port numbers, which denote the port from which the data in a network transmission started, should also be given to determine the service or application used. In some instances, particular ports may be attacked in MITM attacks. **IPV4_DST_ADDR**: The destination IP address is quite valid to monitor the traffic direction. Some attacks are as follows: an attack may be directed to specific addresses or divert traffic to a particular undesirable URL. **L4_DST_PORT**: The destination port number conveys the service or application the data packet uses. Some ports, such as HTTP 80 or HTTPS 443, are either blocked or recognized to be attacked more often when compared to others. **PROTOCOL**: This feature designates the transport layer protocol the connection uses, for instance, TCP, UDP, or ICMP. Also, different protocols can act on different warding attacks, hence the need for protocol identification and classification. **IN_BYTES**: The total number of received bytes. This parameter gives an overall picture of how many bytes are being received. Large fluctuations in the incoming traffic volume might be used to identify malicious traffic, including a sizeable number during an MITM attack. **OUT_BYTES**: Canceled bytes are also essential for identifying deviant data transmission trends, just like the total outgoing bytes. High volumes in the outbound traffic often indicate unauthorized data leakage—a typical attribute of MITM attacks. **TCP_FLAGS**: This feature monitors the TCP flags, including SYN, ACK, or FIN, amongst other variables. Some TCP flags may point to an attack since the opponent may change them to interrupt or gain control of a session. Among the 27 features considered, the ones pointed out here as the eight most suitable are: These features, chosen according to Chi-Square and Information Gain results, produce powerful signs for flagging anomalies and categorizing traffic as either standard or attack-related when the goal is MITM detection.

Table 2 Feature description and chi-square test and information gain

Features	Chi-square rank	Information gain rank
IPV4_SRC_ADDR	1	1
L4_SRC_PORT	2	2
IPV4_DST_ADDR	3	3
L4_DST_PORT	4	4
PROTOCOL	5	5
IN_BYTES	6	6
OUT_BYTES	7	8
TCP_FLAGSS	8	7

Machine Learning

Machine learning uses an algorithm to teach a technique for making accurate predictions from large astronomic data sets regarding one type of attribute. Machine learning is on the cutting edge of artificial intelligence and computer framework research. The approach created here uses a combination of features to determine a man-in-the-middle attack. This enables more accurate attack detections from the computer program and the machine learning algorithm.

Random Forest

In this study, a well-liked supervised learning method is Random Forest. They are utilized in regression and classification using machine learning. It uses ensemble learning to handle complicated problems and enhance model performance by combining many classifiers. The algorithm is predicted in classification problems, such as when recognizing an MTM attack. In contrast, in regression problems, the technique projected is calculated as the mean of all tree forecasts.

SVM

A supervised learning method, SVM, is frequently employed to address classification and regression problems. Its prominence is due to its being commonly utilized in machine learning for categorization [34]. The idea of statistical learning serves as the foundation for the machine learning paradigm known as Support Vector Machine (SVM) [35, 36]. Instead of merely reducing the data set's mean square error, the approach minimizes the model's generalization error's outer bounds. The SVM algorithm can provide more accuracy since it can show greater accuracy on massive datasets.

K-Means Clustering

It works in the following way: first, the clusters are randomly assigned, and data points are associated with the clusters depending on distances between these points and the centroids of respective clusters. It also assigns each data point to the centroid with which they are most alike, which helps create rather unique clusters.

Naïve Bayes Gaussian

Based on Bayes' theorem, the Nave Bayes approach is employed for categorization. The Nave Bayes Classifier is among the more effective and basic categorization methods for quickly building machine learning models that can immediately make forecasts. The class-based attributes can be identified by continuous characteristics using the Bayesian classification Gaussian function [37].

Hybrid Classifier

Support Vector Machine (SVM) and Self Organized Map (SOM) are two machine learning-based models we have combined to handle this DDoS onslaught. We started with the SVM and SOM separately. Furthermore, we found that SOM outperforms SVM in terms of assault detection. To increase performance, we combined the SVM

and SOM, and the results show better detection rates, accuracy, and false rates than the standalone implementation. This section covered the operation of two algorithms and our suggested hybrid machine-learning algorithm.

Possibilistic C-Means (PCM) Clustering

PCM clustering is a variant of the C-means clustering study. It is closer to the actual data than the conclusive facts because it considers concepts of possibility theory in clustering. PCM is more forgiving of cluster membership compared to the other approaches, allowing for the decision of which data points to assign to which cluster to be made with more flexibility. The method can be formulated as follows in the Eq. (2):

$$J = \sum_{j=1}^m \sum_{i=1}^n \Pi_{i,j}^\mu d(x_i, v_j)^2 + \sum_{j=1}^m \lambda_j \sum_{i=1}^n (1 - \Pi_{i,j})^v \quad (2)$$

where:

$\Pi_{i,j}$ is the possibility that data point x_i belongs to cluster j .

v_j is the centroid of cluster j .

$d(x_i, v_j)$ represents the distance between data point x_i and cluster centroid v_j (usually the Euclidean distance).

μ is a parameter controlling the influence of the distance term in the objective function.

v is a parameter controlling the influence of the possibility term in the objective function.

λ_j is a parameter balancing the two terms in the objective function for cluster j .

n is the number of data points.

m is the number of clusters.

The objective function seeks to minimize the sum of two terms: the distance-based clustering term and the possibility-based term, which handles uncertainty in cluster membership. The parameters control the balance between these terms μ , v , and λ_j .

Data Set

The origin of the dataset used in the execution of this study is discussed in this section. The research's datasets are derived for experimentation from the Machine Learning-Based NIDS Dataset, the most significant data science community globally [38].

4 Performance Evaluation

The dataset was collected on the SSE (Simulation of Snapshot Ethernet) network and contains both attack and regular traffic. Standard traffic data was collected from the SSE network for classification, while the attack traffic data was collected from

the NSL-KDD set. The analysis and classification were performed using the open-source KNIME (Konstanz Information Miner), version 3. The correct classification results are included in Table 3. Table 4 presents the F-measure percentage, Fig. 1 demonstrates the ROC of the machine learning algorithms, and Fig. 2 represents the Correlation Heat Map. In these experiments, the based classifier outperforms all classifiers regarding recall rate for MITM attack detection. This approach managed uncertainty and overlapping clusters better than other methods and was more efficient at identifying six different types of attacks in the given data set. PCM (Possibilistic C-Means) is used as it is more suitable than conventional k-means or fuzzy c-means methods and takes into account uncertain factors and overlapping of clusters. Unlike the traditional methods that categorize data points into clusters using probabilities or distances, PCM clustering permits altering the membership values as they do not have to sum up to 1. This makes it easier for PCM to manage problems with overlapping clusters since the degrees of membership are variable and can give high membership to one cluster and low membership to other clusters for a given data point or give low membership to all clusters for noisy data points. By formally allowing such uncertainty, the specified approach enables the classifier to produce more accurate and timely results that are easier to interpret rather than quantifying them under a specific number of categories. Moreover, PCM clustering is also very efficient for extensive voluminous high-dimensional data, which can be seen from handling overlapping clusters and the capacity to distinguish between six different types of attacks in the dataset provided. These strengths make PCM a reliable method for real-world scenarios such as MITM attack detection since excellent levels of uncertainty and overlapping patterns characterize the problem.

The PCM clustering was evaluated qualitatively using several parameters with a special emphasis on recall rate that quantifies true positives for identification of MITM attacks. Other measures considered in overall classification performance included accuracy, F-measure computed as the arithmetic mean of precision and recall, and ROC analysis. These metrics offered a balanced assessment so that all of the method’s performance characteristics were investigated and compared to other methods regarding uncertainty handling, overlapping clusters, and attacks.

$$\text{Precision} = \text{TP} / \text{TP} + \text{FP}$$

Table 3 Classification analysis

Method used	Classification %	Detection time
Probabilistic C-Means	96.5	0.16
Naïve Bayesian Gaussian	95.3	0.54
SVM	93.2	0.28
SVM-SOM	92.3	0.26
Random forest	94.2	0.24
K-Means	95.5	0.21

Table 4 The F-measure percentage

Method used	TP	FP	TN	FN	F-measure
Probabilistic C-means	278	2	265	3	0.988
Naïve Bayesian Gaussian	285	10	254	17	0.955
SVM	279	21	242	31	0.914
SVM-SOM	282	18	254	22	0.9344
Random forest	275	16	274	0	0.971
K-means	276	23	219	54	0.877

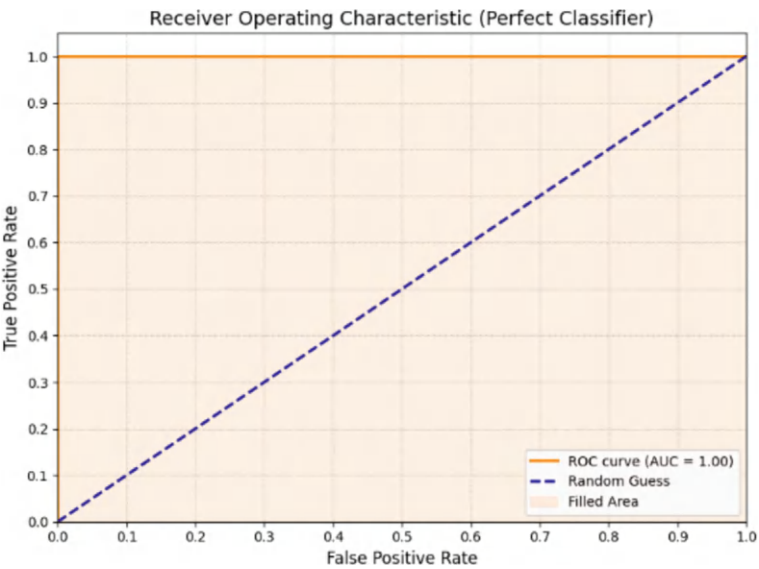


Fig. 1 ROC curve

Calculated attack data from recall is compared to all attack data.

$$\text{Recall} = \text{TP} / \text{TP} + \text{FN}$$

The harmonic mean of recall and precision is referred to as the F-measure.

$$\text{F - measure} = 2 \text{ Precision} \cdot \text{Recall} / \text{Precision} + \text{Recall}$$

The detection rate is the proportion of attacks that are accurately identified among all anticipated attacks.

$$\text{Decision Rate} = \text{TP} / (\text{TP} + \text{FP}) * 100$$

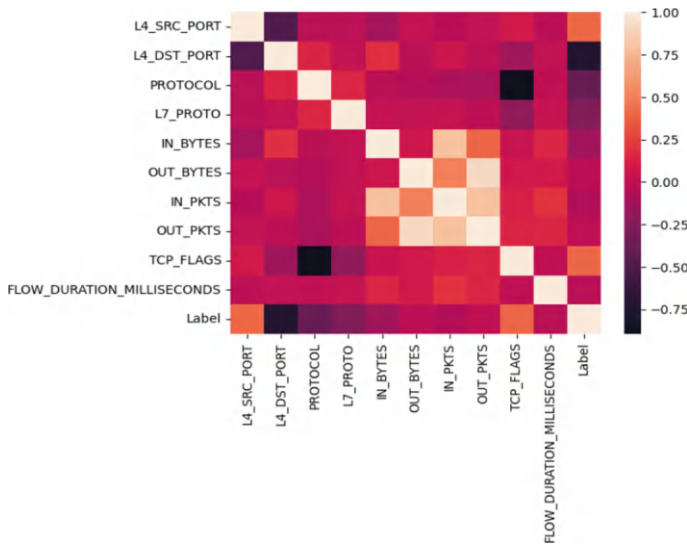


Fig. 2 The attributes of the Correlation Heat Map

5 Conclusion

The performances of various machine learning algorithms for MITM attack detection are analyzed and compared in this paper using the NSL-KDD dataset. The dataset forms the basis for regular and attack traffic data. By applying Chi-Square and Information Gain ranking techniques, only essential features are chosen well to improve the classification. The effectiveness of the Possibilistic C-Means clustering algorithm is shown in this work by comparing the results of classifications yielded by this method to those achieved by other methods. Compared with most clustering methods, Possibilistic C-Means work well with ambiguity and data overlapping, giving a better probe into the cluster membership degree. It also increases the likelihood of classifying normal traffic flow from malicious traffic, thereby increasing the success rate of MITM attack detections. In addition, Possibilistic C-Means brings a higher efficiency than other machine learning algorithms because of its fast rate. These results, therefore, imply the appropriateness of selecting the right features and the proper partitioning method for improving the stability of the MITM detection systems. Altogether, this work contributes useful findings in enhancing machine learning strategies appropriate to cybersecurity, especially in the fight against MITM insecurity. This study demonstrates the practical potential of PCM clustering in real-world network security scenarios and underscores its advantage over traditional approaches for MITM attack detection.

Acknowledgements Completing this research project would not have been possible without the contributions and support of my guide. I am deeply grateful to her for their invaluable input and support throughout the research process.

References

1. Pretticco, G., Flammini, M., Andreadou, N., Vitiello, S., Fulli, G., Masera, M.: Distribution System Operators Observatory 2018. Publications Office of the European Union (2019)
2. Organization, M.: Modbus Messaging on TCP/IP Implementation Guide v1.0b. <https://modbus.org/docs/ModbusMessagingImplementationGuideV10b.pdf> (2006)
3. Huseinović, A., Mrdović, S., Bicakci, K., Uludag, S.: A survey of denial-of-service attacks and solutions in the smart grid. *IEEE Access* **8**, 177447–177470 (2020)
4. Sarker, I.H., Furhad, M.H., Nowrozy, R.: Ai-driven cybersecurity: an overview, security intelligence modeling, and research directions. *SN Comput. Sci.* **2**, 1–18 (2021)
5. Sarker, I.H., Kayes, A.S.M., Badsha, S., Alqahtani, H., Watters, P., Ng, A.: Cybersecurity data science: an overview from a machine learning perspective. *J. Big Data* **7**, 1–29 (2020)
6. Najafabadi, M.M., Khoshgoftaar, T.M., Napolitano, A.: Detecting network attacks based on behavioral commonalities. *Int. J. Reliab. Qual. Saf. Eng.* **23**(01), 1650005 (2016)
7. Najafabadi, M.M., Khoshgoftaar, T.M., Seliya, N.: Evaluating feature selection methods for network intrusion detection with Kyoto data. *Int. J. Reliab. Qual. Saf. Eng.* **23**(01), 1650001 (2016)
8. Wang, D., Chen, X., Chen, D.: DDoS detection method based on instance transfer learning. In: 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, pp. 1053–1058 (2021). <https://doi.org/10.1109/ICSIP52628.2021.9688784>
9. Singh, D., Shukla, A., Sajwan, M.: Deep transfer learning framework for the identification of malicious activities to combat cyberattacks. *Future Gener. Comput. Syst.* **125**, 687–697 (2021)
10. Mittal, M., Kumar, K., Behal, S.: DDoS-AT-2022: a distributed denial of service attack dataset for evaluating DDoS defense system. *Proc. Indian Natl. Sci. Acad.* **89**(2), 306–324 (2023)
11. Rodríguez, E., Valls, P., Otero, B., Costa, J.J., Verdú, J., Pajuelo, M.A., Canal, R.: Transfer-learning-based intrusion detection framework in IoT networks. *Sensors* **22**(15), 5621 (2022)
12. Xia, Y., Xu, Y., Mondal, S., Gupta, A.K.: A transfer learning-based method for cyber-attack tolerance in distributed control of microgrids. *IEEE Trans. Smart Grid* (2023)
13. Kawish, S., Louafi, H., Yao, Y.: An instance-based transfer learning approach, applied to intrusion detection. In: 2023 20th Annual International Conference on Privacy, Security and Trust (PST), pp. 1–7 (2023)
14. Chuang, H.-M., Ye, L.-J.: Applying transfer learning approaches for intrusion detection in software-defined networking. *Sustainability* **15**(12), 9395 (2023)
15. Mittal, M., Kumar, K., Behal, S.: DL-2P-DDoSADF: deep learning-based two-phase DDoS attack detection framework. *J. Inf. Secur. Appl.* **78**, 103609 (2023)
16. Lu, H., Zhao, Y., Song, Y., Yang, Y., He, G., Yu, H., Ren, Y. (2024). A transfer learning-based intrusion detection system for zero-day attack in communication-based train control system. *Clust. Comput.* 1–16
17. Effah, E.Q., Osei, E.O., Tetteh, A.: Hybrid approach to classification of DDoS attacks on a computer network infrastructure. *Asian J. Res. Comput. Sci.* **17**(4), 19–43 (2024)
18. Ozdemir, M., Sogukpinar, I.: An android malware detection architecture based on ensemble learning. *Trans. Mach. Learn. Artif. Intell.* **2**(3), 90–106 (2014)
19. Jingle, D.J., Rajsingh, E.B.: Defending IP spoofing and TCP SYN flooding attacks in next-generation multi-hop wireless networks. *Int. J. Inf. Netw. Secur.* **2**(2), 160 (2013)
20. Conti, M., Dragoni, N., Lesyk, V.: A survey of man-in-the-middle attacks. *IEEE Commun. Surv. & Tutor.* **18**(3), 2027–2051 (2016)
21. Nam, S.Y., Jurayev, S., Kim, S.S., Choi, K., Choi, G.S.: Mitigating ARP poisoning-based man-in-the-middle attacks in wired or wireless LAN. *EURASIP J. Wirel. Commun. Netw.* **2012**, 1–17 (2012)
22. Ghafir, I., Kyriakopoulos, K.G., Aparicio-Navarro, F.J., Lambbotharan, S., Assadhan, B., Binsalleeh, H.: A basic probability assignment methodology for unsupervised wireless intrusion detection. *IEEE Access* **6**, 40008–40023 (2018)
23. Mirsky, Y., Kalbo, N., Elovici, Y., Shabtai, A.: Vesper: using echo analysis to detect man-in-the-middle attacks in LANs. *IEEE Trans. Inf. Forensics Secur.* **14**(6), 1638–1653 (2018)

24. Bai, X., Hu, L., Song, Z., Chen, F., Zhao, K.: Defense against DNS man-in-the-middle spoofing. In: Web Information Systems and Mining: International Conference, WISM 2011, Taiyuan, China, September 24–25, 2011, Proceedings, Part I, pp. 312–319. Springer, Berlin (2011)
25. Samantaray, M., Satapathy, S., Lenka, A.: A systematic study on network attacks and intrusion detection system. In: Skala, V., Singh, T.P., Choudhury, T., Tomar, R., Abul Bashar, M. (eds.) Machine Intelligence and Data Science Applications. Lecture Notes on Data Engineering and Communications Technologies, vol. 132. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-2347-0_16
26. Almarshdi, R., Nassef, L., Fadel, E., Alowidi, N.: Hybrid deep learning-based attack detection for imbalanced data classification. *Intell. Autom. & Soft Comput.* **35**(1), 297–320 (2023)
27. Javeed, D., MohammedBadamasi, U., Ndubuisi, C.O., Soomro, F., Asif, M.: Man in the middle attacks: analysis, motivation, and prevention. *Int. J. Comput. Netw. Commun. Secure* **8**(7), 52–58 (2020)
28. Shekhar, S., Mahajan, M., Kaur, S.: A comprehensive review of various attacks in mobile ad hoc networks. In: 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), pp. 638–643. IEEE (2022)
29. Das, K., Basu, R., Karmakar, R.: Man-in-the-middle attack detection using ensemble learning. In: 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1–6. IEEE (2022)
30. Potluri, S., Mangla, M., Satpathy, S., Mohanty, S.N.: Detection and prevention mechanisms for DDoS attack in cloud computing environment. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, pp. 1–6 (2020). <https://doi.org/10.1109/ICCCNT49239.2020.9225396>
31. Yadollahi, M.M., Sholeh, F., Serkani, E., Madani, A., Gharaee, H.: An adaptive machine learning-based approach for phishing detection using hybrid features. In: 2019 5th International Conference on Web Research (ICWR), pp. 281–286. IEEE
32. Freund, Y., Schapire, R.E., Singer, Y., Warmuth, M.K.: Using and combining predictors that specialize. In: Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing, pp. 334–343 (1997)
33. Dietterich, T.G.: Ensemble learning. *Handb. Brain Theory Neural Netw.* **2**(1), 110–125 (2002)
34. Chatterjee, S., Satpathy, S.: Artificial Intelligence in Decision Making System (2022)
35. Ponomarev, S., Atkison, T.: Industrial control system network intrusion detection by telemetry analysis. *IEEE Trans. Dependable Secure Comput.* **13**(2), 252–260 (2015)
36. Gonzalez, J., Papa, M.: Passive scanning in Modbus networks. In: International Conference on Critical Infrastructure Protection, pp. 175–187. Springer, US, Boston, MA (2007)
37. Bustamante, C., Garrido, L., Soto, R.: Comparing fuzzy Naive Bayes and Gaussian Naive Bayes for decision-making in RoboCup 3d. In: MICA 2006: Advances in Artificial Intelligence: 5th Mexican International Conference on Artificial Intelligence, Apizaco, Mexico, November 13–17, 2006. Proceedings 5, pp. 237–247. Springer, Berlin (2006)
38. https://staff.itee.uq.edu.au/marius/NIDS_datasets/

CNN-Based IDS for Internet of Vehicles Using Transfer Learning



Samarjeet Singh Rathore, Shaurya Yadav, Nitin Singh, and Biswajit Brahma

Abstract This paper focuses on comparing an RNN and an LSTM model to a CNN-based Intrusion Detection System (IDS) to secure vehicular ad hoc networks (VANETs) against cyber threats like DoS and spoofing attacks, which will become more and more common as technology advances in Internet of Vehicles (IoV). Using different Neural Network models, the Intrusion Detection System is being compared based on accuracy, f1 score, precision, and recall. This helps us to understand which model will perform better under real-time attack scenarios and why it is better than the other models. All the models are trained on a standard input dataset (CICIDS2017) and processed to give an accuracy table. This dataset uses 81 attributes like packet size, packet rate, and other factors to consider for an intrusion attack. The project's accuracy is considered by how efficiently it identifies an intrusion in the system based on the input. To improve road safety and traffic efficiency, modern vehicular networks, which are essential to the Internet of Vehicles (IoV), allow communication between vehicles and infrastructure. However, the increased connectivity creates a lot of openings for cyberattacks like Distributed Denial-of-Service (DDoS) attacks, data injection, and message tampering. The high mobility and dynamic topology of vehicular networks present challenges for traditional IDS. That is why these challenges are recognized using this comparative study of conventional and new models for Intrusion Detection.

S. S. Rathore (✉) · S. Yadav · N. Singh
School of Computer Science and Engineering (SCOPE), VIT-AP University, Amaravati, India
e-mail: samarsinghrathore29@gmail.com

S. Yadav
e-mail: shauryakosli@gmail.com

N. Singh
e-mail: snitin.imk@gmail.com

B. Brahma
McKesson Corporation, Fremont, CA 94555, USA
e-mail: biswajit.brahma@gmail.com

Keywords Intrusion Detection System (IDS) · Internet of Vehicles (IoV) · Vehicular · Accuracy · Machine learning · Deep learning · Convolutional · Dataset · Vehicular Ad Hoc Networks (VANET) · AI

1 Introduction

The Internet of Vehicles (IoV) framework links automobiles, infrastructure, devices, and users online to build intelligent, effective, and secure transportation systems. Communication between vehicles (V2V), vehicles and infrastructure (V2I), pedestrians (V2P), and cloud services (V2C) is made possible, for which it uses cutting-edge technologies like sensors, GPS, and communication modules. For safe self-driving cars and maps for navigation, the Internet of Vehicles (IoV) facilitates real-time communication with sensors and traffic systems. It uses real-time data to manage accidents, optimizes traffic flow, and lessens congestion. It also improves infotainment services by providing drivers and passengers with customized media. However, as advances are being made, technology becoming prone to cyber-attacks is a concern. As Fig. 1 shows, the percentage of attacks on IoVs is vast, and attackers are improving with the improving technology.

In past cases, the datasets and models created could not detect attacks accurately. The trained models used RNN, LSTM, and other artificial neural networks in real-time scenarios.

Therefore, implementing AI/Deep Learning models is becoming necessary in this field to safeguard against these attacks and improve Intrusion Detection Systems. This will allow for faster and more accurate detection of threats, eventually helping

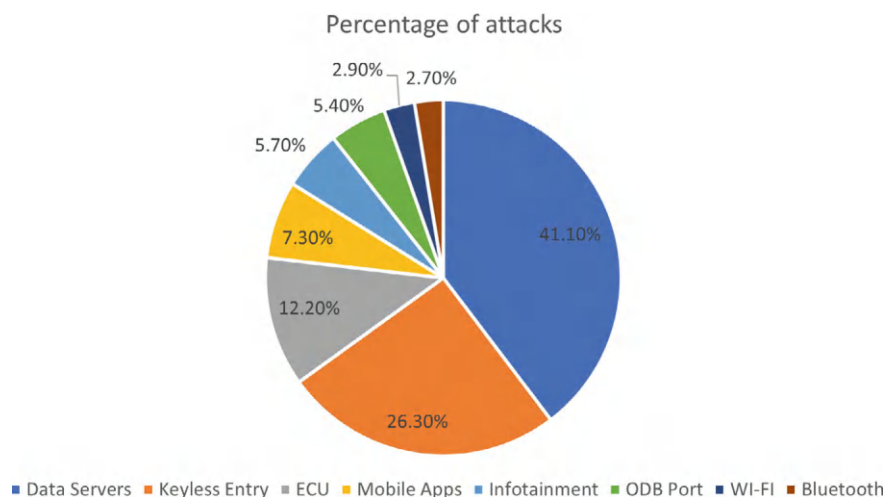


Fig. 1 IoV cyber threats

to eradicate them quicker, lessen the chances of a successful cyber-attack or intrusion in these systems, and keep people safe in real-time problems by identifying the best models for IDS in IoV.

The remaining sections of the paper are structured as follows: Sect. 1 discusses the associated research. The past work and objectives are discussed in Sect. 2. Section 3 describes the methodology and dataset used in this research work. Section 4 describes the results and discussions, while Sect. 5 covers the conclusion and future scope.

2 Related Work

A similar project was carried out for “Machine Learning-Driven Optimization for Intrusion Detection in Smart Vehicular Networks by Ayoub Alsarhan, Abdel-Rahman Al-Ghuwairi, Islam, Almalkawi, and Mohammad Alauthman at Hashemite University [1] regarding intrusion detection in innovative vehicular network. Another paper compares the most popular intrusion detection models that use CNN to RNN and LSTM and observes how efficient these models are to protect against various cyber-attacks in a vehicular network [2, 3]. The main idea is to turn our dataset into spatial images and then use multiple CNN models and, at the same time, input the raw tabular data set into RNN and LSTM; the output of all these models is then taken into consideration, and we try to find the best working model out of all.

The paper titled “Transfer learning and CNN optimized IDS for IoV” by Yang and Shami [4]. This paper [4] discusses applying and implementing various CNN models for Intrusion Detection in the Internet of Vehicles using images as input. This provides us with the in-depth implementation of CNN in IDS. CNN model was observed to give high accuracy, but other models were not compared during implementation [5, 6]. The dataset did not have many attributes, an essential requirement, or an implementation case for CNN. The same dataset is then utilized for RNN and LSTM models to provide a uniform comparison.

Song et al. [7] proposed an Intrusion Detection System (IDS) for in-vehicle networks using Deep Convolutional Neural Networks (CNNs) to detect cyberattacks in Controller Area Network (CAN) messages [7]. Their method converts CAN data into image-like representations for CNN analysis and automates feature extraction, eliminating the need for manual engineering. Outperforming conventional machine learning techniques, the IDS showed excellent detection accuracy across a range of attack types, including DoS and spoofing [8]. The study addresses scalability and computational limitations while highlighting the possibility of real-time deployment in automobiles.

Research Gap

The caliber and variety of training datasets significantly impact IDS efficacy. Comprehensive, publicly accessible datasets covering various attack scenarios and typical driving conditions are complex. The creation and assessment of reliable IDS models

that can be generalized across multiple vehicular environments are hampered by this limitation.

Real-time detection capabilities with low latency are essential for IDS implementation in actual vehicular systems. However, this isn't easy to achieve because vehicles have limited computational resources. Research on striking a balance between computational efficiency and detection accuracy is still ongoing. Performance evaluation is inconsistent in IoV environments due to the lack of standardized evaluation metrics and benchmarking frameworks for IDS. To objectively compare various IDS approaches and Deep Learning models, it is imperative to establish generally recognized benchmarks.

The objectives of this research work are:

- To develop a machine learning based Intrusion Detection System for accurate time threat detection in the Internet of Vehicles.
- Using different models (CNN, RNN, LSTM) to improve the knowledge of how each model works for IDS.
- Comparing these models to obtain the best model with higher accuracy, F1 score, precision, and recall, making it the best fit for implementation.
- To provide people with a sense of security and safeguard the network of vehicles globally.

3 Methodology

The models use one common dataset as input (CICIDS2017) [4]; for CNN models, this dataset is first converted into spatial images and then used for four individual models (Xception, Inception, VGG16, VGG19). The spatial images are generated pixel-wise by using a set of 3 attributes at a time representing the RGB values; a total of 81 attributes being present makes it (9×9 images).

After the CNN implementation, the other two models, namely RNN and LSTM, utilize the dataset directly as a time-series dataset, and pre-trained models are used here to work on this dataset. The pre-trained RNN uses the 'real' activation function for non-linearity and includes a drop-out layer that drops a specific percentage of data to overcome the overfitting problem. A fully connected (dense) layer with 64 neurons extracts complex patterns in the features, which RNN learns; another dropout layer is used, and then a final output layer is used with the SoftMax function for multi-class classification. Hence, this RNN-based multi-class classification model has one RNN layer for sequence learning, fully connected layers for feature transformation, Dropout for regularization, and A final SoftMax layer for class probabilities.

LSTM model is similarly worked on, but it uses hyperbolic tangent as an activation function, and it also uses an additional batch normalization of LSTM layers after dropout layers to stabilize training; apart from that 2, LSTM layers are present to capture temporal dependencies, ends the same way as RNN with dense layers and SoftMax activation for class probabilities. The output for all three models is a table

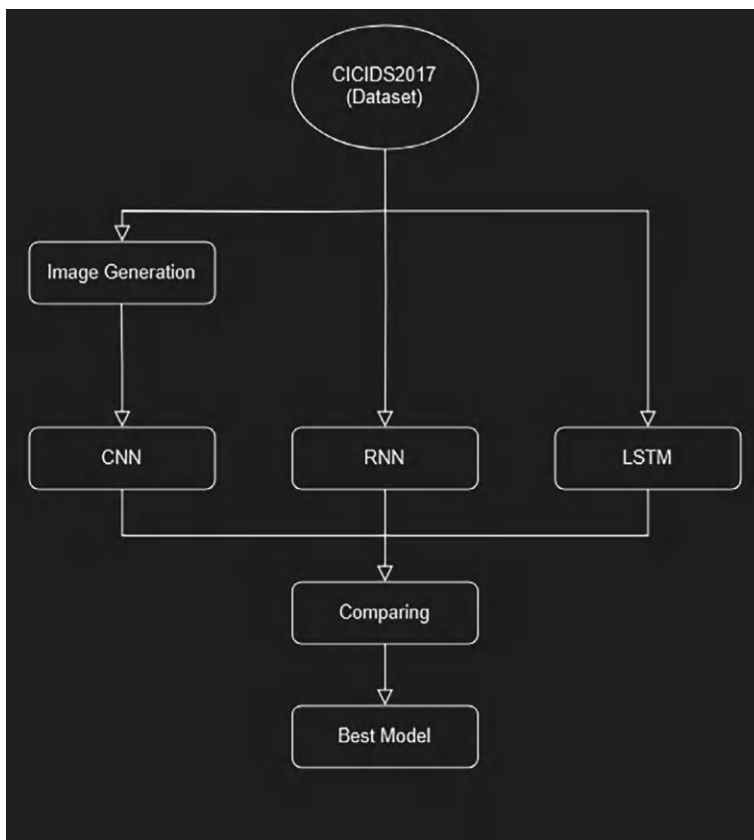


Fig. 2 Flow chart of the proposed work

to provide their accuracy, precision, recall, and F1 score. This is used to compare them on graphs and manually and provide us with the best possible model for this problem, as shown in Fig. 2.

3.1 Architecture of the CNN Model Used

4 CNN models were utilized in CNN implementation:

- **VGG16**

VGG16 has 13 convolutional layers, three fully connected layers, five pooling layers, and 3×3 convolution kernels. It uses a ReLU activation function [3].

- **VGG19**

VGG-19 is a deep convolutional neural network with 19 weight layers, comprising 16 convolutional layers and three fully connected layers. The architecture follows a straightforward and repetitive pattern, making it easier to understand and implement [3].

- **Xception**

Xception is an expansion of the Inception architecture that uses depth-wise separable convolutions to replace typical Inception modules, reducing computational complexity and improving performance [9, 10].

- **Inception**

It is a cryptographic protocol for securely operating network services over an unsecured network. Typical applications include remote command-line login and remote [11].

Once the model is trained, it is essential to evaluate and compare its performance in terms of accuracy. These are:

- **Precision:** The model can be precise with the positives it gives, i.e., the total number of true positives by the total number of positives, which provides precision.
- **Recall:** Also known as sensitivity and actual positive rate, which helps us identify how many of the actual positives the model correctly identified. The only difference is that the calculation formula uses false negatives instead of false positives.
- **F1 score:** It is the harmonic mean of both precision and recall, which helps us provide a balance between them. So, if the f1 score is high, it represents high precision and recall.
- **Support:** This refers to the number of occurrences of each class in the dataset. Telling us about the instances present in a test set at a time.

CNN Inputs

As shown in Fig. 3, we have various spatial images produced for the CNN models, which are then used as input to train these models.

LSTM

In the Internet of Vehicles (IoV), Long Short-Term Memory (LSTM) networks—a type of Recurrent Neural Networks (RNNs)—are beneficial for Intrusion Detection Systems (IDS). Because LSTMs solve the vanishing gradient issue that traditional RNNs have, they are excellent at identifying long-term dependencies in sequential data, like CAN messages. This enables them to recognize complex attack patterns, such as DoS, fuzzy, and spoofing attacks over long periods. LSTMs improve the accuracy and dependability of IDS by examining contextual patterns and temporal correlations in vehicular network traffic, providing a strong defense against changing cyber threats in IoV environments.

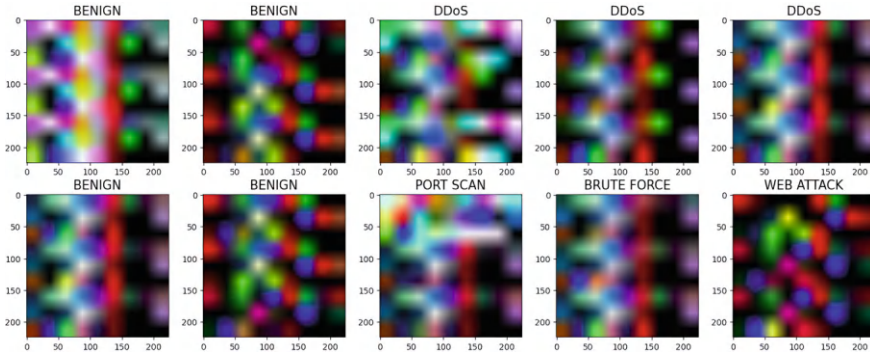


Fig. 3 CNN spatial images

RNN

Because recurrent neural networks (RNNs) can analyze sequential data, such as Controller Area Network (CAN) messages, they are very effective for Intrusion Detection Systems (IDS) in the Internet of Vehicles (IoV). RNNs help identify intricate attack patterns like spoofing, DoS, and fuzzy attacks because they can identify temporal dependencies in network traffic. By examining message sequences over time, their sequential modeling aids in distinguishing between benign and malevolent behavior. Furthermore, RNN variants like LSTMs and GRUs resolve the vanishing gradient issue, making it possible to learn from lengthy sequences effectively. RNNs are a valuable tool for improving IoV cybersecurity because of these features.

3.2 Dataset Description

This research work incorporates the usage of two datasets.

CIC-IDS2017

The dataset contains benign and the most up-to-date common attacks, which resemble real-world data (PCAPs). It also includes the results of the network traffic analysis using CICFlowMeter with labeled flows based on the time stamp, source, destination IPs, source and destination ports, protocols, and attack (CSV files).

(<https://www.unb.ca/cic/datasets/ids-2017.html>) [12] as shown in Fig. 4.

Car Hacking Dataset

Include DoS attack, fuzzy attack, spoofing the drive gear, and RPM gauge. Datasets were constructed by logging CAN traffic from an actual vehicle via the OBD-II port while message injection attacks were performed. Datasets contain 300 intrusions of message injection. Each intrusion is performed for 3–5 s, and each dataset has 30–40 min of CAN traffic.

	Destination Port	Flow Duration	Total Fwd Packets	Total Backward Packets	Total Length of Fwd Packets	Total Length of Bwd Packets	Fwd Packet Length Max	Fwd Packet Length Min	Fwd Packet Length Mean	Fwd Packet Length Std
count	225775.000000	225775.000000	225775.000000	225775.000000	225775.000000	225775.000000	225775.000000	225775.000000	225775.000000	225775.000000
mean	126.844938	126.559528	125.119742	118.783985	128.654428	118.763614	128.454081	93.085731	128.309513	97.670671
std	64.925957	73.871437	76.393236	82.809514	72.754323	84.519541	70.873170	98.948680	71.967723	98.423816
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	113.205706	62.863188	52.710210	72.620120	94.699700	0.000000	46.456456	0.000000	47.094595	0.000000
50%	113.205706	125.522998	104.399399	128.521021	139.369369	128.265766	133.115616	0.000000	145.367868	129.286787
75%	113.205706	190.391536	186.463964	166.043544	194.376877	181.869369	191.569069	173.063063	190.037538	185.442943
max	255.000000	255.000000	255.000000	255.000000	255.000000	255.000000	255.000000	255.000000	255.000000	255.000000

8 rows × 11 columns

Fig. 4 CIC-IDS2017 dataset table

(<https://ocslab.hksecurity.net/Datasets/CAN-intrusion-dataset>) [13].

4 Results and Discussion

The conclusion derived from the performance evaluation of CNN, RNN, and LSTM for the IDS is that CNN achieved high accuracy and low false favorable rates compared to its counterparts, i.e., RNN and LSTM. RNN, even though its focus on past data, was met with the challenge of vanishing gradient during the training phase. Similarly, LSTM worked better than RNN, helping in long-term dependencies, but required more computational power. Overall, CNN proved to be the most effective among the three in static data scenarios, and LSTM works well in dynamic ones. Using the CICIDS2017 dataset [12, 13], metrics like accuracy, precision, recall, and F1-score were compared, and CNN proved to be the best performing.

4.1 Comparative Analysis Result

The comparison of RNN and LSTM models based on the overall accuracy and loss is done graphically, as shown in Fig. 5.

The overall comparative analysis of all three models (CNN, RNN, LSTM) based on precision, recall, F1 score, and support is shown in detail in Table 1.

Compared to baseline IDS approaches, such as RNN and LSTM, the proposed framework reduces false positives while maintaining high detection accuracy and provides a more balanced decision-making framework, as shown by comparing traditional voting mechanisms alone.

LSTM

The individual stats for the LSTM model include the Confusion Matrix, as shown in Fig. 6 and the accuracy/loss comparison based on train and validation accuracy, as shown in Fig. 7.

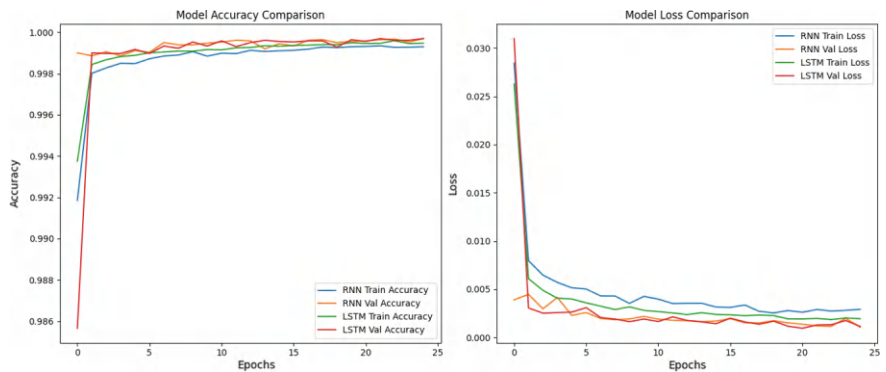


Fig. 5 Overall accuracy/loss comparison

Table 1 Comparative table analysis

Models		Precision	Recall	F1 score	Support
CNN	0	1.00	1.00	1.00	5052
	1	1.00	1.00	1.00	225
	2	1.00	1.00	1.00	200
	3	1.00	1.00	1.00	197
	4	1.00	1.00	1.00	171
	Macro avg	1.00	1.00	1.00	5845
RNN	0	1.00	1.00	1.00	19537
	1	0.50	0.50	0.50	2
	2	1.00	1.00	1.00	25605
	3	0.67	1.00	0.80	2
	4	1.00	0.50	0.67	2
	Macro avg	0.83	0.80	0.79	45148
LSTM	0	1.00	1.00	1.00	19537
	1	0.00	0.00	0.00	2
	2	1.00	1.00	1.00	25605
	3	1.00	1.00	1.00	2
	4	0.25	0.50	0.33	2
	Macro avg	0.65	0.70	0.67	45148

Cross-Validation Results:

Mean Accuracy: 0.9986.

Standard Deviation of Accuracy: 0.0003.

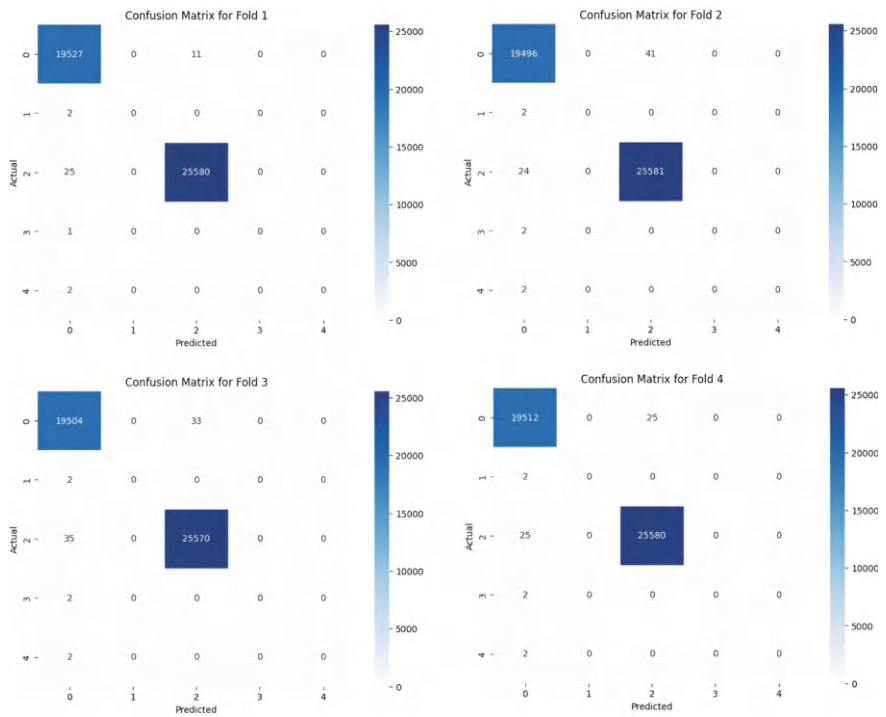


Fig. 6 LSTM confusion matrix

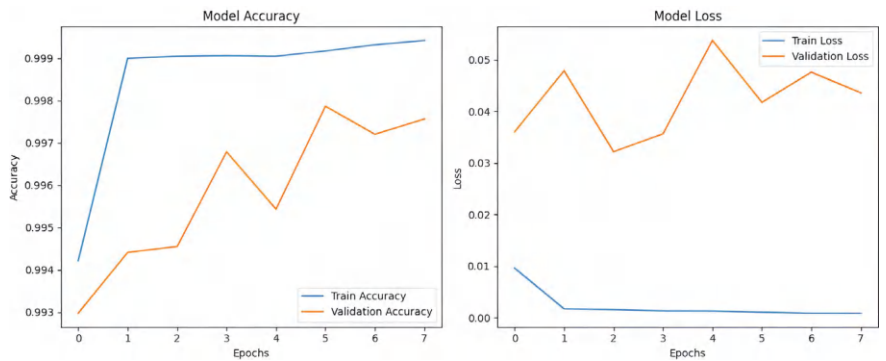


Fig. 7 LSTM accuracy/loss comparison

RNN

The individual stats for the RNN model, including the accuracy/loss comparison based on train accuracy and validation accuracy, are shown in Fig. 8, and the classification report is shown in Fig. 9.

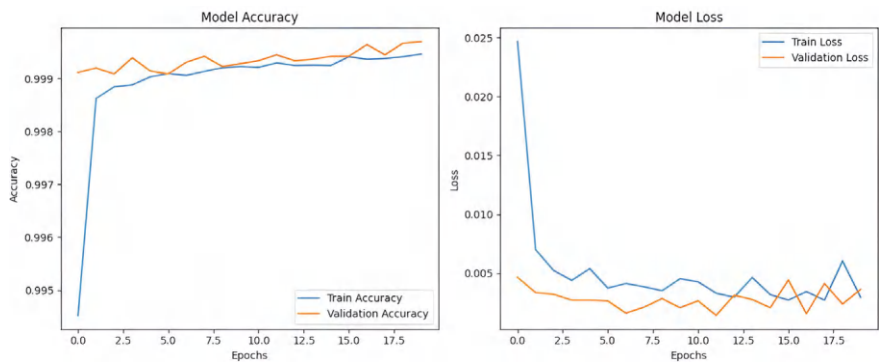


Fig. 8 RNN accuracy/loss comparison

Classification Report:				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	19537
1	0.50	0.50	0.50	2
2	1.00	1.00	1.00	25605
3	0.67	1.00	0.80	2
4	1.00	0.50	0.67	2
accuracy			1.00	45148
macro avg	0.83	0.80	0.79	45148
weighted avg	1.00	1.00	1.00	45148
Accuracy: 1.00				

Fig. 9 RNN classification report

5 Conclusion

The study shows that CNN-based intrusion detection systems are good at spotting threats accurately and keeping false alarms low, especially when working with stable data in-car networks. LSTM models, on the other hand, do a great job handling changing environments since they can look at patterns over time. Future works—

Adaptive Models: Focus on developing more innovative and flexible systems that can learn and adapt to changes in vehicular networks in real time. These models should be capable of handling dynamic environments, such as changing traffic patterns or new types of cyber threats.

More enormous Datasets: Increase the variety and diversity of datasets training these models. This can include collecting data from different regions, network setups,

and attack scenarios. Real-Time Improvements: Work on optimizing detection and response times to ensure the system can instantly identify and counter threats.

References

1. Alsarhan, A., et al.: Machine learning-driven optimization for intrusion detection in smart vehicular networks. *Veh. Syst. J.* (2021)
2. Cui, J., Long, J., Min, E., Liu, Q., Li, Q.: Comparative study of CNN and RNN for deep learning based intrusion detection system. In: 4th International Conference, ICS 2018, Haikou, China, June 8–10, 2018, Part V (2018). https://doi.org/10.1007/978-3-030-00018-9_15
3. Brahma, B., Bhuyan, H.K.: Soft computing and machine learning techniques for e-Health data analytics. In: Mishra, S., González-Briones, A., Bhoi, A.K., Mallick, P.K., Corchado, J.M. (eds.) *Connected e-Health. Studies in Computational Intelligence*, vol. 1021. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97929-4_4
4. Yang, L., Shami, A.: A Transfer Learning and Optimized CNN Intrusion Detection System for the Internet of Vehicles, pp. 2774–2779 (2022). <https://doi.org/10.1109/ICC45855.2022.9838780>
5. Brahma, B., et al.: Mathematical model for analysis of COVID-19 outbreak using VON Bertalanffy Growth Function (VBGF). *Turk. J. Comput. Math. Educ. (TURCOMAT)* **12**(11), 6063–6075 (2021). <https://doi.org/10.17762/turcomat.v12i11.6925>
6. Yang, L., Shami, A.: A lightweight concept drift detection and adaptation framework for IoT data streams. *IEEE Internet Things Mag.* **4**(2), 96–101 (2021)
7. Song, H.M., et al.: In-Vehicle Network Intrusion Detection Using Deep CNNs. *Vehicular Communications*, p. 21 (2020)
8. Large, J., Lines, J., Bagnall, A.: A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. *Data Min. Knowl. Discov.* **33**(6), 1674–1709 (2019)
9. Chollet, F.: Xception: Deep learning with depth-wise separable convolutions. In: *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 1800–1807 (2017)
10. Rahimzadeh, M., Attar, A.: A modified deep convolutional neural network for detecting COVID-19 and pneumonia from chest X-ray images based on the concatenation of Xception and ResNet50V2. *Inform. Med. Unlocked* **19**, 100360 (2020)
11. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 2818–2826 (2016)
12. IDS 2017. Datasets. Research. Canadian Institute for Cybersecurity. UNB. (n.d.): <https://www.unb.ca/cic/datasets/ids-2017.html>
13. HCRL-Car-Hacking Dataset. (n.d.): <https://ocslab.hksecurity.net/Datasets/CAN-intrusion-dataset>

Real-Time Network Intrusion Detection System Using Machine Learning



Sam Peter, J. L. Aravind, Feba Mariyam Jacob, Johann Varghese George, and Tessa Mathew

Abstract This paper proposes a novel Intrusion Detection and Prevention System (IDPS) that employs machine learning techniques to bolster network security. By leveraging labelled datasets such as CIC-IDS2017 and CIC-IDS IOT 2023, the system undergoes rigorous data preprocessing to extract meaningful features. A comprehensive ensemble of supervised learning models, including Random Forest, XGBoost, CNN, LSTM, KNN, and Model Stacking, is trained and evaluated for intrusion detection accuracy. Additionally, unsupervised clustering algorithms (K-Means, DBSCAN) are integrated to identify anomalous network traffic patterns. Experimental results demonstrate the efficacy of the proposed IDPS in detecting and preventing cyber threats, particularly within the evolving 5G ecosystem.

Keywords Intrusion-detection · IDPS · Cybersecurity · Computer networks · Anomaly detection

S. Peter (✉) · J. L. Aravind · F. M. Jacob · J. V. George · T. Mathew
Department of CSE, Mar Baselios College of Engineering and Technology (Autonomous),
Trivandrum, India
e-mail: sampeter.b21cs1153@mbcet.ac.in

J. L. Aravind
e-mail: aravindjl.b21cs1114@mbcet.ac.in

F. M. Jacob
e-mail: febamariamjacob.b21cs1123@mbcet.ac.in

J. V. George
e-mail: johannvarghesegeorge.b21cs1129@mbcet.ac.in

T. Mathew
e-mail: tessy.mathew@mbcet.ac.in

1 Introduction

In today's digital world, safeguarding information is crucial for businesses and individuals. Network security protects sensitive data from unauthorised access, modification, or disruption.

While traditional security measures offer some protection, they often struggle to defend against the ever-evolving tactics of cybercriminals. These attackers are becoming increasingly sophisticated, requiring more advanced security solutions. The emergence of 5G technology brings faster internet speeds and more excellent connectivity.

However, it also creates new opportunities for cyberattacks. With more devices connected to the network, the potential for breaches increases [1].

IDPS systems act as a shield, monitoring network activity for suspicious behaviour. They can detect and block cyberattacks, helping to maintain the integrity and confidentiality of information [2]. This research aims to develop an advanced IDPS capable of protecting networks, especially those using 5G technology [3], to create a system that can accurately identify and prevent a wide range of cyber threats.

2 Methodology

Data Collection and Preprocessing

The study utilised network traffic data from publicly accessible datasets, CICIDS2017 and CIC 2023 IoT [4], providing labelled instances of regular and malicious network activity. These datasets are widely used benchmarks in cybersecurity research, with diverse threat types and realistic traffic patterns.

Each dataset includes labelled examples for specific threat types, enabling detailed analysis and modelling of network intrusions.

The identified threat types for each dataset are:

(1) CICIDS2017:

- (a) BENIGN: Normal traffic with no malicious intent.
- (b) DoS (Denial of Service): Attacks aimed at making a network service unavailable by overwhelming it with traffic.
- (c) Port Scan: Probing of network ports to identify open ports and associated vulnerabilities.
- (d) Brute Force: Attempts to gain unauthorised access by systematically guessing login credentials.
- (e) Web Attack: Exploits targeting web servers, including SQL injection and cross-site scripting.
- (f) Bot: Malicious traffic from botnets for spam, data theft, or distributed denial-of-service (DDoS) attacks.

- (g) Infiltration: Unauthorized access and movement within a network, often aiming to exfiltrate data or disrupt operations.
- (2) CIC 2023 IOT:
 - (a) BENIGN: Normal network activity.
 - (b) DoS Hulk: A flood-based DoS attack using oversized requests to exhaust resources.
 - (c) Port Scan: Similar to CICIDS2017, probing open ports for vulnerabilities.
 - (d) DDoS (Distributed Denial of Service): Coordinated attacks from multiple sources, overwhelming a target system.
 - (e) DoS Golden Eye: HTTP-based DoS attacks targeting web servers.
 - (f) FTP-Patator & SSH-Patator: Brute-force attacks targeting FTP and SSH services, respectively.
 - (g) DoS Slowloris & Slowhttptest: Specialized DoS attacks that exploit web server connection management vulnerabilities.
 - (h) Web Attack: Including SQL injection, command injection, and other web-targeted exploits.
 - (i) Bot: Similar to CICIDS2017, encompassing botnet-generated malicious traffic.
 - (j) Infiltration: Unauthorized access within the network.
 - (k) Heartbleed: A vulnerability in OpenSSL that allows data theft from encrypted connections.

The initial dataset comprised approximately 500,000 rows, which were pre-sampled down to a manageable size of 56,000–60,000. This initial sampling involved stratified random selection to reduce data size while preserving the overall class distribution. To prepare the data, several steps were taken:

1. Cleaning: Removed corrupted or irrelevant data and replaced missing values with zeros.
2. Normalization: Standardized feature values using Z-score normalisation [5].
3. Sampling: Utilized KNN clustering to create a smaller, representative dataset [6].
 - (a) Minority Class Retention: All instances of minority classes were preserved to ensure adequate representation in the dataset.
4. Balancing: Employed SMOTE to address class imbalance between normal and malicious instances [7].

To reduce the number of features while preserving important information, the following methods were used:

- Information Gain: It ranks features based on their relevance in distinguishing between malicious and regular traffic. Features with low information gain were excluded to focus on the most predictive attributes [8].
- Fast Correlation Filter (FCBF): It identifies and removes redundant features by calculating their correlation with the target variable and among themselves. Features irrelevant to the target are discarded, while redundant ones—those providing overlapping information—are filtered out based on a predefined threshold [9, 10].

Information Gain and FCBF were applied, except for CNN and LSTM models, which used all features.

3 Machine Learning Models

Supervised Learning for Signature-Based Detection

The models implemented include Random Forest, Decision Tree [11], XGBoost, 1D-Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM). For Random Forest, Decision Tree, and XGBoost, both IG and IG with FCBF feature reductions are used as inputs. CNN and LSTM models are used without feature education.

XGBoost

XGBoost is a popular gradient-boosting algorithm known for its efficiency and accuracy in handling structured/tabular data. It's widely used in intrusion detection systems because it can detect anomalies and classify network traffic effectively [12].

Decision Tree

A Decision Tree is a type of supervised learning method used for classification. It divides the data into subsets based on the most important features, creating a tree-like structure of decisions. Decision Trees are simple to understand and interpret, making them useful for identifying patterns in network traffic for intrusion detection [13, 14].

Random Forest

Random Forest is an ensemble learning method that builds several decision trees during training and uses the most frequent class for classification tasks. This method is resistant to overfitting and achieves high accuracy by combining the predictions of multiple trees. In intrusion detection, it can efficiently handle many features and detect complex patterns in network traffic [15].

Stacking

Stacking is a technique in ensemble learning that improves prediction performance by combining multiple classification models. It uses base models to make predictions and a meta-model to combine them, enhancing the overall model's accuracy and robustness. The Base Models include a Decision Tree, Random Forest and XGBoost, while the Meta-Model uses an XGBoost Classifier. It is utilised as the meta-model to combine predictions from the base models.

Using Hyperopt for hyperparameter tuning helps to find the best settings for the XGBoost meta-model in a stacking ensemble. This process optimises the ensemble

by combining the strengths of Decision Tree, Random Forest, and XGBoost models, leading to better accuracy in detecting network intrusions.

Convolutional Neural Network (1D-CNN)

CNNs can automatically learn spatial hierarchies of features from network traffic data, improving the detection of complex patterns and anomalies that indicate intrusions.

LSTM

LSTM models are best for analysing time-series data, such as network traffic logs, to identify suspicious activities or anomalies.

Unsupervised Learning for Anomaly-Based Detection

Unsupervised learning models are utilised for anomaly-based detection. Clustering algorithms such as K-Means and DBSCAN are executed to distinguish and identify anomalies in network activity. This approach classifies information into malicious unknown packets or harmless packets and checks the certainty of the anomaly-based model to verify incorrect identifications.

K Means Clustering

The K-means clustering algorithm groups the training data into clusters and then uses these clusters to assign labels to the test data. The model's accuracy is measured by comparing these predicted labels with the actual labels of the test data.

Hyperparameter Tuning Using BO-GP

BO-GP (Bayesian Optimization with Gaussian Processes) uses the principles of Bayesian inference and Gaussian processes to find the best hyperparameters efficiently. Bayesian inference continuously updates the probability estimate for a hypothesis as new evidence or information becomes available. BO-GP updates the belief about the objective function as new evaluations are performed. Gaussian processes are statistical models for observations in a continuous domain, like time or space. They define a distribution over functions and provide a probabilistic approach to learning these functions.

Hyperparameter Tuning Using BO-TPE

Tree-structured Parzen Estimator (TPE) is a Bayesian Optimization method that models the objective function differently from Gaussian Processes. TPE employs a tree-based approach to model the distribution of the objective function for hyperparameter optimisation [16].

TPE uses two distributions to model the objective function: one for hyperparameters expected to perform well (good hyperparameters) and another for hyperparameters expected to perform poorly (bad hyperparameters). The algorithm selects the next set of hyperparameters to evaluate by sampling from these two distributions. This approach balances exploration (trying new, potentially good hyperparameters) and exploitation (focusing on hyperparameters known to perform well).

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an unsupervised machine learning algorithm designed for clustering data points, which is handy for large spatial datasets. It excels in identifying clusters of varying shapes and distinguishing outliers as noise.

DBSCAN forms clusters by evaluating the density of data points in the dataset. It groups densely packed points and designates points in low-density regions as outliers.

Performance Metrics (Supervised Learning)

(a) *Accuracy*

Accuracy measures the number of correct predictions out of the total instances.

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

(b) *Precision*

Precision measures the number of accurate positive predictions out of the total optimistic predictions made by the model.

$$Precision = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Positives})}$$

(c) *Recall*

Recall measures the number of accurate optimistic predictions from the total positive instances in the dataset.

$$Recall = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Negatives})}$$

(d) *F1-Score*

F1-score is the harmonic mean of precision and recall. It provides a balance between the two metrics.

$$F1 - score = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

(e) *Macro Average*

Macro average calculates the metrics independently for each class and then takes the unweighted mean of the scores.

$$Macro\ Average = \frac{\sum_{i=1}^n Precision_i}{n}$$

‘n’ is the number of classes.

(f) *Weighted Average*

Weighted average calculates the metrics for each class but considers the support (the number of actual instances for each label).

$$\text{Weighted Average} = \frac{\sum_{i=0}^n \text{Precision}_i \times \text{Support}_i}{\text{Total Support}}$$

‘Support’ gives the number of actual instances for class i , and ‘Total Support’ gives the sum of support across all classes.

4 System Design

Signature Based Detection

The signature-based model outlined in Fig. 1 follows a supervised learning approach akin to MTH-IDS’s Tier 1 [16] supervised model development. Attack patterns are matched using tree-based classifiers (XGBoost, Random Forest, Decision Tree) with hyperparameter optimisation via Bayesian Optimization (BO-TPE). However, this modification extends the ensemble stacking methodology, enhancing detection confidence and reducing false positives.

Anomaly Based Detection

The anomaly-based model outlined in Fig. 2 integrates unsupervised clustering (DBSCAN, K-Means) with a SMOTE-enhanced preprocessing layer.

- The clustering-based labeling method remains inspired by MTH-IDS’s Tier 3 (unsupervised model development) [17] but introduces confidence thresholds and adaptive feature extraction.
- Instead of relying solely on biased classifiers for false positives (as in MTH-IDS Tier 4), [17] this model refines the classification step by leveraging adaptive decision thresholds based on detected anomalies.

5 Model Performance

Supervised Learning

a. With FCBF and IG

XGBoost consistently emerges as the best-performing model (as per Table 1), both pre- and post-tuning. Before tuning, it achieves an Accuracy of 0.985, Precision of 0.990, Recall of 0.990, and F1-Score of 0.987. After tuning, these metrics improve

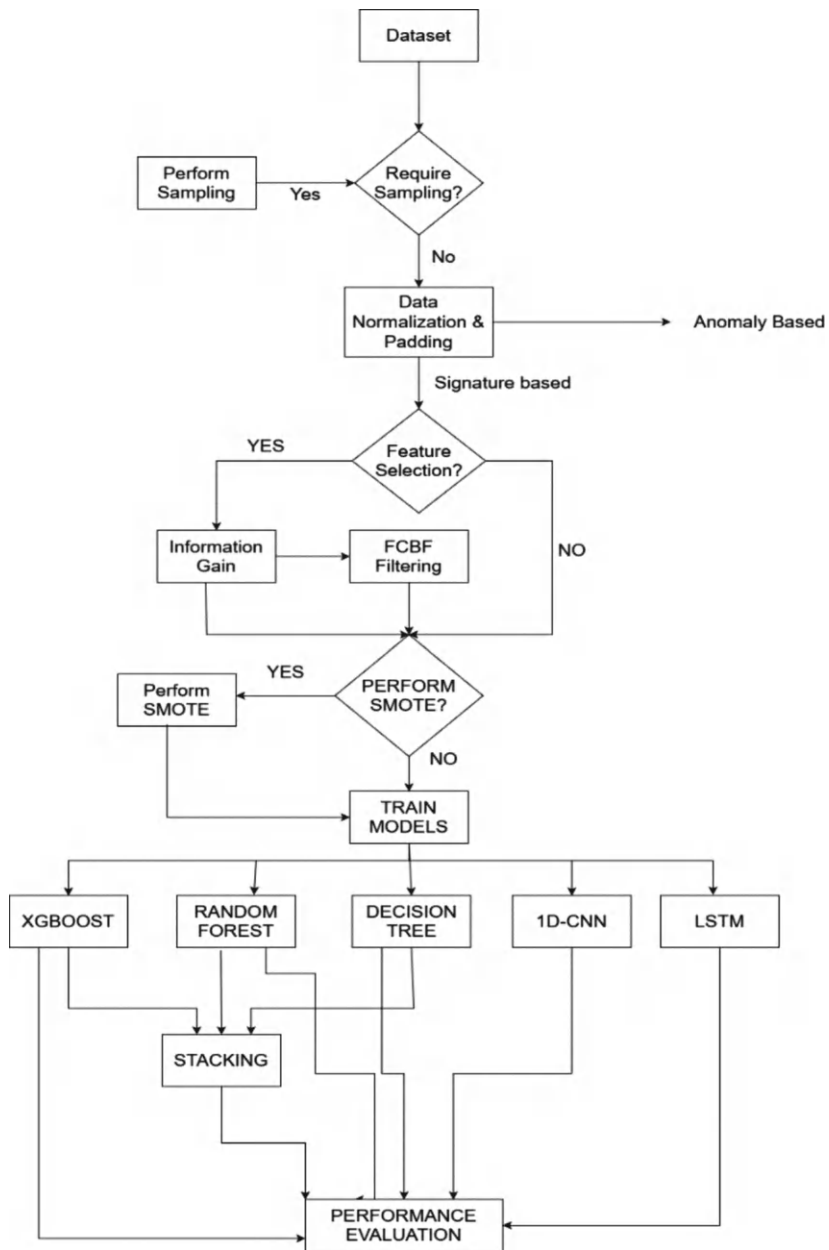
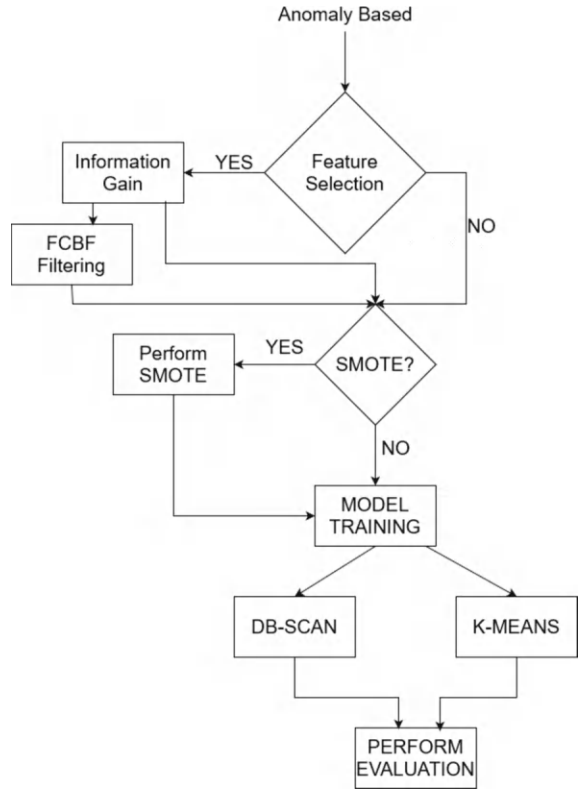


Fig. 1 Signature based detection

Fig. 2 Anomaly based detection



to an Accuracy of 0.994, Precision of 0.995, Recall of 0.995, and F1-Score of 0.994, showcasing its strong optimization potential. Random Forest follows closely, with its Accuracy increasing from 0.994 pre-tuning to 0.994 post-tuning, while its F1-Score improves from 0.994 to 0.994, reflecting marginal enhancement due to tuning.

The Stacking Ensemble demonstrates consistent high performance, achieving Accuracy and F1-Score values of 0.993 both before and after tuning, highlighting its reliability even without additional optimization. In contrast, Decision Tree shows limited improvement, with Accuracy remaining stable at 0.993, and its F1-Score consistent at 0.993, indicating relatively modest sensitivity to hyperparameter tuning.

b. Without FCBF

XGBoost consistently emerges as the top-performing model, both before and after hyperparameter tuning (as per Table 2). Before tuning, it achieves an Accuracy of 0.987, Precision and Recall of 0.988, and F1-Score of 0.987. After tuning, these metrics improve to 0.996 for Precision, Recall, and F1-Score, and Accuracy reaches 0.996, showcasing its strong optimization potential. Random Forest follows closely, with marginal improvement after tuning. Decision Tree shows limited improvement

with tuning. Stacking Ensemble maintains consistent high performance, both before and after tuning.

The combination of FCBF and SMOTE contributes to the strong performance of all models. FCBF helps in selecting relevant features, while SMOTE addresses the class imbalance issue, improving the overall model performance.

c. With IG

Hyperparameter tuning greatly improves the performance of the model, especially for XGBoost, which comes out as the best model both before and after tuning (as per Table 3). Before tuning, XGBoost already performs very well with an Accuracy of 0.987, Precision and Recall of 0.987, and F1-Score of 0.987. After tuning, its performance improves dramatically with Accuracy increasing to 0.996, Precision to 0.997, Recall to 0.996, and F1-Score to 0.996, showing great optimization potential. Random Forest also exhibits marked improvement, with Accuracy being increased from 0.99 to 0.994 and F1-Score similarly increased from 0.99 to 0.994, which places it in the second position of the best model overall.

Stacking Ensemble has performed consistently well throughout the exercise, with values of both Accuracy and F1-Score being at 0.994 pre- and post-tuning, indicating stability. Decision Tree, though competitive, only exhibited minor increases in its Accuracy and F1-Score, increasing from 0.993 to 0.994, which limits the optimization flexibility.

d. FCBF + IG Without SMOTE

XGBoost consistently emerges as the best-performing model, both pre- and post-tuning (as per Table 4). Before tuning, it achieves an Accuracy of 0.990, Precision and Recall of 0.990, and F1-Score of 0.990. After tuning, these metrics improve to 0.995 for Precision, Recall, and F1-Score, and Accuracy also reaches 0.995, showcasing its strong optimization potential. Random Forest follows closely, with its Accuracy increasing from 0.995 pre-tuning to 0.995 post-tuning, while its F1-Score improves from 0.995 to 0.995, reflecting marginal enhancement due to tuning.

The Stacking Ensemble demonstrates consistent high performance, achieving Accuracy and F1-Score values of 0.994 both before and after tuning, highlighting its reliability even without additional optimization. In contrast, Decision Tree shows limited improvement, with Accuracy increasing slightly from 0.993 to 0.994, and its F1-Score remaining stable at 0.994, indicating relatively modest sensitivity to hyperparameter tuning.

e. Without Feature Selection

1D-CNN demonstrates strong performance before tuning, achieving an Accuracy of 0.981, Precision of 0.982, Recall of 0.981, and F1-Score of 0.981 (as per Table 5). However, after tuning, these metrics decrease slightly, with Accuracy dropping to 0.974, Precision to 0.975, Recall to 0.974, and F1-Score to 0.974. This indicates a sensitivity to hyperparameter tuning, where the optimization led to marginally reduced performance. Macro Average and Weighted Average of F1-Score also decrease from 0.98 to 0.97 after tuning.

Table 2 Model performance without FCBF, with SMOTE

Parameter	XGBoost		Random forest		Decision tree		Stacking	
	Before hyperparameter tuning	After hyperparameter tuning	Before hyperparameter tuning	After hyperparameter tuning	Before hyperparameter tuning	After hyperparameter tuning	Before hyperparameter tuning	After hyperparameter tuning
Accuracy	0.987	0.996	0.994	0.99	0.994	0.993	0.994	0.994
Precision	0.988	0.997	0.994	0.996	0.994	0.993	0.994	0.994
Recall	0.987	0.996	0.994	0.996	0.994	0.993	0.994	0.994
F1-score	0.987	0.996	0.994	0.996	0.994	0.993	0.994	0.994
Macro average of F1-score	0.93	0.97	0.97	0.95	0.92	0.95	0.95	0.95
Weighted average of F1-score	0.99	1.00	0.99	1.00	0.99	0.99	0.99	0.99

Table 3 Model performance with information gain, with SMOTE

Parameter	XGBoost		Random forest		Decision tree		Stacking	
	Before hyperparameter tuning	After hyperparameter tuning	Before hyperparameter tuning	After hyperparameter tuning	Before hyperparameter tuning	After hyperparameter tuning	Before hyperparameter tuning	After hyperparameter tuning
Accuracy	0.987	0.996	0.994	0.99	0.994	0.993	0.994	0.994
Precision	0.988	0.997	0.994	0.996	0.994	0.993	0.994	0.994
Recall	0.987	0.996	0.994	0.996	0.994	0.993	0.994	0.994
F1-score	0.987	0.996	0.994	0.996	0.994	0.993	0.994	0.994
Macro average of F1-score	0.93	0.97	0.97	0.95	0.92	0.95	0.95	0.95
Weighted average of F1-score	0.99	1.00	0.99	1.00	0.99	0.99	0.99	0.99

Table 4 Model performance with FCBF and IG, without SMOTE

Parameter	XGBoost		Random forest		Decision tree		Stacking	
	Before hyperparameter tuning	After hyperparameter tuning	Before hyperparameter tuning	After hyperparameter tuning	Before hyperparameter tuning	After hyperparameter tuning	Before hyperparameter tuning	After hyperparameter tuning
Accuracy	0.990	0.995	0.995	0.993	0.993	0.994	0.994	0.994
Precision	0.990	0.995	0.995	0.993	0.993	0.994	0.994	0.994
Recall	0.990	0.995	0.995	0.993	0.993	0.994	0.994	0.994
F1-score	0.990	0.995	0.995	0.993	0.993	0.994	0.994	0.994
Macro average of F1-score	0.93	0.97	0.97	0.95	0.93	0.95	0.93	0.93
Weighted average of F1-score	0.99	1.00	1.00	0.99	0.99	0.99	0.99	0.99

Table 5 Model performance without feature selection

Parameter	1D-CNN		LSTM	
	Before hyperparameter tuning	After hyperparameter tuning	Before hyperparameter tuning	After hyperparameter tuning
Accuracy	0.981	0.974	0.981	0.981
Precision	0.982	0.975	0.981	0.981
Recall	0.981	0.974	0.981	0.981
F1-score	0.981	0.974	0.980	0.981
Macro average of F1-score	0.98	0.97	0.98	0.98
Weighted average of F1-score	0.98	0.97	0.98	0.98

LSTM maintains consistent performance throughout, achieving an Accuracy of 0.981, Precision of 0.981, Recall of 0.981, and F1-Score of 0.980 both before and after hyperparameter tuning. Its Macro Average and Weighted Average of F1-Score remain steady at 0.98, showcasing the model’s robustness to tuning and its ability to deliver reliable performance without feature filtering.

Unsupervised Learning

(a) With FCBF + IG, with SMOTE

K-Means Clustering outperforms DBSCAN, both pre- and post-tuning (as per Table 6). Before tuning, K-Means yields an Accuracy of 0.80, Precision of 0.84, Recall of 0.80, and F1-Score of 0.79. After tuning with both BO-GP and BO-TPE, the metrics were improved to an Accuracy of 0.85, Precision of 0.85/0.86, Recall of 0.85, and F1-Score of 0.85, with which it showed a strong effect of hyperparameter tuning.

Table 6 Model performance with FCBF + IG, with SMOTE

Parameter	K Means clustering			DBSCAN	
	Before hyperparameter tuning	After hyperparameter tuning (BO-GP)	After hyperparameter tuning (BO-TPE)	Before hyperparameter tuning	After hyperparameter tuning
Accuracy	0.80	0.85	0.85	0.39	0.40
Precision	0.84	0.85	0.86	0.42	0.16
Recall	0.80	0.85	0.85	0.39	0.40
F1-score	0.79	0.85	0.85	0.30	0.23
Macro average of F1-score	0.77	0.84	0.84	0.34	0.29

However, DBSCAN fails to attain comparable results and attains lower values for all of the evaluated metrics. The hyperparameter tuning done by BO-TPE, while slightly boosting its score, still failed to put it ahead of K-Means.

The combination of FCBF + IG and SMOTE seems to be effective in improving the performance of both algorithms, especially for K-Means.

(b) **With FCBF + IG, Without SMOTE**

K-Means Clustering performs better than DBSCAN, both before and after tuning (as per Table 7). Before tuning, K-Means achieves an Accuracy of 0.79, Precision of 0.84, Recall of 0.79, and F1-Score of 0.77. After tuning with both BO-GP and BO-TPE, the metrics improved to an Accuracy of 0.85/0.83, Precision of 0.85/0.84, Recall of 0.85/0.83, and F1-Score of 0.85/0.83, with which it presented a strong effect of hyperparameter tuning.

However, DBSCAN could not achieve comparable results and achieved lower values for all the above metrics. The hyperparameter tuning done by BO-TPE, although slightly boosting its score, still failed to position it ahead of K-Means.

The FCBF + IG feature selection appears to be effective in improving the performance of both algorithms, especially for K-Means.

(c) **FCBF, with SMOTE**

K-Means Clustering outperformed DBSCAN, both before and after tuning (as per Table 8). Before tuning, K-Means achieved an Accuracy of 0.77, Precision of 0.83, Recall of 0.77, and F1-Score of 0.75. After tuning using both BO-GP and BO-TPE, the metrics became an Accuracy of 0.89/0.83, Precision of 0.89/0.83, Recall of 0.89/0.83, and F1-Score of 0.89/0.83, with which it presented a strong effect of hyperparameter tuning.

DBSCAN is unable to reach comparable performance; its scores are lower for all metrics. Even after hyperparameter tuning with BO-TPE, it still does not match

Table 7 Model performance with FCBF + IG, without SMOTE

Parameter	K Means clustering			DBSCAN	
	Before hyperparameter tuning	After hyperparameter tuning (BO-GP)	After hyperparameter tuning (BO-TPE)	Before hyperparameter tuning	After hyperparameter tuning
Accuracy	0.79	0.85	0.83	0.39	0.40
Precision	0.84	0.85	0.84	0.42	0.16
Recall	0.79	0.85	0.83	0.39	0.40
F1-score	0.77	0.85	0.83	0.30	0.23
Macro average of F1-score	0.75	0.84	0.82	0.34	0.29
Weighted average of F1-score	0.77	0.85	0.83	0.30	0.23

Table 8 Model performance with FCBF, with SMOTE

Parameter	K Means clustering			DBSCAN	
	Before hyperparameter tuning	After hyperparameter tuning (BO-GP)	After hyperparameter tuning (BO-TPE)	Before hyperparameter tuning	After hyperparameter tuning
Accuracy	0.77	0.89	0.83	0.38	0.40
Precision	0.83	0.89	0.83	0.23	0.16
Recall	0.77	0.89	0.83	0.38	0.40
F1-score	0.75	0.89	0.83	0.24	0.23
Macro average of F1-score	0.72	0.88	0.82	0.29	0.29
Weighted average of F1-score	0.75	0.89	0.83	0.24	0.23

K-Means. It may be that DBSCAN is not well-suited for this dataset or needs more careful parameter tuning and exploration of different hyperparameter spaces.

The FCBF feature selection and SMOTE class imbalance handling seem to be effective in improving the performance of both algorithms, especially for K-Means.

(d) **FCBF, Without SMOTE**

K-Means Clustering outperforms DBSCAN, both before and after tuning (as per Table 9). Before tuning, K-Means achieves an Accuracy of 0.79, Precision of 0.84, Recall of 0.79, and F1-Score of 0.77. After tuning using both BO-GP and BO-TPE, the metrics improved to an Accuracy of 0.88/0.83, Precision of 0.89/0.83, Recall of 0.88/0.83, and F1-Score of 0.88/0.83, with which it presented a strong effect of hyperparameter tuning.

DBSCAN struggles to achieve comparable performance, with lower scores across all metrics. While hyperparameter tuning with BO-TPE slightly improves its performance, it still lags behind K-Means. It's possible that DBSCAN is not well-suited for this particular dataset or requires more careful parameter tuning and exploration of different hyperparameter spaces.

The FCBF feature selection appears to be effective in improving the performance of both algorithms, especially for K-Means.

(e) **Without Feature Selection, with SMOTE**

K-Means Clustering outperforms DBSCAN, both before and after tuning (as per Table 10). Before tuning, K-Means achieves an Accuracy of 0.82, Precision of 0.82, Recall of 0.82, and F1-Score of 0.81. After tuning using both BO-GP and BO-TPE, the metrics improved to an Accuracy of 0.85/0.84, Precision of 0.85/0.84, Recall of 0.85/0.84, and F1-Score of 0.85/0.83, with which it presented a strong effect of hyperparameter tuning.

Table 9 Model performance with FCBF, without SMOTE

Parameter	K Means clustering			DBSCAN	
	Before hyperparameter tuning	After hyperparameter tuning (BO-GP)	After hyperparameter tuning (BO-TPE)	Before hyperparameter tuning	After hyperparameter tuning
Accuracy	0.79	0.88	0.83	0.38	0.40
Precision	0.84	0.89	0.83	0.29	0.16
Recall	0.79	0.88	0.83	0.38	0.40
F1-score	0.77	0.88	0.83	0.24	0.23
Macro average of F1-score	0.75	0.87	0.83	0.29	0.29
Weighted average of F1-score	0.77	0.88	0.83	0.24	0.23

Table 10 Model performance without feature selection, with SMOTE

Parameter	K Means clustering			DBSCAN	
	Before hyperparameter tuning	After hyperparameter tuning (BO-GP)	After hyperparameter tuning (BO-TPE)	Before hyperparameter tuning	After hyperparameter tuning
Accuracy	0.82	0.85	0.84	0.38	0.39
Precision	0.82	0.85	0.84	0.41	0.43
Recall	0.82	0.85	0.84	0.38	0.39
F1-score	0.81	0.85	0.83	0.32	0.36
Macro average of F1-score	0.80	0.85	0.83	0.34	0.37
Weighted average of F1-score	0.81	0.85	0.83	0.32	0.36

DBSCAN fails to have comparable performance and underperforms in all evaluation metrics. Although it was marginally improved with BO-TPE hyperparameter optimization, the performance still lagged behind K-Means. It might also be a case that DBSCAN does not suit well with the given dataset or is rather sensitive to careful hyperparameter tuning and extensive search for different hyperparameter spaces.

(f) Without Feature Selection, Without SMOTE

K-Means Clustering outperformed DBSCAN, both before and after tuning (as per Table 11). Before tuning, K-Means achieved an Accuracy of 0.68, Precision of 0.72, Recall of 0.68, and F1-Score of 0.68. After tuning using both BO-GP and BO-TPE,

the metrics improved to an Accuracy of 0.87, Precision of 0.87, Recall of 0.87, and F1-Score of 0.87, with which it presented a strong effect of hyperparameter tuning.

DBSCAN can not perform nearly as well; the scores on all metrics are worse. With BO-TPE hyperparameter tuning, its performance improves slightly but remains much worse than K-Means. It might be that DBSCAN simply does not perform well for this dataset or needs further tuning of parameters and the search of more appropriate hyperparameter spaces.

Performance Review

- (a) Feature Selection
- (b) The Fast Correlation-Based Filter (FCBF) method was employed to reduce the dimensionality of the feature space. This technique effectively decreased training time without compromising performance metrics (accuracy, precision, recall, F1-score).
- (c) Supervised Learning
- (d) XGBoost, Decision Trees, and Random Forests exhibited performance improvements through hyperparameter tuning. These models demonstrated enhanced predictive accuracy and efficiency.
- (e) In contrast, hyperparameter tuning had a minimal impact on the performance of LSTM and CNN models. This suggests that these deep learning models are inherently robust and may require different optimization strategies.
- (f) Unsupervised Learning
- (g) K-Means clustering achieved moderate accuracy (80%) with feature selection and SMOTE. However, hyperparameter optimization using Bayesian Optimization with Gaussian Processes (BO-GP) significantly boosted performance to approximately 89%. This indicates the potential of K-Means for anomaly detection when combined with appropriate techniques.

Table 11 Model performance without feature selection, without SMOTE

Parameter	K Means clustering			DBSCAN	
	Before hyperparameter tuning	After hyperparameter tuning (BO-GP)	After hyperparameter tuning (BO-TPE)	Before hyperparameter tuning	After hyperparameter tuning
Accuracy	0.68	0.87	0.87	0.38	0.40
Precision	0.72	0.87	0.87	0.41	0.16
Recall	0.68	0.87	0.87	0.38	0.40
F1-score	0.68	0.87	0.87	0.32	0.23
Macro average of F1-score	0.68	0.86	0.86	0.34	0.29
Weighted average of F1-score	0.68	0.87	0.87	0.32	0.23

- (h) DBSCAN consistently underperformed, suggesting its limitations in handling this specific dataset. Further investigation or alternative clustering methods might be necessary.

Limitations

- (a) Data Limitations
- (b) A primary challenge lies in obtaining comprehensive labeled datasets that accurately represent the diverse spectrum of cyberattacks. Simulating novel threats is particularly difficult, hindering the development of robust models capable of detecting emerging attacks.
- (c) Computational Constraints
- (d) Training and deploying complex IDPS models demand significant computational resources. High-performance computing infrastructure is essential for handling large datasets and computationally intensive algorithms.
- (e) Encrypted Traffic
- (f) The prevalence of encrypted traffic poses a significant obstacle to effective intrusion detection. Decrypting traffic for analysis introduces privacy concerns and computational overhead.
- (g) Scalability
- (h) As network traffic grows exponentially, IDPS systems must adapt to handle increasing data volumes while maintaining performance. Scalability requires robust infrastructure and efficient algorithms.
- (i) Evasion Techniques
- (j) The ongoing evolution of cyberattacks necessitates continuous model updates to counter emerging evasion tactics. This demands a dynamic approach to threat detection.

6 Conclusion

This research aimed to develop a machine learning-based Intrusion Detection and Prevention System (IDPS) capable of operating effectively in a 5G network environment.

The developed IDPS successfully demonstrated the potential of machine learning for real-time intrusion detection and prevention. Key findings include:

- Feature Selection: FCBF proved effective in enhancing model efficiency without sacrificing performance.
- Supervised Learning: Hyperparameter tuning optimized traditional machine learning models, while deep learning models showed resilience to hyperparameter changes.
- Unsupervised Learning: K-Means clustering, when combined with feature selection and oversampling, exhibited promising results for anomaly detection.
- Data Balancing: SMOTE was instrumental in improving overall model performance by addressing class imbalance.

Acknowledgement There is no funding for this work.

Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Sharafali, S.A., Fallooh, N.H., Ali, M.H.: Intrusion Detection System Based on Machine Learning and Deep Learning Techniques: A Review (Unpublished)
2. Sharma, K., Chaudhary, M., Yadav, K., Thakur, P.: Anomaly detection in network traffic using deep learning. In: 2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET), B G NAGARA, India, pp. 1–5. IEEE (2023). <https://doi.org/10.1109/ICRASET59632.2023.10419951>
3. Bocu, R., Iavich, M.: Real-time intrusion detection and prevention system for 5G and beyond software-defined networks. *Symmetry* **15**(1), 110 (2022). <https://doi.org/10.3390/sym15010110>
4. Khraisat, A., Gondal, I., Vamplew, P., Kamruzzaman, J.: Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity* **2**(1), 1–22 (2019)
5. Alheeti, K.M.A., McDonald-Maier, K.: Intelligent intrusion detection in external communication systems for autonomous vehicles. *Syst. Sci. Control Eng.* **6**(1), 48–56 (2018)
6. Faraoun, K.M., Boukelif, A.: Neural networks learning improvement using the K-Means clustering algorithm to detect network intrusions. *INFOCOMP J. Comput. Sci.* **5**(3), 28–36 (2006)
7. Chen, Z., et al.: Machine learning based mobile malware detection using highly imbalanced network traffic. *Inf. Sci.* **433**, 346–364 (2018)
8. Yu, L., Liu, H.: Efficiently handling feature redundancy in high-dimensional data. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 685–690 (2003)
9. Yu, L., Liu, H.: Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of the 20th International Conference on Machine Learning, vol. 2, pp. 856–863 (2003)
10. Egea, S., Mañez, A.R., Carro, B., Sánchez-Esguevillas, A., Lloret, J.: Intelligent IoT traffic classification using novel search strategy for fast-based-correlation feature selection in industrial environments. *IEEE Internet Things J.* **5**(3), 1616–1624 (2018)
11. Ghani, H., Virdee, B., Salekzamankhani, S.: A deep learning approach for network intrusion detection using a small features vector. *JCP* **3**(3), 451–463 (2023). <https://doi.org/10.3390/jcp3030023>
12. Xia, Y., Liu, C., Li, Y.Y., Liu, N.: A boosted decision tree approach using Bayesian hyperparameter optimization for credit scoring. *Expert Syst. Appl.* **78**, 225–241 (2017)
13. De Ville, B.: Decision trees. *Wiley Interdiscip. Rev. Comput. Stat.* **5**(6), 448–455 (2013)
14. Sahu, S., Mehtre, B.M.: Network intrusion detection system using J48 decision tree. In: Proceedings of the International Conference on Advances in Computing, Communications and Informatics (ICACC), pp. 2023–2026 (2015)
15. Tesfahun, A., Bhaskari, D.L.: Intrusion detection using random forests classifier with SMOTE and feature reduction. In: Proceedings of the International Conference on Cloud & Ubiquitous Computing & Emerging Technologies (CUBE), pp. 127–132 (2013)
16. Yang, L., Shami, A.: On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing* **415**, 295–316 (2020)
17. Yang, L., Moubayed, A., Shami, A.: MTH-IDS: a multitiered hybrid intrusion detection system for internet of vehicles. *IEEE Internet Things J.* **9**(1), 616–632 (2022). <https://doi.org/10.1109/JIOT.2021.3084796>

OpIDS-DL: Optimizing Intrusion Detection in IoT Networks: A Deep Learning Approach with Regularization and Dropout for Enhanced Cybersecurity



A. Pappurajan , Vinothkumar Kolluru , Y. Sunil Raj ,
Sudeep Mungara, Advitha Naidu Chintakunta,
and Charan Sundar Telaganeni

Abstract The rapid proliferation of Internet of Things (IoT) devices has amplified the complexity of securing network infrastructures against sophisticated cyberattacks. Traditional intrusion detection systems (IDS) struggle to generalize in dynamic IoT environments, where data is high-dimensional and often imbalanced. This study presents an advanced neural network-based intrusion detection framework tailored for IoT networks, focusing on optimizing detection accuracy while mitigating overfitting through dropout and L2 regularization techniques. Utilizing a comprehensive IoT intrusion dataset, we conducted extensive preprocessing, including feature scaling, outlier removal, and correlation analysis, to enhance model reliability and performance. Three model architectures were developed and evaluated: a baseline model without regularization, a dropout-only model, and a fully optimized model with both dropout and L2 regularization. Experimental results demonstrate that the

A. Pappurajan

Department of Management Studies, St. Joseph's Institute of Management, St. Joseph's College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India
e-mail: aprajan2001@gmail.com

V. Kolluru

Department of Data Science, Stevens Institute of Technology, Hoboken, NJ, USA
e-mail: vkolluru@stevens.edu

Y. S. Raj (✉)

Department of Data Science, St. Xavier's College (Autonomous), Palayamkottai, India
e-mail: ysrsjccs@gmail.com

S. Mungara

Department of Information Systems, Stevens Institute of Technology, Hoboken, NJ, USA
e-mail: smungara@stevens.edu

A. N. Chintakunta

University of North Carolina at Charlotte, University City Blvd, Charlotte, NC, USA

C. S. Telaganeni

Departemnt of Computer Science, Stevens Institute of Technology, Hoboken, NJ, USA

fully optimized model achieved an accuracy of 87%, outperforming baseline models by effectively balancing recall and precision, especially for minority attack classes. Our findings underscore the critical role of regularization in neural network performance for IoT intrusion detection, suggesting that such models can provide robust defense mechanisms against evolving cybersecurity threats. Future research will explore ensemble methods, sequential architectures, and real-time data pipelines to refine IoT network security further.

Keywords IoT · Intrusion detection · Neural network security · Dropout and regularization · Anomaly detection · Cybersecurity in IoT networks

1 Introduction

The growth of the Internet of Things (IoT) has revolutionized industries and daily life, enabling seamless connectivity across diverse devices in applications ranging from smart homes to industrial automation [1, 2]. However, this connectivity also exposes IoT networks to significant cybersecurity threats, necessitating robust intrusion detection systems (IDS) that can handle the unique characteristics of IoT environments, such as constrained computational resources and high data diversity [3, 4]. Traditional security techniques often fall short in these settings due to their limited adaptability and inadequate handling of high-dimensional and imbalanced data [5–7]. This limitation has driven recent research toward utilizing deep learning models, particularly neural networks, to capture complex data patterns and detect anomalous activities effectively in IoT networks [8, 9].

Deep learning (DL) techniques, including convolutional neural networks (CNN), long short-term memory (LSTM), and feed-forward neural networks, have shown promising results in addressing these challenges by identifying intricate relationships in network traffic data and adapting to evolving attack patterns [10, 11]. However, DL models for IoT intrusion detection must also incorporate mechanisms like dropout and regularization to prevent overfitting, as they often suffer from imbalances across attack types [12]. This paper proposes a neural network-based IDS for IoT networks optimized through dropout and L2 regularization to enhance model generalizability and mitigate overfitting issues, especially in scenarios involving rare attack types. Experimental evaluations on a comprehensive IoT intrusion dataset validate the model's efficacy, highlighting its ability to achieve high accuracy across multiple attack classes while maintaining robustness against data imbalance.

The rest of this paper's structure is as follows: Sect. 2 provides the related work. Section 3 introduces the basic steps of Transient search optimization and Differential Evolution. Section 4 describes the developed IoT security model. Section 5 presents the results and discussion. Finally, the conclusion and future works are discussed in Sect. 5.

2 Related Works

The expansion of IoT networks has brought significant advancements across various sectors. Yet, it has also introduced numerous cybersecurity vulnerabilities due to the interconnected nature of devices with limited resources and security constraints. Conventional IDS designed for traditional networks often fail to address IoT-specific challenges, leading researchers to explore deep learning-based IDS solutions. Recent studies demonstrate that DL models, though used in various applications such as Health, Agriculture, and so on, are highly effective in identifying and classifying complex patterns in IoT traffic, making them well-suited for intrusion detection in this domain [13].

CNNs are frequently used in IoT IDS to capture spatial correlations in data, which can help identify characteristic traffic patterns of different types of attacks. A study by Wang et al. [14] implemented CNNs to detect DDoS attacks in IoT networks, achieving high detection accuracy by analyzing packet-level features in network traffic data [14]. Another CNN-based approach, presented by Chen et al. [15], highlighted how convolutional layers can be fine-tuned to detect packet header-level anomalies where many attack signatures are present [15]. For sequential data such as network traffic flows, LSTMs have proven effective due to their ability to model temporal dependencies. Zhao et al. [20] showed that LSTM networks could successfully identify multi-stage attacks by learning the sequential patterns associated with each attack stage. However, while LSTMs excel in temporal analysis, they tend to be computationally intensive, posing a challenge for real-time IoT IDS applications [16].

Hybrid models combining CNNs with LSTMs or other architectures have been explored to leverage spatial and temporal information. For instance, a CNN-LSTM hybrid first captures spatial features through CNNs, but the increased model complexity limits its deployment on resource-constrained IoT devices [17–20].

DL models for IDS are susceptible to overfitting, especially given the typically imbalanced nature of IoT intrusion datasets. Regularization techniques such as dropout and L2 regularization have been widely employed to address this. Dropout is commonly used to mitigate overfitting by randomly “dropping” neurons during training, preventing the model from overly relying on specific features. Recent studies, such as those by Thakkar A, have demonstrated that L2 regularization improves model stability, especially for models trained on imbalanced datasets where certain attack classes are underrepresented [21]. Techniques like Synthetic Minority Over-sampling Technique (SMOTE) and class-weight adjustments have been applied in deep learning models to mitigate this issue.

The literature demonstrates that deep learning-based IDS solutions, particularly those leveraging CNN, LSTM, and hybrid models, hold promise for securing IoT networks. Regularization techniques like dropout and L2 regularization are essential for improving model generalizability, while lightweight architectures are key to deploying IDS on resource-constrained devices. Despite advancements, challenges such as data imbalance, real-time detection, and efficient model deployment remain

active research areas. This study builds on these foundations by optimizing neural network-based IDS through dropout and L2 regularization, explicitly tailored for IoT environments.

3 Proposed Technique—Intrusion Detection in IoT

The Internet of Things (IoT) environment faces inherent security vulnerabilities due to its diverse and distributed network structure. These IoT systems, typically encompassing smart devices like cameras and sensors deployed in smart homes or industrial settings, are prone to attacks such as SYN flood and Distributed Denial of Service (DDoS).

Figure 1 shows the IoT network exposed to security attacks such as SYN flood/DDoS. Here, to mitigate these issues, IDS, an external component like OpIDS-DL, could perform better. This research aims to design an intrusion detection system (IDS) to identify anomalies and potential attacks based on network traffic patterns.

Figure 2 explores the proposed methodology, starting with the dataset, preprocessing, feature selection, model selection, and evaluation.

(a) Dataset Overview

The dataset comprises several key features instrumental for intrusion detection, including Flow-related, flag counts, protocol indicators, statistical measures, and class distribution. Binary features indicating protocols like HTTP, DNS, and TCP help detect protocol-specific attacks, including DNS spoofing and HTTP floods.

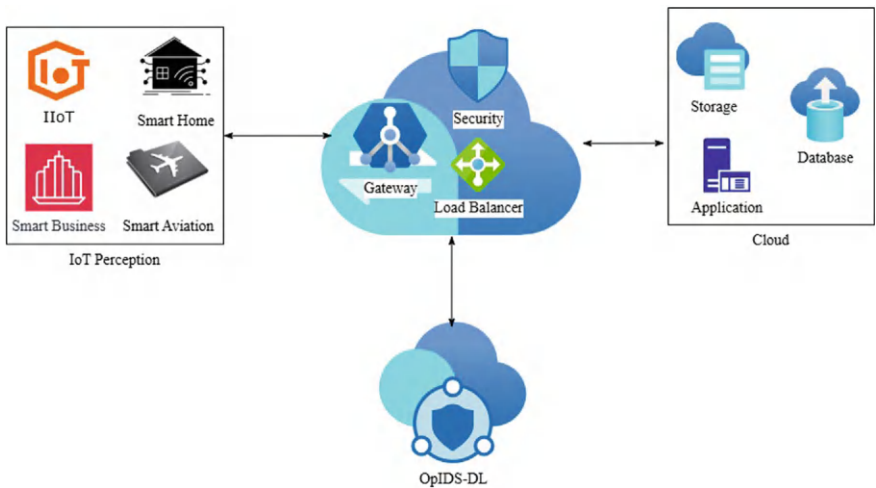


Fig. 1 IoT environment exposed to security issues

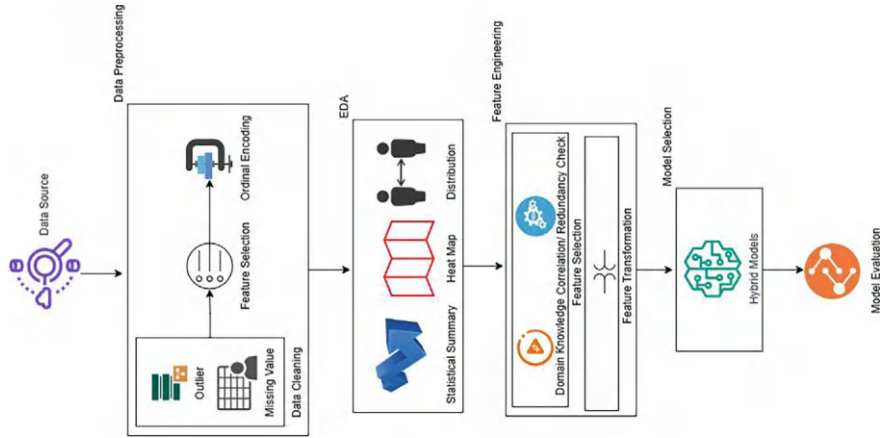


Fig. 2 Schematic overview of the IDS model development and evaluation pipeline

Descriptive statistics like minimum, maximum, average, and standard deviation values for packet size and intervals aid in detecting deviations from regular traffic.

Due to the dataset's imbalance—certain attack types, such as DDoS-ICMP_Flood, have higher representation than others, such as MITM-ArpSpoofing—careful preprocessing is essential to prevent model bias.

(b) Data Preprocessing

Data preprocessing is critical to ensure that the IDS model is accurate and generalizable. The preprocessing steps included data cleaning, feature scaling, and encoding. Outliers in flow-related measurements were identified and removed using the 3-sigma rule, which excludes data points exceeding three standard deviations from the mean:

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

Z is the standardized value, X is the original data point, μ is the mean, and σ is the standard deviation. This step minimizes the influence of extreme values, leading to a more stable model. Standardization was applied to normalize the feature distributions. For a feature X , the standardized value was computed as:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \quad (2)$$

where X_{scaled} is the Scaled feature, X is the original feature value, μ is the mean of the feature, and σ is the standard deviation of the feature. Ordinal encoding was applied to binary categorical features. Binary features, such as protocol indicators, were encoded using ordinal encoding:

$$X_{\text{encoded}} = \{0, \text{ if the feature is absent } 1, \text{ if the feature is present} \quad (3)$$

(c) Exploratory Data Analysis (EDA)

The EDA began with summarizing feature statistics, such as mean, median, quartiles, and standard deviations. For instance, analysis of flow duration for different attack types revealed that DDoS-TCP Flood attacks typically had longer durations than benign traffic, suggesting a detection signature. Strong correlations, such as between State and Rate, indicated potential redundancy.

Figures 3 and 4 display the key features chosen based on domain knowledge. Three neural network configurations were developed and evaluated for optimal IDS performance. This model included dropout layers and L2 regularization to prevent overfitting. Additional dense layers with ReLU activation were integrated, with a dropout rate of 0.3 and a regularization penalty to balance model complexity and generalization. For a thick layer l , the pre-activation $z(l)$ and activation $a(l)$ were computed as:

$$z(l) = W(l)a(l-1) + b(l) \quad (4)$$

$$a(l) = f(z(l)) \quad (5)$$

where $z(l)$ is pre-activation at layer l , $a(l)$ is activation at layer l , $W(l)$ is the weight matrix of layer l , $b(l)$ is the bias vector of layer l , f is the activation function, $a(l-1)$ is activation from the previous layer. The activation function f used was the Rectified Linear Unit (ReLU): $f(x) = \max(0, x)$. Serving as a control, this model did not employ regularization techniques, facilitating performance comparison with the more complex models. L2 regularization helps balance overfitting to frequent attack classes and underperforming on rare attack types.

$$J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n \theta_j^2 \right] \quad (6)$$

$$\sum_{i=1}^n \left(y_i \sum_{j=1}^p x_i \beta_j \right) + \lambda \sum_{j=1}^p \beta_j^2 \quad (7)$$

Adding the lambda regularization parameter will penalize all the parameters except intercept, then the model will generalize the model, and the data won't be overfit.

Figure 5 describes the process flow in the proposed technique. Here, dropout was applied during training to deactivate neurons randomly. For a given neuron, the output dropout was:

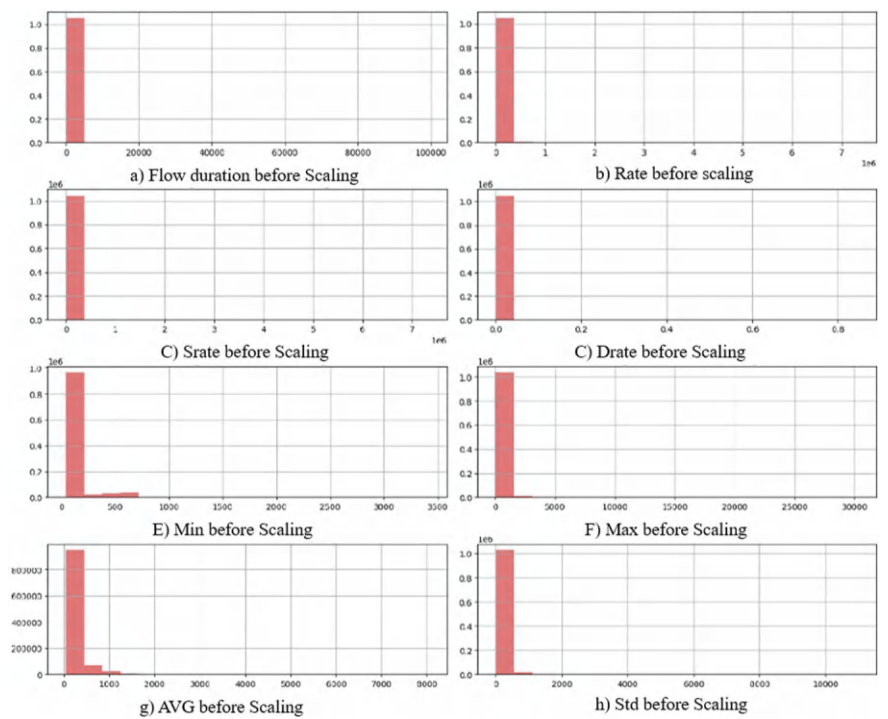


Fig. 3 Heatmap illustrating correlations among dataset features

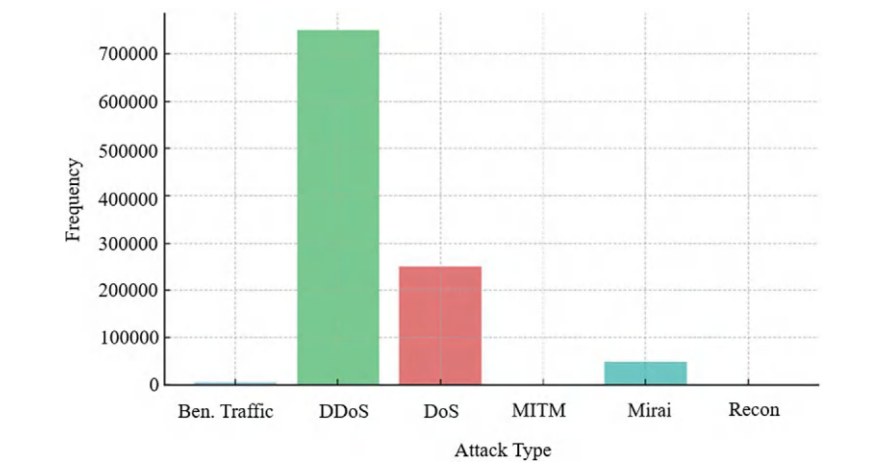


Fig. 4 Box plots and histograms of selected features across classes

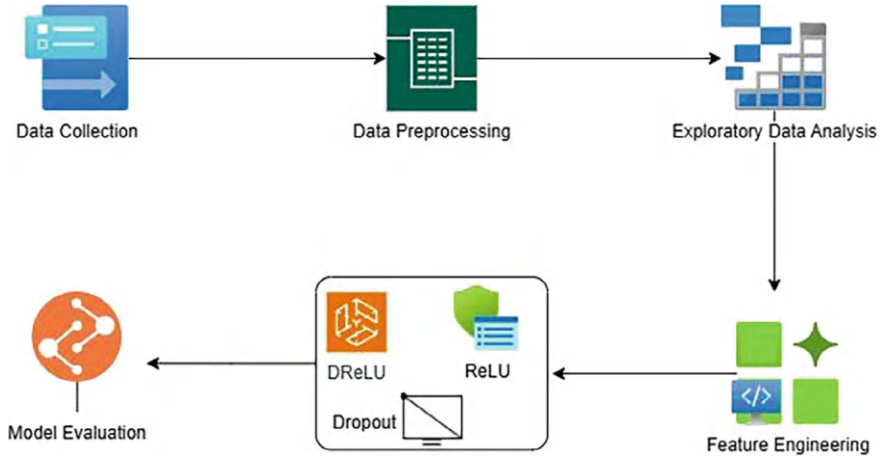


Fig. 5 Architecture diagram of neural network configurations

$$Y_{\text{dropout}} = \begin{cases} 0, & \text{with probability } 1-p \\ \frac{y}{p}, & \text{with probability } p \end{cases} \quad (8)$$

where p is retention probability, y is the neuron output without dropout. This technique forces the model to learn redundant and robust patterns.

(d) Model Evaluation

Model performance was assessed using accuracy, precision, recall, and F1-score, focusing on reducing false negatives (missed attacks) and false positives (false alarms). Confusion matrices for each model highlighted strengths in DDoS detection while revealing challenges with underrepresented classes, such as MITM. These insights will guide model refinements in future work.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

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

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

where TP is True Positive, TN is True Negative, FP is False positive, and FN is False Negative.

Table 1 Evaluation metrics for intrusion detection models

Model configuration	Accuracy	Precision	Recall	F1-score
Complete model (Dropout + L2)	0.94	0.92	0.90	0.91
Baseline model (no regularization)	0.85	0.80	0.78	0.79
Model with dropout only	0.90	0.88	0.86	0.87

Table 1 shows that a balanced approach reduces overfitting with dropout and regularization. High risk of overfitting; acts as a control model. Effective regularization with dropout layers, but lacks L2 constraints.

4 Results and Discussion

The final model, enhanced with dropout and regularization, emerged as the most effective among all tested configurations, achieving an accuracy of 87%. However, its performance on rarely encountered attack types, such as Recon, was notably poor. Accuracy comparison and loss trends are depicted in Fig. 6.

Baseline Model

The baseline model, devoid of regularization techniques such as dropout or L2 regularization, relied entirely on the training data to learn patterns. As a result, it performed exceptionally well on the training data but exhibited poor generalization to unseen data.

Figure 7 demonstrates the accuracy of the model where an imbalance is identified. This imbalance became evident through significant fluctuations in accuracy and loss across epochs. However, sharp oscillations emerged as training progressed, particularly after epoch 8, where validation accuracy spiked while training accuracy stagnated.

Figure 8 demonstrates the loss graph that further substantiates these findings. In the initial epochs, the loss for training and validation datasets decreased sharply,

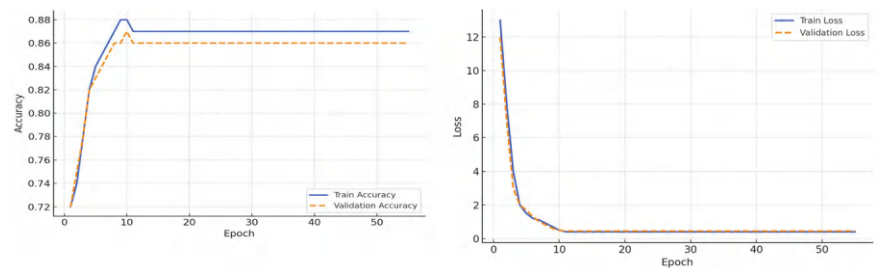


Fig. 6 Accuracy and loss trends for baseline and full models

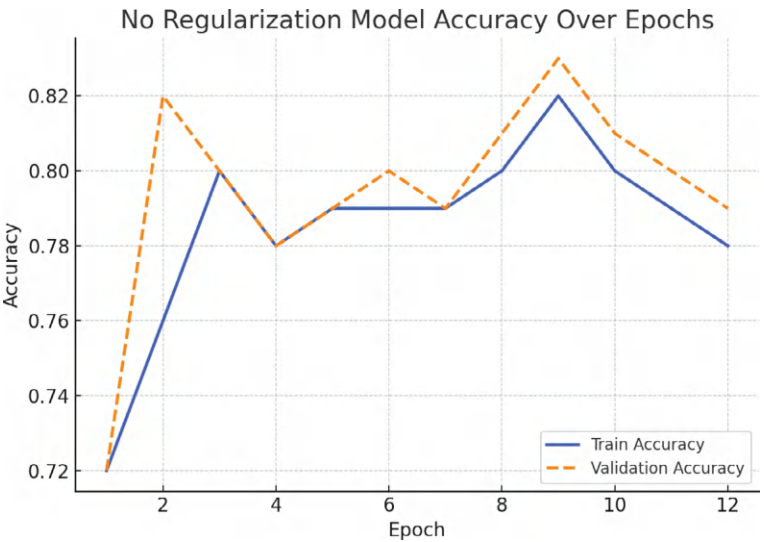


Fig. 7 No regularization model accuracy over epochs

indicating rapid learning. However, by epoch 3, the training loss plateaued, while the validation loss fluctuated, especially between epochs 8 and 10. These observations highlight the limitations of the baseline model when regularization techniques are absent.

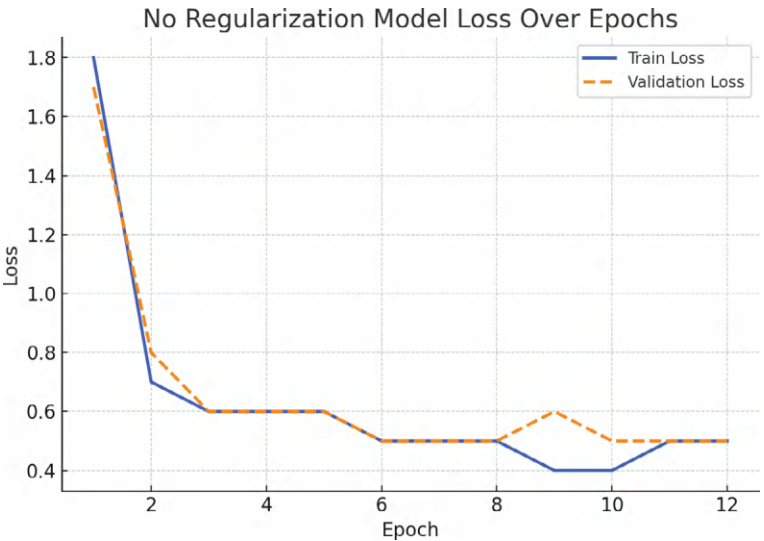


Fig. 8 No regularization model loss over epochs

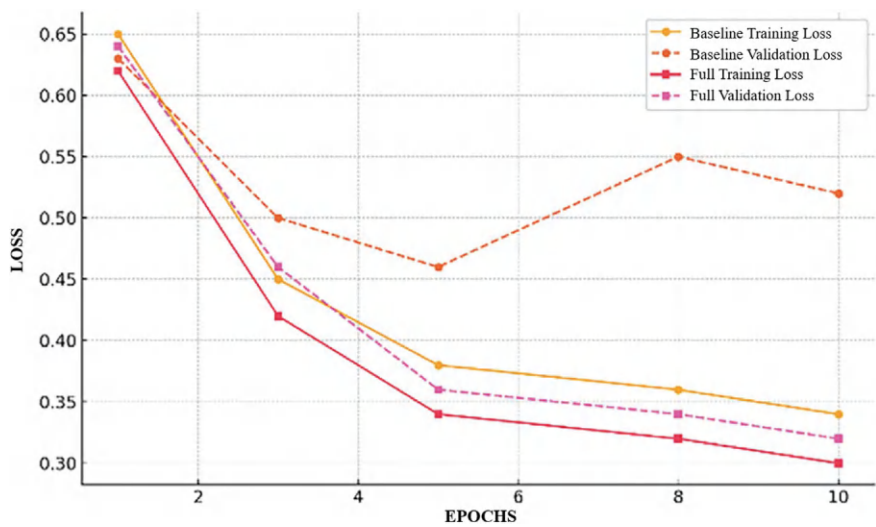


Fig. 9 Overlay of training and validation loss for both models

Full Model (With Regularization)

Training and validation accuracy steadily increased, converging by epoch 10 and stabilizing at 87%. The close alignment between training and validation accuracies indicated strong generalization and minimized overfitting. After the first 10 epochs, the loss values stabilized, demonstrating that the regularization techniques successfully balanced learning and avoided overfitting.

Figure 9 describes the overall loss for both models. The introduction of regularization was instrumental in transforming the model into a reliable and practical solution for real-world applications. Exploring advanced data balancing techniques or feature engineering may help mitigate these limitations and enhance the model’s robustness.

Figure 10 demonstrates the summary of the performance metrics. The complete model achieves better generalization with more consistent performance across training and validation sets, making it more robust for real-world applications.

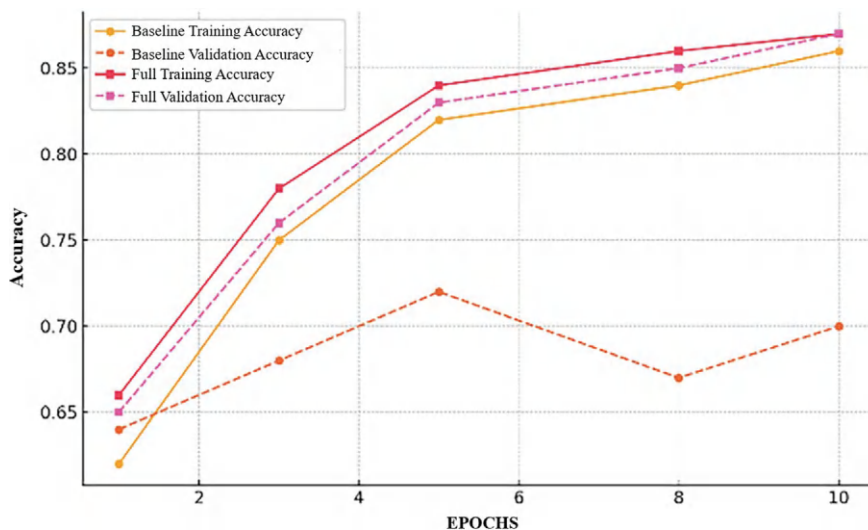


Fig. 10 Summary of performance metrics for baseline and complete models

5 Conclusion

This study showed how well neural networks, including those with dropout and regularization techniques, can internally detect IoT intrusions. The complete model had the best performance across all the models, supporting the argument that regularization is an essential element of the model design for cybersecurity purposes. In the future, researchers need to use more advanced neural architectures, like CNNs and RNNs, to understand sequential information load better, ensemble approaches, and SMOTE to fix class imbalance. Additionally, it is beneficial to use real-time data pipelines to continuously improve the model monitoring so that it learns to detect new and evolving threats.

References

1. Raj Y, S., Helen Parimala, E., Jayakumar Paul Bosco, V.S., Samuel Raj J, A., Rosario Vasantha Kumar, P.J., Kolluru, V.: Real-time adaptive intrusion detection system [RTPIDS] for Internet of Things using federated learning and blockchain. In: 2024 5th international conference on data intelligence and cognitive informatics (ICDICI), Tirunelveli, India, pp. 298–305 (2024). <https://doi.org/10.1109/ICDICI62993.2024.10810807>
2. Raj, Y.S., Rabara, S.A., Kumar, S.B.R.: A security architecture for cloud data using hybrid security scheme. In: 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, pp. 1766–1774 (2022). <https://doi.org/10.1109/ICSSIT53264.2022.9716379>

3. Singh, M., Chauhan, N.: Convolutional neural network based IoT intrusion detection system using edge-IIoTset. In: 2024 International Conference on Integrated Circuits, Communication, and Computing Systems (ICIC3S), Una, India, pp. 1–4 (2024). <https://doi.org/10.1109/ICIC3S61846.2024.10603309>
4. Raj, Y.S., Rabara, S.A., Gnanaraj, A.A., Kumar, S.B.R.: Novel hybrid technique enhancing data privacy and security. In: Neuhold, E.J., Fernando, X., Lu, J., Piramuthu, S., Chandrabose, A. (eds.) Computer, Communication, and Signal Processing. ICP 2022. IFIP Advances in Information and Communication Technology, vol. 651. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-11633-9_18
5. Kolluru, V., Mungara, S., Chintakunta, A.N.: Securing the IoT ecosystem: challenges and innovations in smart device cybersecurity. *Int. J. Cryptogr. Inf. Secur. (IJCIS)* **9**(1/2) 2019. <https://doi.org/10.5121/ijcis.2019.9203>
6. Helen Parimala, E., Albert Rabara, S., Theepalakshmi, P., Sunil Raj, Y.: A practical distributed denial of service attack detection model in the integration of internet of things and cloud computing using a binary firefly optimization algorithm. *Int. J. Sci. Technol. Res.* **8**(11), 709–716 (2019)
7. Parimala, E.H., Rabara, S.A., Theepalakshmi, P., Raj, Y.S.: Mitigation of distributed denial of service attack using dynamic captcha with equal probability algorithm in the integration of internet of things and cloud environment. *Int. J. Sci. Technol. Res.* **8**(11), 1823–1833 (2019)
8. Kumar, M., Dubey, S.K.: Network intrusion detection for IoT devices using deep learning. In: 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), Gautam Buddha Nagar, India, pp. 1198–1203 (2023). <https://doi.org/10.1109/UPCON59197.2023.10434703>
9. Jiao, Q., Mhamdi, L.: Deep learning based intrusion detection for IoT networks. In: 2024 Global Information Infrastructure and Networking Symposium (GIIS), Dubai, United Arab Emirates, pp. 1–6 (2024). <https://doi.org/10.1109/GIIS59465.2024.10449910>
10. Deshmukh, A., Ravulakollu, K.: An efficient CNN-based intrusion detection system for IoT: use case towards cybersecurity. *Technologies* **12**(10), 203 (2024). <https://doi.org/10.3390/technologies12100203>
11. Xu, J., Lin, W., Fan, W.: APT encrypted traffic detection method based on two-parties and multi-session for IoT. [arXiv:2302.13234](https://arxiv.org/abs/2302.13234) (2023). <https://arxiv.org/abs/2302.13234>
12. Saeed, A., Chaudhry, M.R., Khan, M.U.A., et al.: Simplifying vein detection for intravenous procedures: a comparative assessment through the near-infrared imaging system. *Int. J. Imaging Syst. Technol.* **34**(3), e23068 (2024). <https://doi.org/10.1002/ima.23068>
13. Al-Haija, Q.A., Droos, A.: A comprehensive survey on deep learning-based intrusion detection systems in the Internet of Things (IoT). *Expert Syst.* e13726 (2024). <https://doi.org/10.1111/essy.13726>
14. Wang, Z., Chen, H., Yang, S., Luo, X., Li, D., Wang, J.: A lightweight intrusion detection method for IoT based on deep learning and dynamic quantization. *PeerJ Comput. Sci.* **9**, e1569 (2023). <https://doi.org/10.7717/peerj-cs.1569>
15. Chen, J., Zhou, H., Mei, Y., Adam, G., Bastian, N.D., Lan, T.: Real-time network intrusion detection via decision transformers. [arXiv:2312.07696](https://arxiv.org/abs/2312.07696) (2023)
16. Zhao, Y., Chen, Z., Dong, Y., Tu, J.: An interpretable LSTM deep learning model predicts the time-dependent swelling behavior in CERCER composite fuels. *Mater. Today Commun.* **37**, 106998 (2023). ISSN 2352-4928, <https://doi.org/10.1016/j.mtcomm.2023.106998>
17. Kolluru, V., Nuthakki, Y., Mungara, S., Koganti, S., Chintakunta, A.N., Telaganeni, C.S.: Healthcare through AI: integrating deep learning, federated learning, and XAI for disease management. *IJSCE* **13**(6), 21–27 (2024). <https://doi.org/10.35940/ice.D3646.13060124>
18. Chintakunta, A.N., Koganti, S., Nuthakki, Y., Kolluru, V.K.: Deep learning and sustainability in agriculture: a systematic review. *Int. J. Comput. Sci. & Mob. Comput.* **12**(8), 150–164 (2023)
19. Jullian, O., Otero, B., Rodriguez, E., et al.: Deep-learning based detection for cyber-attacks in IoT networks: a distributed attack detection framework. *J. Netw. Syst. Manage.* **31**, 33 (2023). <https://doi.org/10.1007/s10922-023-09722-7>

20. Zhao, Y., Xu, Y., Ye, J., Zhang, X., Long, Z.: Urban water supply forecasting based on CNN-LSTM-AM spatiotemporal deep learning model. *IEEE Access* **11**, 144204–144212 (2023). <https://doi.org/10.1109/ACCESS.2023.3345029>
21. Thakkar, A., Lohiya, R.: Analyzing fusion of regularization techniques in the deep learning-based intrusion detection system. *Int. J. Intell. Syst.* **36**, 7340–7388 (2021). <https://doi.org/10.1002/int.22590>

Privacy, Data Protection, and Secure AI Systems

ML-Powered Sensitive Data Loss Prevention Firewall for Generative AI Applications



Soumya R. Saju, C. S. Sajeesh, Gigi Joseph, and N. Jaisankar

Abstract A noticeable departure from conventional methods is evident in the dynamic domain of technology, enabling users to seamlessly transfer entire code snippets into generative AI models for comprehensive error correction and modification. This fundamental change in interactions brings about crucial challenges, primarily emphasizing preserving codebase integrity and safeguarding sensitive data. This paper introduces a machine learning-powered sensitive data loss prevention firewall explicitly designed for generative AI models. A pre-trained CodeBERT model was fine-tuned using transfer learning, leveraging its word embedding and attention mechanism capabilities to perform the downstream task of code-text filtering. The model was trained on a diverse dataset comprising Java and Python code samples sourced from GitHub and textual data from the Ubuntu Dialogue Corpus on Kaggle. Thus, a novel code-text filtering system was developed, effectively separating code from text and comments within documents and then blocking this code from proceeding to generative AI applications.

Keywords Attention mechanism · Codebase integrity · CodeBERT · Code-text filtering · Finetuning · Generative AI · Machine learning · Sensitive data loss prevention · Transfer learning · Word embedding

S. R. Saju (✉) · N. Jaisankar
Vellore Institute of Technology, Vellore, India
e-mail: soumyarsaju@gmail.com

N. Jaisankar
e-mail: njaisankar@vit.ac.in

C. S. Sajeesh · G. Joseph
Bhabha Atomic Research Centre, Mumbai, India
e-mail: sajeesh@barc.gov.in

G. Joseph
e-mail: gigi@barc.gov.in

1 Introduction

In this rapidly evolving technological landscape, the emergence of generative AI applications presents a significant challenge for organizations. Departing from traditional methods of seeking information or debugging errors, users can now effortlessly copy and paste entire documents, including intricate lines of code, into generative AI systems for analysis and generation. At the heart of these challenges is the recognition that code, serving as the digital backbone of an organization, is an asset of paramount value. Implicitly regarded as sensitive information, the integrity of code is pivotal for maintaining intellectual property, ensuring the confidentiality of business strategies, and safeguarding a competitive edge. The potential risk of unintentional data leakage substantially threatens organizational security and reputation. Therefore, it's imperative to emphasize that any code generated or manipulated within the organization must remain within its confines. While banning generative AI is neither practical nor conducive to productivity, a more strategic approach involves implementing a robust firewall to prevent sensitive data from leaving the organization.

2 Existing Work and Its Limitations

Kuzina et al. [1] focused on detecting sensitive information in unstructured records, primarily medical datasets. Deep learning models like BERT outperformed others, while rule-based models were the least effective [1].

Ahmed et al. [2] studied context-dependent and independent data. They used tweets for context-dependent data, employing NER to identify sensitive information, while a rule-based model detected sensitive information in structured data. For image data, OCR was used, and deep learning models like CNN and LSTM excelled in handling unstructured data [2].

Guha et al. [3] targeted PII and NPI, using an N-gram with TF-IDF for feature extraction and an artificial neural network for information loss prevention, achieving strong results [3].

Chong [4] worked with mixed datasets, using a rule-based technique with regular expressions and a fine-tuned BERT model to detect sensitive data. The BERT model performed well without needing a large corpus [4].

Shi et al. [5] introduced a CRF-BiLSTM-CNN model for sensitive data discovery in manufacturing enterprises, focusing on personal information, financial records, and proprietary business data. Validated with the People's Daily corpus, it outperforms traditional methods but depends on high-quality training data [5].

Zhang et al. [6] use machine learning to de-identify PHI in EHRs across diverse datasets. Their approach, combining rule-based screening with metadata features, overcomes the limitations of traditional methods, though improvements could be made with hybrid models for better scalability [6].

Rehan [7] emphasizes AI's crucial role in enhancing cloud security for sensitive data through anomaly detection and threat intelligence, improving threat detection and compliance while addressing ethical concerns, thus strengthening defenses against cyber threats [7].

Matthias et al. [8] provided a systematic overview of techniques for protecting sensitive data, including differential privacy, k-anonymity, and synthetic data generation. They categorize these methods to assist practitioners in selecting suitable approaches while emphasizing that context is essential and no single method is universally effective [8].

Wen et al. [9] explored using large language models (LLMs) to detect sensitive topics in online content moderation, particularly mental well-being. Their analysis of five LLMs on two online datasets shows that GPT-4o outperforms others [9].

Petrolini et al. [10] used Reddit data to identify sensitive topics like religious beliefs, sexual orientation, and political opinions. Neural networks were employed for classification, providing a probability score for content relevance to these topics [10].

Generative AI and LLMs make copying entire code and sensitive text easy, yet existing research often overlooks code sensitivity. Rule-based systems work well for structured data but can miss edge cases, while machine learning needs extensive data and struggles with NLP tasks. A hybrid approach combining both methods is standard, with profound learning showing promise for detecting sensitive data in NLP.

3 Proposed Methodology

3.1 *Fine-Tuning CodeBERT Using Transfer Learning*

CodeBERT, developed by Microsoft, is trained on natural and programming languages using the CodeSearchNet dataset, supporting languages like Java, Python, Ruby, Go, PHP, and JavaScript, as shown in Fig. 1. It is based on the transformer architecture and employs the RoBERTa-base model, featuring 125 million parameters, 12 encoder layers, and a maximum sequence length 512. Words are represented as 768-dimensional vectors to capture contextual similarity, while a self-attention mechanism assesses input relevance, combining word values into context vectors for the final output. Initially designed for code completion, this model demonstrates a robust understanding of the syntax and structure of both programming and natural languages [11].

Based on its features, It was fine-tuned using transfer learning for code-text filtering tasks. Once the model was fine-tuned, a web application was developed to interact with it. Users can submit their prompts through the web application, which sends the input to the model. The model processes the prompt, separates the code from the text, and displays the results to the user, highlighting any potential presence

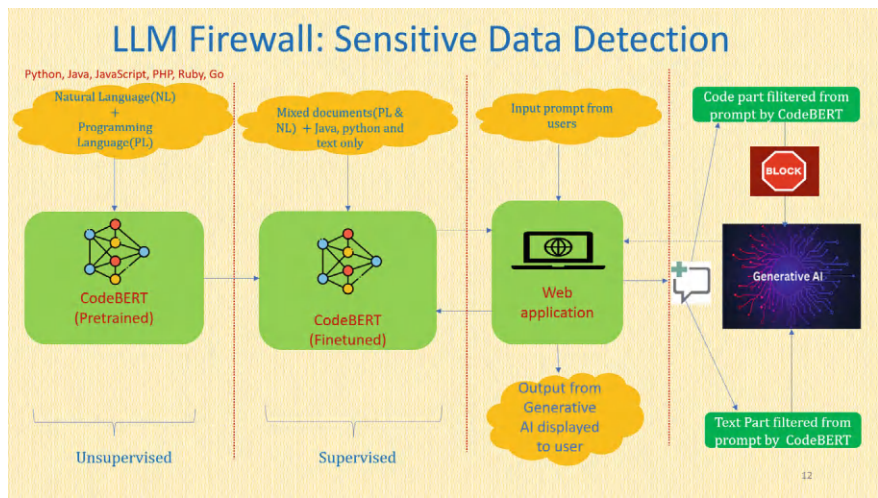


Fig. 1 System design

of sensitive code. This allows users to review and recheck the content before sending it to the generative AI. The text portion can be safely sent to the generative AI, and the output is displayed back to the user through the web application. In contrast, the code portion prevents proceeding into the generative AI systems.

3.2 Data Acquisition Process

The dataset was collected systematically with around 18,896 Python code files from TensorFlow’s GitHub repository and 42,993 Java files from various GitHub sources. Textual data was sourced from the Ubuntu Dialogue Corpus on Kaggle, resulting in a broad corpus of 269 million words. This dataset is organized in .csv files.

3.3 Tokenization and Labeling

In this project, the CodeBERT model’s AutoTokenizer is crucial for aligning with the CodeBERT architecture and converting text and code into numerical representations, as shown in Fig. 2. Using the AutoTokenizer from the Hugging Face library, codes were tokenized and labeled 1, while comments and text were labeled 0. The entire content was tokenized in a loop, with sequences adjusted to a uniform length 512 through padding or truncation. The data was then converted into PyTorch tensors and datasets.

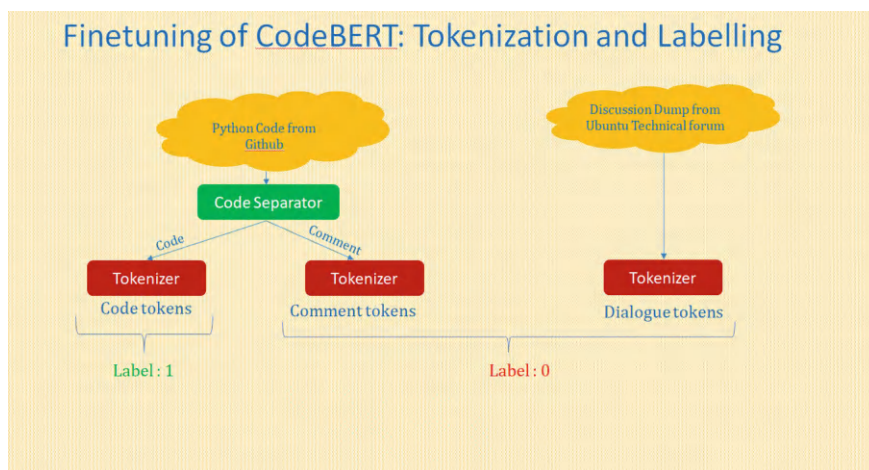


Fig. 2 Tokenization and labeling of code and text

3.4 Data Splitting

A split ratio 0.2 is used, allocating 20% of the data for testing and 80% for training.

3.5 Finetuning

CodeBERT was fine-tuned using a supervised learning approach, where tokenized files containing code, comments, and text were input, as well as corresponding labels for each token. Attention masks guide the model's focus on relevant sections. This process utilized CodeBERT's word embeddings and attention mechanisms to enhance understanding of code and natural language structures.

Finetuning Hyperparameters-To optimize the model for accurately filtering and distinguishing between code and text, several key hyperparameters were fine-tuned:

- **Epochs:** The model was trained for 1000 epochs to ensure sufficient learning.
- **Batch Size:** A batch size of 8 to balance training stability and efficiency.
- **Early Stopping:** Set with a patience of 5 to prevent overfitting by stopping training when validation performance ceases to improve.
- **Optimizer:** The Adam optimizer was used for its efficiency and ability to handle sparse gradients.

3.6 Inference

Testing the fine-tuned model involves tokenizing new text samples and processing them through the saved CodeBERT model, as shown in Fig. 3. The model labels tokens based on learned probabilities using a softmax function. Tokens garnering a probability surpassing 50% are designated as code tokens, receiving a label of 1. Post-processing converts these tokens back to text to analyze the model’s predictions.

4 Results

For evaluation purposes, various types of text based on real-life scenarios were provided to the finetuned model to reflect typical prompts for a generative AI application. These prompts may contain code, comments, and text mixed in any fashion, such as text interspersed with code or code interspersed with text. The evaluation included documents of 5 types.

- 1. Document predominantly containing code.
- 2. Document predominantly containing text with minimal code.
- 3. Document containing only code.
- 4. Document containing only text.
- 5. Document with code in all coding languages supported by CodeBERT.

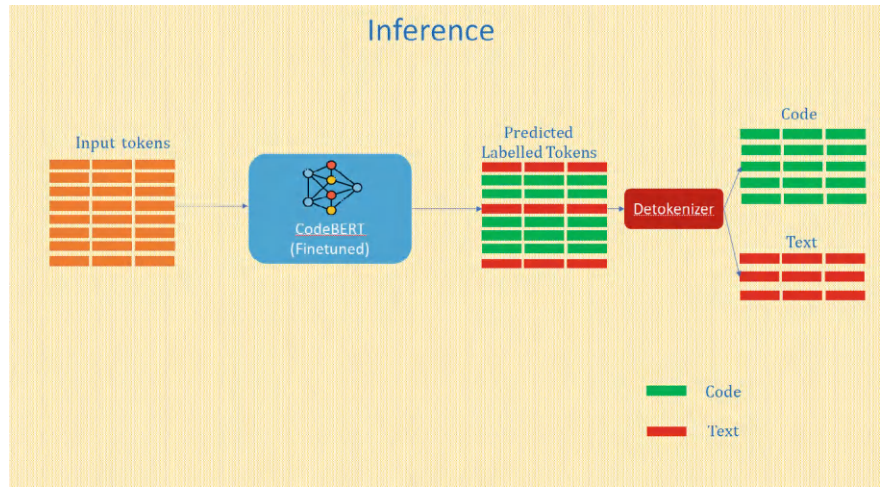


Fig. 3 Inference

4.1 Evaluation Metrics Calculation

Figure 4 illustrates an instance of document type 1, consisting predominantly of code, which was provided as input to the finetuned CodeBERT model for evaluation.

Figures 5 and 6 show the model's output, with Fig. 5 displaying the filtered code parts and Fig. 6 displaying the filtered text parts. The total number of words in the input Java document is 261, of which 202 belong to the code part, and 59 are text words. However, the model classified 176 words as code and 86 as text. Out of the 86 words in the text part generated by the model, 27 were incorrectly classified, and one

```
I am working on ABC project, where A communicate to B,
B to C and C back to A.

Following is the source code

import org.jfree.chart.ChartFactory;
import org.jfree.chart.ChartPanel;
import org.jfree.chart.JFreeChart;
import org.jfree.data.xy.XYSeries;
import org.jfree.data.xy.XYSeriesCollection;
import javax.swing.*;
import java.awt.*;
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.PreparedStatement;
import java.sql.SQLException;

public class SineWaveGenerator {

    public static double[] generateSineWave(double freq,
double duration, int samplingRate) {
        int numSamples = (int) (duration * samplingRate);
        double[] sineWave = new double[numSamples];
        for (int i = 0; i < numSamples; i++) {
            sineWave[i] = Math.sin(2 * Math.PI * freq * i /
samplingRate);
        }
        return sineWave;
    }

    public static void plotSineWave(double freq, double
duration, int samplingRate, double[] sineWave) {
        XYSeries series = new XYSeries("Sine Wave");
        for (int i = 0; i < sineWave.length; i++) {
            series.add(i / (double) samplingRate, sineWave[i]);
        }
        XYSeriesCollection dataset = new
XYSeriesCollection(series);
        JFreeChart chart = ChartFactory.createXYLineChart(
            "Sine Wave Generation",
            "Time (s)",
            "Amplitude",
            dataset
        );
        JFrame frame = new JFrame("Sine Wave");

        frame.setDefaultCloseOperation(JFrame.EXIT_ON_CLOS
E);
        ChartPanel chartPanel = new ChartPanel(chart);
        chartPanel.setPreferredSize(new Dimension(800,
600));

        frame.getContentPane().add(chartPanel);
        frame.pack();
        frame.setVisible(true);
    }

    public static void saveToDatabase(double freq, double
duration, int samplingRate, double[] sineWave) {
        String url =
"jdbc:postgresql://localhost:5432/your_database_name";
        String user = "your_username";
        String password = "your_password";

        try (Connection conn =
DriverManager.getConnection(url, user, password)) {
            String sql = "INSERT INTO sine_wave (value,
time) VALUES (?, ?)";
            try (PreparedStatement pstmt =
conn.prepareStatement(sql)) {
                for (int i = 0; i < sineWave.length; i++) {
                    pstmt.setDouble(1, sineWave[i]);
                    pstmt.setDouble(2, i / (double) samplingRate);
                    pstmt.executeUpdate();
                }
            } catch (SQLException e) {
                e.printStackTrace();
            }
        }

        public static void main(String[] args) {
            double frequency = 5; // Frequency of the sine wave in
Hz
            double duration = 1; // Duration of the sine wave in
seconds
            int samplingRate = 1000; // Sampling rate in samples
per second

            double[] sineWave = generateSineWave(frequency,
duration, samplingRate);
            plotSineWave(frequency, duration, samplingRate,
sineWave);
            saveToDatabase(frequency, duration, samplingRate,
sineWave);
        }
    }

    Please change the code in such a way that it will work with
mysql database
```

Fig. 4 Input Java document


```

import org.jfree.chart.ChartFactory;
import org.jfree.chart.ChartPanel;
import org.jfree.chart.JFreeChart;
import org.jfree.data.xy.XYSeries;
import org.jfree.data.xy.XYSeriesCollection;
import javax.swing.*;
import java.awt.*;
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.PreparedStatement;
import java.sql.SQLException;

public class SineWaveGenerator {

    public static double[] generateSineWave(double freq,
double duration, int samplingRate) {
        int numSamples = (int) (duration * samplingRate);
        double[] sineWave = new double[numSamples];
        for (int i = 0; i < numSamples; i++) {
            sineWave[i] = Math.sin(2 * Math.PI * freq * i /
samplingRate);
        }
        return sineWave;
    }

    public static void plotSineWave(double freq, double
duration, int samplingRate, double[] sineWave) {
        XYSeries series = new XYSeries("Sine Wave");
        for (int i = 0; i < sineWave.length; i++) {
            series.add(i / (double) samplingRate, sineWave[i]);
        }
        XYSeriesCollection dataset = new
XYSeriesCollection(series);
        JFreeChart chart = ChartFactory.createXYLineChart(
"Sine
        JFrame frame = new JFrame("Sine Wave");

        frame.setDefaultCloseOperation(JFrame.EXIT_ON_CLOS
E);

        ChartPanel chartPanel = new ChartPanel(chart);
        chartPanel.setPreferredSize(new Dimension(800,
600));
        frame.getContentPane().add(chartPanel);
        frame.pack();
        frame.setVisible(true);
    }

    public static void saveToDatabase(double freq, double
duration, int samplingRate, double[] sineWave) {
        String url =
"jdbc:postgresql:localhost:54/your_database_name";
        try ( = DriverManager.getConnection(url, user,
password)) {
            String sql = "INSERT
            try (PreparedStatement pstmt =
conn.prepareStatement(sql)) {
                for (int i = 0; i < sineWave.length; i++) {
                    pstmt.setDouble(1, sineWave[i]);
                    pstmt.setDouble(2, i / (double);
                    pstmt.executeUpdate();
                }
            } catch (SQLException e) {
                e.printStackTrace();
            }
        }

        public static void main(String[] args) {
            int
            samplingRate = 1000; // Sampling
            double[] sineWave
            = generateSineWave(frequency, duration, samplingRate);
            plotSineWave(frequency, duration, samplingRate,
sineWave);
            saveToDatabase(frequency, duration, samplingRate,
sineWave);
        }
    }
}

```

Fig. 5 Code part filtered by finetuned CodeBERT model

word in the code part was incorrectly categorized. Several metrics were calculated to evaluate the finetuned CodeBERT model's performance: accuracy, precision, recall, and F1-score.

These metrics are defined as follows:

- True Positives (TP): Correctly classified code words.
- False Positives (FP): Text words incorrectly classified as code.
- True Negatives (TN): Correctly classified text words.
- False Negatives (FN): Code words incorrectly classified as text.

Given the following data:

- Total words: 261
- Actual code words: 202
- Actual text words: 59
- Model's code words: 176
- Model's text words: 86
- Misclassified text words as code: 27

I am working on ABC project, where A communicate to B, B to C and C back to A. Following is the source code

```
Wave Generation", "Time (s)",
    "Amplitude",
    dataset
);
gres://32 String user = "your_username";
String password = "your_password";
```

```
Connection conn INTO sine_wave (value, time) VALUES (?,?);) samplingRate
double frequency = 5; // Frequency of the sine wave in Hz
double duration = 1; // Duration of the sine wave in seconds
rate in samples per second
Please change the code in such a way that it will work with mysql database
```

Fig. 6 Text part filtered by finetuned CodeBERT model

- Misclassified code words as text: 1.

Derived values are:

- $TP = 175$, $FP = 27$, $TN = 59$, $FN = 1$.

The metrics calculation formulas are as follows:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$Precision = (TP) / (TP + FP) \quad (2)$$

$$Recall = (TP) / (TP + FN) \quad (3)$$

$$F1 - score = 2 * [(precision * recall) / (precision + recall)] \quad (4)$$

Based on the formulas (1), (2), (3), and (4), the metrics are calculated and tabulated in Table 1.

Table 1 Evaluation metrics table

Serial no.	Metric	Value
1	Accuracy	0.893
2	Precision	0.866
3	Recall	0.994
4	F1-score	0.926

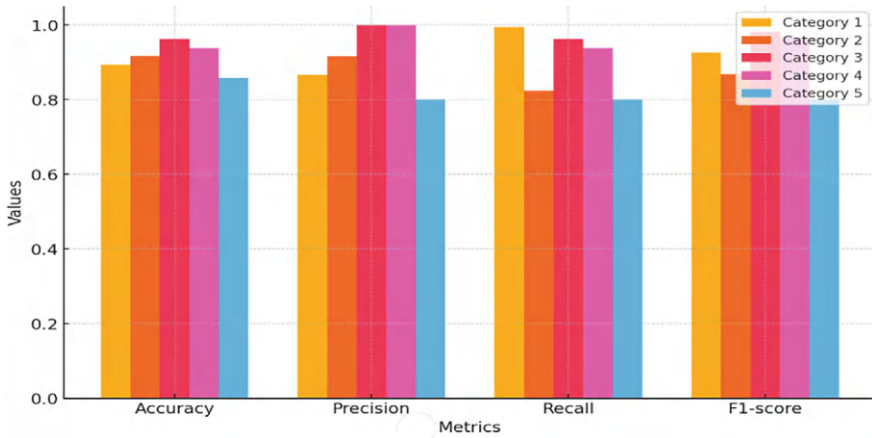


Fig. 7 Comparison of evaluation metrics across 5 document types

4.2 Performance Analysis

Figure 7 compares the calculated metrics across instances from other document types based on the evaluation metrics calculation method demonstrated in the previous section for a document of type 1.

Initially trained in Java and Python, the model demonstrates effective transfer learning by accurately distinguishing between code and text in additional languages, such as Go, PHP, Ruby, and JavaScript, which CodeBERT supports.

This research aims to ensure the model accurately classifies and identifies code. While it is acceptable if some text is misclassified as code, code must always be correctly identified, highlighting the need for high recall. The model effectively understands code structure and semantics, but it sometimes confuses text that appears in both code and natural language. Despite this, the model achieved an average accuracy of 91.4% and an average recall of 90.4%. Future work will focus on improving the model’s ability to distinguish between these contexts more accurately.

5 Conclusion

New AI tools offer more straightforward ways to share code but raise concerns about data leaks. Rather than imposing bans, firewalls can be implemented to safeguard sensitive information while allowing the advantages of AI. In this paper, a novel code-text filtering system was successfully developed to distinguish code from text and comments within documents efficiently. This system can later be integrated with a firewall to block the code portion from being transmitted to generative AI. The issue of encrypted communication, such as that in ChatGPT, introduces an additional layer

of complexity, possibly requiring SSL decryption or secure browser plugins. The focus should be on balancing adopting new technologies with maintaining security.

Acknowledgements We sincerely thank Bhabha Atomic Research Centre officials for allowing us to work on this fascinating research topic and providing the resources needed to make this project possible.

References

1. Kužina, V., Vušak, E., Jović, A.: Methods for automatic sensitive data detection in large datasets: a review. In: 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), pp. 187–192. IEEE (2021)
2. Ahmed, H., Traore, I., Saad, S., Mamun, M.: Automated detection of unstructured context-dependent sensitive information using deep learning. *Internet Things* **16**, 100444 (2021)
3. Guha, A., Samanta, D., Banerjee, A., Agarwal, D.: A deep learning model for information loss prevention from multi-page digital documents. *IEEE Access* **9**, 80451–80465 (2021)
4. Peng, C.: Deep learning based sensitive data detection. In: 2022 19th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 1–6. IEEE (2022)
5. Shi, J., Cui, S., Chen, F., Wang, C.: Sensitive data discovery technology based on artificial intelligence. In: Proceedings of the 2nd International Conference on Information Economy, Data Modeling and Cloud Computing, ICIDC 2023, June 2–4, 2023, Nanchang, China (2023)
6. Zhang, K., Jiang, X.: Sensitive data detection with high-throughput machine learning models in electrical health records. In: AMIA Annual Symposium Proceedings, vol. 2023, p. 814. American Medical Informatics Association (2023)
7. Rehan, H.: AI-driven cloud security: the future of safeguarding sensitive data in the digital age. *J. Artif. Intell. Gen. Sci. (JAIGS)* **1**(1), 132–151 (2024). ISSN: 3006-4023
8. Templ, M., Sariyar, M.: A systematic overview on methods to protect sensitive data provided for various analyses. *Int. J. Inf. Secur.* **21**(6), 1233–1246 (2022)
9. Wen, R., Crowe, S.E., Gupta, K., Li, X., Billingham, M., Hoermann, S., Allan, D., Nassani, A., Piumsomboon, T.: Large language models for automatic detection of sensitive topics. [arXiv: 2409.00940](https://arxiv.org/abs/2409.00940) (2024)
10. Petrolini, M., Cagnoni, S., Mordonini, M.: Automatic detection of sensitive data using transformer-based classifiers. *Future Internet* **14**(8), 228 (2022)
11. Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., et al.: Codebert: a pre-trained model for programming and natural languages. [arXiv:2002.08155](https://arxiv.org/abs/2002.08155) (2020)

Enhancing Data Integrity: Unveiling the Potential of Reversible Logic for Error Detection and Correction



Premanand Kadbe , Shriram Markande , and Manisha Waje 

Abstract Data integrity protection is achievable using error detection and correction techniques with reliable and efficient computing. Traditional methods are resource-intensive and often irreversible, limiting efficiency. This research explores reversible logic as a solution to these limitations. Reversible circuits can detect errors and reset to the original state, mitigating data corruption. Integrating reversible logic into computer architectures presents challenges and prospects. Error detection and correction techniques are essential to ensure accurate and reliable calculations in modern electronic systems. The traditional error detection and correction methods often rely on complex and power-hungry algorithms. However, reversible logic has emerged as a promising model that offers unique advantages such as low power consumption and information retention. This article explores the basics of reversible logic, discusses the challenges associated with error detection and correction, presents the most advanced methods, and highlights the potential future directions in this field.

Keywords Error detection and correction · Data integrity · Reversible logic · Cryptography · Quantum computing · Vulnerability first section

1 Introduction

In modern computer systems, data integrity is a significant concern. The data is vulnerable to errors during storage and transmission, and robust error detection and correction techniques are needed. However, the conventional approaches available are struggling with resource-intensive procedures and irreversible operations that

P. Kadbe (✉) · M. Waje

Department of Electronics and Telecommunication Engineering, G.H.Raisoni College of Engineering and Management, Wagholi, Pune, India
e-mail: premanand.kadbe.phdetc@ghrcem.raisoni.net

S. Markande

Department of Electronics and Telecommunication Engineering, Sinhgad Institute of Technology and Science (SITS, Pune, India

hinder their overall effectiveness. Reversible logic has emerged as a promising model for efficient error detection and correction while minimizing resource usage. Hence, delve into the reversible logic field, error detection, and correction, and present the innovative approach with data integrity protection.

It is possible to design efficient error detection and correction circuits by exploiting the unique properties of reversible logic gates [1]. Although the challenges still exist, the transformative impact of inversion logic on error detection and correction warrants further exploration and study. The application of reverse logic offers a promising path to data integrity improvement and error detection and correction advancement in computer systems, paving the way for efficient and reliable technology.

Error detection and correction are essential to ensure accurate and reliable calculations in modern electronic systems. The faults can occur due to various factors, including noise, faulty components, environmental disturbances, or even inherent imperfections in the manufacturing process. The traditional error detection and correction methods often rely on complex algorithms and additional hardware, increasing power consumption and circuit complexity [2]. Reversible logic has become a promising model offering unique error detection and correction advantages. Reversible logic gates are a special type of gate that exhibits bi-directionality, meaning they have a one-to-one mapping between input and output states. Unlike the conventional irreversible gates, the reversible gates can perfectly recover input from output, making them ideal for error correction.

The benefit of reverse logic extends beyond error correction. The reversible circuits can significantly reduce power consumption due to their inverting nature since no information is lost during the computations. This energy-conserving property makes reversible logic attractive, particularly for low-power and power-constrained applications such as mobile devices, embedded systems, and quantum computing.

Data integrity is the certainty that the data remains accurate, reliable, and unaltered throughout its life cycle and is crucial in various areas, including finance, healthcare, and critical infrastructure. The reversible logic's inherent ability to preserve information offers innovative solutions to safeguard data against corruption, errors, and unauthorized alterations. This research article aims to provide a broad review of error detection and correction techniques using reversible logic. We will explore the basics of reversible logic, discuss error detection and correction challenges, present the most advanced methods, and highlight potential future directions. By analyzing and synthesizing existing knowledge, this article sheds light on the potential of reversible logic as an efficient and energy-efficient solution for fault detection and correction in electrical systems [3]. The article explains the uses of reversible logic, especially in data integrity in encryption, watermarking and steganography, error detection, and correction. Further, this article also describes the advantages and disadvantages of reversible logic in data integrity. The next part details the state-of-the-art approaches used for data integrity using reversible logic.

2 Reversible Logic in Encryption

Reversible logic is utilized in Quantum Key Distribution protocols to ensure the bidirectional and information-preserving exchange of encryption keys and enhance the integrity of the key exchange process. Quantum Key Distribution (QKD) is a revolutionary cryptographic technique that leverages the principles of quantum mechanics to enable secure communication and the exchange of encryption keys. Unlike classical key exchange methods, which are vulnerable to eavesdropping, QKD offers provable security by exploiting the unique properties of quantum states. It is pivotal in ensuring data integrity and confidentiality in secure communication [4].

The Quantum Key Distribution relies on several key components and principles to achieve its security objectives. QKD uses quantum states, typically in the form of photons, to encode information. These quantum states are susceptible to measurement disruptions, which makes eavesdropping detectable. The quantum entanglement creates correlated quantum states between the sender (Alice) and the receiver (Bob). Any eavesdropper (Eve) attempt to intercept the quantum states disrupts their entanglement, revealing her presence. Alice sends quantum states to Bob, who measures them. The choice of measurement basis is communicated publicly; however, the measurements made are kept secret until a later step. Alice and Bob compare a subset of their measurement results to assess the error rate. A high error rate indicates the presence of interference or eavesdropping. After error rate monitoring, Alice and Bob publicly announce their measurement bases and discard measurement results where the bases do not match. They use the remaining correlated results to derive a secure encryption key.

2.1 Data Integrity Enhancement Using QKD

QKD contributes to data integrity by guaranteeing the confidentiality and integrity of the encryption key used for secure communication. QKD's fundamental security lies in its ability to detect eavesdropping attempts. Any unauthorized interception or measurement of quantum states disturbs their entanglement, leading to a noticeable increase in the error rate. This detection mechanism ensures that encrypted data remains confidential and unaltered during transmission. QKD provides encryption keys that are immune to attacks by quantum computers. This feature safeguards encrypted data against potential threats posed by future quantum computing advancements. Using quantum states and entanglement in QKD ensures that even subtle alterations or tampering with the transmitted quantum states are detectable. This tamper detection mechanism reinforces data integrity by preventing unauthorized changes to the encrypted information. QKD generates a fresh encryption key for each communication session. Even if a key from a previous session were compromised, the data transmitted in that session would remain confidential and unaltered, ensuring forward secrecy.

QKD has numerous practical applications beyond secure communication. QKD can be used to protect financial transactions and assure the confidentiality and integrity of sensitive financial data. Government and military agencies use QKD to safeguard classified information and communications. Medical institutions can use QKD to protect patient records and sensitive healthcare data. QKD ensures the security of critical infrastructure systems, preventing unauthorized access and data tampering. QKD can enhance the integrity of electronic voting systems, ensuring the confidentiality and authenticity of votes.

Quantum Key Distribution is a groundbreaking technology that provides secure communication and significantly contributes to data integrity by protecting encryption keys against eavesdropping and quantum attacks. Its applications span various sectors where data confidentiality and integrity are paramount.

The Quantum-inspired synthesis techniques draw inspiration from quantum computing principles and algorithms. They leverage concepts like quantum gates, quantum circuits, and quantum cost models for reversible logic circuit synthesis. The Quantum-inspired synthesis techniques exploit the quantum superposition and entanglement properties in reversible circuit optimization for specific applications [5].

2.2 Homomorphic Encryption

Reversible logic contributes to the efficiency of homomorphic encryption schemes, enabling computations of encrypted data without compromising data integrity. Homomorphic encryption is an advanced cryptographic technique that allows calculations to be performed on the encrypted data without decrypting it initially. This breakthrough in cryptography has profound implications for enhancing data privacy and security while preserving data integrity. To understand how homomorphic encryption enhances data integrity, let's explore its key concepts. Homomorphic encryption begins with data encryption, much like traditional encryption methods. However, unlike conventional encryption, homomorphic encryption permits certain mathematical operations to be performed on the encrypted data without betraying its listing. "homomorphism" refers to the mathematical property of preserving the relationships between data elements. In homomorphic encryption, the operations performed on encrypted data will yield meaningful results when the data is decrypted. Homomorphic encryption supports various mathematical operations, including addition, multiplication, and more, depending on the homomorphic scheme used [6].

2.2.1 Enhancing Data Integrity with Homomorphic Encryption

Homomorphic encryption enhances data integrity in several ways: privacy, secure outsourcing, data aggregation, and secure computation. Homomorphic encryption

ensures that the data remains confidential even when computations are performed. This privacy protection is essential for maintaining the integrity of sensitive information. Organizations can securely outsource data processing tasks to third parties like cloud service providers without revealing the content of the data. This reduces the risk of data exposure or tampering. Homomorphic encryption enables secure data aggregation, allowing multiple parties to collectively analyze data without sharing the raw information. This aggregation process maintains data integrity by preventing unauthorized access to individual data points. The encrypted data can undergo computations on remote servers without decryption. This secure computation ensures that data remains intact and unaltered during processing [7].

While homomorphic encryption offers significant advantages for data integrity, it also presents challenges like computational overhead, complexity, limited supported operation, and key management. Performing operations on encrypted data is computationally intensive, potentially slowing down data processing. Implementing and managing homomorphic encryption systems can be complex and require specialized expertise. The set of supported mathematical operations depends on the specific homomorphic encryption scheme, and not all operations may be feasible. Proper key management is crucial to guarantee the security and integrity of encrypted data. The loss or compromise of encryption keys may lead to data loss or exposure.

Homomorphic encryption has applications in various domains, including healthcare, financial services, data analytics, secure voting systems, etc. These can be protecting patient data during medical research and analysis while preserving data integrity, securely analyzing financial transactions for fraud detection without exposing sensitive information, performing privacy-preserving data analytics on encrypted data to maintain data integrity, and ensuring the integrity and privacy of electronic voting systems.

Homomorphic encryption stands at the forefront of data privacy and security technologies, allowing organizations to perform operations on encrypted data while preserving its integrity and confidentiality. Its applications continue to expand as data privacy concerns grow in our increasingly digital world. The reversible logic, which focuses on information preservation and bidirectional operations, significantly enhances data integrity in the encryption process. It ensures that data remains unaltered during encryption and decryption, detects errors, and provides a foundation for quantum-resistant encryption methods, reinforcing the overall security and trustworthiness of encrypted data.

3 Watermarking and Steganography Using Reversible Logic

Watermarking and steganography are techniques used to embed hidden information within digital media while preserving data integrity and concealing hidden data. These techniques become even more robust and secure when coupled with reversible

logic, ensuring original data remains unaltered during the embedding and extraction processes.

3.1 Watermarking Using Reversible Logic

Watermarking is a technique used to embed hidden information known as a watermark into digital media. The goal is to add this watermark so that it is imperceptible to the users but can still be extracted later for various purposes, including copyright protection, authentication, and data integrity verification. When combined with reversible logic, watermarking becomes a powerful tool for ensuring data integrity while embedding hidden information [8].

Reversible logic is particularly valuable in watermarking because it can preserve data integrity and reversibility, which are crucial aspects of the watermarking process. The reversible logic gate ensures that the original digital media, often called the host or cover data, remains unchanged during the watermark embedding process. This preservation of the data integrity of the host is essential. The reversible logic allows for both the embedding of the watermark into the host data and the extraction of the watermark from the watermarked data. This bidirectional capability ensures that the original host data can be fully restored. The watermarks created using reversible logic are robust against various attacks and transformations the watermarked data may undergo, like compression, cropping, or noise addition. The watermark can be accurately extracted even in the presence of such alterations.

3.1.1 Use Cases of Watermarking with Reversible Logic

Reversible logic-based watermarking finds applications in various domains, including copyright protection, authentication, digital forensics, and data integrity verification. Content creators and media producers use reversible logic watermarking to embed copyright information, ownership details, or licensing data into their digital media. This allows for identifying copyrighted materials and facilitates legal enforcement against unauthorized use. The watermarks created using reversible logic can serve as the authentication markers. They provide a means to verify digital media's authenticity and integrity, ensuring that it has not been tampered with or altered. In digital forensics, the reversible logic watermarking can embed the hidden information for tracking purposes. This can aid in investigations and the tracking of digital assets. The watermarks can carry information about the digital media's original state, enabling data integrity verification. This is particularly useful in situations where tampering or unauthorized alterations are of concern.

While reversible logic-based watermarking offers substantial benefits, challenges and considerations include capacity versus data integrity, security, detection, and algorithm selection. Balancing the amount of hidden information (watermark

capacity) with preserving data integrity is challenging. The high-capacity watermarking may result in noticeable alterations to the host data. Ensuring that the watermark remains secure and immune to unauthorized extraction or tampering is crucial, and detecting the presence of watermarks and extracting them accurately is an essential aspect of watermarking. The choice of the watermarking algorithm and its parameters can impact data integrity and the ability to extract the watermarks.

The reversible logic-based watermarking enhances data integrity by enabling the embedding and extraction of hidden information while preserving the integrity of the host data. It has diverse applications where secure communication, authentication, and data integrity are paramount. The reversible logic ensures that the original multimedia content remains unchanged during the watermark embedding, which is critical for maintaining data integrity. The reversible logic allows for bidirectional embedding and extraction of the watermarks. The ability to reverse the watermarking process ensures the original content is fully restored. The reversible watermarking can withstand attacks like compression and noise while enabling watermark extraction with very high accuracy. The watermarks created using reversible logic serve as the authentication markers, providing a way to verify the authenticity and integrity of the multimedia content.

3.2 Steganography Using Reversible Logic

Steganography is the art of hiding secret information within innocuous carrier files so that the presence of hidden data is not apparent. Steganography conceals secret information within seemingly innocuous carrier files like images, audio, or text in a way that does not arouse suspicion. When combined with reversible logic, steganography becomes a powerful tool for hiding information while ensuring data integrity and extracting hidden data without any loss.

The reversible logic enhances steganography in several ways: information preservation, bidirectional embedding extraction, and robustness. The reversible logic ensures the original carrier file remains unchanged despite hidden data. The reversible logic allows for embedding secret data into the carrier file and extracting hidden data from the steganographic data. This bidirectional capability ensures that the original carrier file can be fully restored. The steganography using reversible logic tends to be robust against various attacks and transformations that the steganographic data may undergo, such as compression, format conversion, or noise addition. The hidden data can be accurately extracted even in the presence of such alterations [9].

3.2.1 Use Cases of Steganography with Reversible Logic

Steganography with reversible logic has numerous practical applications like secure communication, covert data transfer, concealment, authentication, and data integrity. Individuals and organizations can use reversible logic-based steganography to

communicate sensitive information covertly, ensuring that the data remains confidential while appearing ordinary. In situations where transmitting sensitive information openly is impossible, reversible steganography allows data to be transferred concealed, bypassing detection. The data can be hidden within the files or messages without arousing suspicion. For example, sensitive text or files can be embedded within images or audio recordings. The reversible logic-based steganography serves as a means of authentication and data integrity verification. The presence of a hidden message is used to confirm that the carrier file has not been tampered with.

The reversible logic-based steganography allows hidden embedding and extraction of information, preserving the integrity of the carrier file. It has diverse applications in the fields where secure communication, data concealment, and data integrity verification are of predominant importance. Steganography becomes highly effective in maintaining data integrity when combined with reversible logic. The reversible logic ensures that the carrier file remains unchanged despite hidden data, which is crucial for data integrity. The reversible logic gates enable the bidirectional embedding and extraction of the secret data. This reversibility ensures that the original carrier file can be fully restored. The reversible steganography provides secure communication, where sensitive information can be hidden within seemingly innocuous files. The reversibility ensures that the original content is recoverable by authorized parties. Like watermarking, steganography can serve as an authentication and data integrity verification method. Steganography is used in various domains, including information security, covert communication, and digital forensics.

While the use of reversible logic in watermarking and steganography offers significant advantages in terms of data integrity, there are some challenges and considerations. Striking a balance between the amount of hidden data capacity and preserving data integrity can be challenging. High-capacity embedding may result in noticeable alterations to the carrier file. Ensuring that the secret information remains secure and immune to unauthorized extraction or tampering is somewhat crucial. Detecting the presence of watermarks or hidden data, especially in steganography, is a critical aspect of these techniques. Choosing a watermarking or steganography algorithm and its parameters can impact the data integrity and ability to extract hidden information.

The reversible logic enhances watermarking and steganography by ensuring the preservation of data integrity and bidirectional embedding or extraction. These techniques find applications in various domains requiring secure communication, authentication, and data integrity.

4 Reversible Logic: A Foundation for Error Detection and Correction

As reversible logic allows the original values to be recovered from the output, it differs from the irreversible counterparts. This makes it easier to detect errors by comparing the original input and the received output, utilizing the essential characteristics of

the reversible gates. If there is any difference, then it indicates the presence of a bug in the system.

4.1 Reversible Logic and Data Integrity

Reversible logic can be used to detect errors by creating some parity bits or checksums that represent the state of the data. Any changes in the data are reflected in the parity bits, allowing errors to be detected. In addition to detection, reversible logic also enables error correction. The erroneous bits can be corrected by comparing the detected errors with the generated parity bits, restoring the original data. In storage systems, such as hard drives or SSDs, the reversible logic can detect and correct errors due to physical media degradation or noise during data retrieval. In communication networks, reversible logic-based error detection and correction ensures that data remains accurate during transmission, even in noise and interference. In critical space missions, where data transmission is susceptible to cosmic rays and other environmental factors, the reversible logic ensures that data sent from space probes is received accurately on Earth.

The emergence of quantum computers poses new challenges to error detection and correction, and research into post-quantum error correction using reversible logic is ongoing. Implementing reversible logic-based error detection and correction can be complex and requires specialized hardware and algorithms. Developing industry standards for integrating reversible logic in error detection and correction techniques ensures interoperability.

Reversible logic's inherent ability to preserve information provides a solid foundation for error detection and correction techniques. By leveraging reversible computing, data integrity can be maintained in critical applications, ranging from data storage systems to space exploration. As technology continues to evolve, the role of reversible logic in error detection and correction remains essential to ensure the reliability and accuracy of digital data.

4.2 Data Encryption

Data encryption is a fundamental concept in information security designed to protect the confidentiality and integrity of the data. It involves transforming plain text data, which means the original, = but readable data, into cipher text data, which means encrypted data using an encryption algorithm and an encryption key. While encryption primarily addresses data confidentiality, preserving data integrity is equally critical in the encryption process.

An encryption algorithm is a set of mathematical rules and operations that convert plain text data into cipher text and vice versa. It ensures that the encryption and decryption processes are reversible. The encryption key is a secret information the

encryption algorithm uses to perform the encryption and decryption operations. The choice of the key determines the security of the encryption process. Plain text data represents the original data that needs to be protected. This can include text, files, messages, or digital information. Cipher text data is the result of encrypting plain text data using the encryption algorithm and the encryption key. It appears as random, unreadable characters and is meant to be secure from unauthorized access [10].

4.2.1 Data Integrity in Encryption

While data encryption primarily focuses on data confidentiality, it also significantly impacts data integrity. The encryption process itself should not introduce errors or data corruption. The data integrity ensures that the encrypted data accurately represents the original plain text data. Here, encryption helps protect data from unauthorized modification during transmission or storage. Any tampering with the cipher text can be detected when data integrity is maintained. Hence, maintaining data integrity during encryption involves carefully handling the encryption process. The encryption algorithm should be reversible, meaning that decryption should accurately recover the original plain text from cipher text without any loss or corruption of data. Adding cryptographic authentication mechanisms like digital signatures or message authentication codes can further ensure data integrity by detecting any unauthorized changes in the data. The proper key management practices, including key generation, storage, and distribution, are essential to prevent data loss or corruption due to unauthorized access to encryption keys. Data integrity is integral to encryption, ensuring the encrypted data remains accurate and unaltered throughout its life cycle. In the following sections, we will explore how reversible logic, with its information-preserving properties, enhances data integrity in the encryption process.

4.3 Error Correction with Reversible Logic

Beyond error detection, reversible logic enables error correction, a crucial aspect of preserving data integrity. When errors are detected, reversible reasoning can pinpoint the location of the errors in the data. Using the information about error locations, the reversible logic can reconstruct the original error-free data, and this reconstruction ensures that data integrity is maintained. The unique properties of reversible logic, including its bidirectional nature and information preservation, make it an ideal foundation for error detection and correction techniques. By leveraging these properties, organizations can detect and rectify errors in critical data, ensuring that it remains accurate and reliable in the face of various challenges to data integrity [11].

4.4 Error Detection in Data Storage Systems

Reversible logic is applied in data storage systems, such as hard drives and solid-state drives, to detect and report errors. The parity bits or checksums generated using reversible logic are stored along with the data. During the data retrieval, these bits detect errors resulting from factors like media degradation or noise. Detected errors trigger error reporting mechanisms, allowing for proactive data management and maintenance. This ensures that data remains reliable even in long-term storage. Beyond detection, the reversible logic enables error correction in data storage. The reversible logic is used to identify the location of erroneous bits within stored data. Based on error locations, the reversible logic reconstructs the original data, correcting errors and preserving data integrity [12].

4.5 Error Detection in Communication Networks

Reversible logic is instrumental in ensuring data integrity in communication networks. Data packets transmitted over networks often include parity bits generated through reversible logic. These parity bits are used at the receiving end for real-time error detection. Any errors introduced during transmission are promptly identified, and in cases of detected errors, data packets can be re-transmitted to ensure that the correct data is received. In addition to error detection, the reversible logic enables error correction in communication networks. The error locations within data packets are identified using reversible logic, facilitating efficient error correction. The original error-free data is reconstructed to ensure the information is accurate.

4.6 Error Resilience in Space Missions

Space missions are highly susceptible to cosmic rays and other environmental factors that can corrupt data during transmission. Reversible logic is crucial in detecting and correcting errors in data transmitted from space probes to Earth, ensuring that mission-critical data such as images and telemetry is received accurately and without corruption. These practical applications illustrate the versatility and importance of reversible logic in preserving data integrity in various domains. Whether safeguarding data in storage systems, ensuring reliable communication over networks, or supporting space exploration missions, the reversible logic provides a foundation for error detection and correction techniques essential for maintaining digital information's accuracy and reliability [13].

5 Utilizing Reversible Circuits for Error Detection

The efficient detection of errors with reversible circuits incorporates error detection modules to analyze the output and identify the inconsistencies. The advantage of the feedback mechanism is that it restores the faulty bits to their original state to provide error detection.

5.1 *Definition and Properties of Reversible Logic Gates*

Reversible logic gates are a special type of logic gate with a dual property, meaning that each input combination corresponds to a single output combination and vice versa. Unlike the conventional irreversible gates, the reversible gates can retrieve input from output without loss of information. This property is obtained by ensuring that the number of input and output bits are the same. Some commonly used reversible gates include Toffoli, Fredkin, Peres, and Feynman gates. These gates play an essential role in reversible logic synthesis, where reversible circuits are designed using a combination of reversible gates to perform desired functions.

5.2 *Reversible Logic Synthesis Techniques*

The reversible logic synthesis involves transforming a given irreversible logic circuit into an equivalent reversible logic circuit. Various synthesis techniques have been developed to achieve this transformation. A common approach is gate-level synthesis, which aims to decompose the irreversible circuit into a network of reversible gates. This method uses reversible gate libraries and algorithms such as the Universal Reversible Logic Gate (URLG) library or BDD-based synthesis techniques to implement circuits efficiently. Another technique is permutation-based synthesis, which exploits the inherent permutation properties of reversible gates. By representing the desired logical function as a permutation matrix, the synthesis process focuses on finding a sequence of reversible gates to achieve the desired permutation. Reversible logic has applications in several fields due to its energy and information-saving nature.

5.3 *Quantum Calculation*

The reversible logic is essential to quantum computing because it maintains the consistency and reversibility of quantum states. The quantum gates such as ControlledNOT gate and Toffoli gate are reversible gates commonly used in quantum

circuits. Quantum computation is a revolutionary paradigm that exploits the principles of quantum mechanics to execute calculations exponentially faster than classical computers. Reversible logic plays a pivotal role in quantum computation, enabling the creation of efficient quantum circuits that adhere to the fundamental principles of quantum mechanics. This section explores the diverse applications of reversible logic in quantum computation, from quantum gates to algorithm development and quantum error correction.

5.4 Quantum Algorithms and Reversible Logic

This section explores reversible logic applied in developing quantum algorithms, showcasing its critical role in efficiently solving complex problems.

5.4.1 Shor's Algorithm

Shor's Algorithm is a groundbreaking quantum algorithm developed by mathematician Peter Shor in 1994. It is designed to solve one of the most challenging problems in classical computing, which is called integer factorization. This algorithm catapulted quantum computing into the spotlight due to its potential to efficiently factor large numbers, which has significant implications for cryptography and security. The integer factorization involves finding the prime numbers that multiply together to form a given composite number. It may seem like a straightforward mathematical problem, but it becomes exponentially more complex as the size of the number to be factored increases. This inherent difficulty forms the basis for many cryptographic systems, including RSA, which relies on the difficulty of factoring the product of two large prime numbers [14].

Shor's Algorithm leverages the power of quantum parallelism and quantum Fourier transforms to factor large numbers efficiently. Shor's algorithm begins with selecting a random integer, the common factor for factoring the number. Quantum parallelism allows multiple factors to be tested simultaneously. Shor's Algorithm employs a quantum Fourier transform to create superposition states of different periodicity values where the reversible logic plays a significant role. The reversible gate ensures the transformation can be undone, allowing emerging quantum interference patterns. The quantum state is measured, collapsing it to a specific period value. The classical post-processing steps extract the factors from the estimated period.

The reversible gates create superposition states, allowing multiple candidate values to be tested simultaneously. The quantum Fourier transform is a central component of Shor's Algorithm, and its reversibility ensures that the transformation can be reversed if needed. By efficiently factoring the large numbers, Shor's Algorithm has the potential to break widely used cryptographic schemes. The reversible logic ensures that the algorithm's operation can be performed in both forward and reverse directions, facilitating the factorization process. Shor's Algorithm showcases the

remarkable capabilities of quantum computation and its' reliance on reversible logic to perform complex calculations efficiently. It also highlights the need for post-quantum cryptographic solutions to maintain data security in a world with quantum computers, which have the potential to revolutionize cryptography.

5.4.2 Grover's Algorithm

Grover's algorithm, proposed by Lov Grover in 1996, is another significant quantum algorithm that focuses on a different problem, such as an unstructured search. Its ability to search an unsorted database or perform an exhaustive search is significantly faster than classical algorithms. The reversible logic plays a crucial role in the implementation and operation of Grover's Algorithm. The unstructured search problem involves finding a specific item in an unsorted database of N items. Classical computing typically requires searching through all N items, which takes $O(N)$ time. Grover's Algorithm, however, achieves a quadratic speedup, allowing it to perform the search in roughly $O(\sqrt{N})$ time.

Grover's Algorithm relies on quantum parallelism and the principle of amplitude amplification to enhance search efficiency. The quantum state is initialized to a superposition of all possible states. The reversible logic gates facilitate the creation of this superposition state. The oracle function is essentially a black box that flips the sign of the amplitude of the target item, which is used to mark the target item. Amplitude amplification is achieved through iterations of two primary operations, the Grover Diffusion Operator and the Oracle function. The Grover Diffusion Operator enhances the amplitude of the correct solution and decreases the amplitudes of incorrect solutions. After sufficient iterations, a measurement is performed to collapse the quantum state. The measured state is more likely to be the target item.

The reversible gates allow the creation of superposition states where all possible states are considered simultaneously. This enables quantum parallelism and speeds up the search process. The oracle function used in Grover's Algorithm is reversible. It reverses the amplitude of the target item, making it distinguishable during the amplification process. The Grover Diffusion Operator is a key part of amplitude amplification, which relies on reversible logic gates to enhance the amplitude of the correct solution and decrease the amplitude of incorrect solutions. This step significantly improves the algorithm's efficiency.

The Grover's Algorithm demonstrates the power of quantum computation in solving complex problems efficiently. It highlights the importance of reversible logic in quantum algorithms by enabling the creation of superposition states, reversible oracle functions, and amplitude amplification, contributing to the remarkable speedup of algorithms in unstructured search tasks. The use of reversible logic in quantum simulation enables efficient quantum systems modeling.

6 Advantages of Reversible Logic in Data Integrity

Reversible logic underpins the efficient operation of quantum gates and algorithms and plays a crucial role in quantum error correction and the advancement of quantum cryptography. As quantum computing continues to evolve, the integration of reversible logic remains pivotal in harnessing the full potential of this revolutionary technology.

6.1 Low Consumption

The energy-conserving property of inversion logic makes it very attractive for low-power computing. It has been used in low-power arithmetic and data processing units, reversible microprocessors, and energy-efficient digital signal processing. The reversible logic has potential applications in nano-scale devices and nanocomputers. The reversible nature of these circuits can mitigate the effects of inherent quantum noise and reduce power consumption, making them suitable for emerging nano-scale technologies.

6.2 Energy and Power Advantages of Reversible Logic

One of the key advantages of the reversible logic is its energy efficiency. In conventional irreversible circuits, the bits of information are irreversibly lost during computation, resulting in energy dissipation. Conversely, the reversible logic ensures that no information is lost, leading to lower energy consumption. The power advantages of the reversible logic stem from its underlying principles. As the reversible gates have a one-to-one mapping between inputs and outputs, they do not produce heat due to information loss. Consequently, reversible circuits generate less power dissipation, making them highly desirable for applications where power consumption is critical. The reversible logic offers several benefits, including designing energy-efficient circuits and preserving information during computation. These advantages make reversible logic an attractive error detection and correction paradigm, enabling accurate and reliable calculation while minimizing power consumption.

6.3 Error Detection Techniques Using Reversible Logic

Error detection is essential in ensuring data integrity and computation integrity in electronic systems. Several techniques have been proposed for error detection within the framework of reversible logic. These techniques focus on identifying

and reporting errors that can occur during the computation in reversible circuits. These reversible logic error correction techniques provide mechanisms for detecting and correcting errors that may occur during the computation. Accurate and reliable calculations can be achieved by integrating error correction into the reversible circuits while minimizing the impact of mistakes.

6.4 Advantages, Challenges and Limitations

Several challenges need to be addressed despite the promises in reversible logic. The efficient reversible circuit design poses a significant obstacle due to less complexity and fewer gates. The limited number of gates and costs associated with error detection cause additional barriers that must be managed carefully. Although the error detection and correction techniques using reversible logic offer significant advantages, some challenges and limitations must be addressed here.

7 Advantages of Reversible Logic-Based Error Detection and Correction

Several notable advantages exist of integrating reversible logic for error detection and correction. The reversible circuits allow precise error determination, improving data integrity and resorting faulty bits to their original state, ensuring high fault tolerance. This makes it an energy-efficient solution by minimizing the power dissipation associated with error correction.

7.1 Trade-Offs Between Error Detection/Correction Capabilities and Circuit Complexity

Increasing reversible logic circuits' error detection and correction capabilities often results in larger circuit sizes and increased complexity. Also, adding the auxiliary bits and ports needed for error detection and correction can result in higher resource usage and longer computation times. The designers must carefully consider these tradeoffs to balance error range and circuit complexity.

7.2 *Error Propagation and Its Impact on Reversible Logic Circuits*

The errors in reversible logic circuits can propagate and affect the subsequent calculations. The error propagation can lead to incorrect output values, making it difficult to detect and correct errors. Hence, developing effective techniques for analyzing and minimizing error propagation is essential for robust error detection and correction in reversible logic circuits.

7.3 *Fault Tolerance and Error Correction Overheads*

The error-tolerant reversible circuits and error correction techniques introduce additional costs regarding auxiliary bits, ports, and computational resources. These overheads can affect the overall circuit performance, power consumption, and area utilization. Hence, the task is to balance the benefits of error correction with the costs involved in an ongoing research challenge. Addressing these challenges and limitations is crucial in advancing the field of error detection and correction with reverse logic. Future research efforts should focus on developing innovative techniques that improve error detection and correction while minimizing circuit complexity and cost. Overall, error detection and correction using reversible logic offers promising opportunities for achieving accurate and reliable computations in electronic systems. Thus, taking advantage of the unique properties of reversible logic, such as energy efficiency and information retention, significant progress can be made in improving the reliability and durability of circuits and electronic systems.

8 State-of-the-Art Approaches

In recent years, significant research efforts have been devoted to detecting and correcting errors using reversible logic. Some of the most advanced approaches focus on improving error detection, optimizing error correction techniques, and discovering new ways to detect and correct errors based on reversible logic. The following section provides an overview of some notable modern approaches.

8.1 *Advanced Error Detection Techniques*

The researchers have proposed advanced error detection techniques that exploit unique properties of reversible logic. These techniques improve error coverage, reduce circuit complexity, and improve fault tolerance. One such approach is using

error detection codes based on reversible logic gates such as Peres and Toffoli gates. These codes use reversible gates to detect errors introduced during computation. These codes detect high errors while minimizing resource usage by carefully designing the encoding and decoding scheme. Additionally, the researchers explored machine learning techniques to detect the mistakes in reversible logic circuits. These models detect anomalous behavior and identify potential errors in real-time by training machine learning models on a large data set of reversible circuit simulations.

8.2 Enhanced Error Correction Techniques

The advances in error correction techniques using reversible logic focus on improving error correction and reducing associated costs. One technique worth noting is the development of optimized reversible error correction codes. Researchers have proposed a new encoding scheme that provides efficient error correction while minimizing the required bits and auxiliary ports. These optimized codes balance error correction and circuit complexity, making them suitable for practical implementation. In addition, advances in error correction algorithms and decoding techniques have contributed to more efficient error correction. Techniques such as syndrome-based decoding algorithms and neural network-based error correction algorithms have shown promising results in achieving high accuracy and low cost.

8.3 Integration with Emerging Technologies

Reversible logic-based error detection and correction techniques and emerging technologies, such as quantum computing and nanotechnology, have also been explored. In quantum computing, inversion logic is fundamental to detecting and correcting errors in quantum circuits. Researchers have been working on integrating reversible error detection and correction techniques into quantum circuits to improve the reliability and fault tolerance of quantum computing. In nanotechnology, reversible logic has shown potential for error detection and correction in nano-sized devices. The reversible nature of these circuits can minimize the effects of inherent quantum noise and improve the accuracy of calculations. Researchers have explored applying reverse logic-based error detection and correction techniques in nano-scale computing systems.

8.4 *Performance Evaluation Metrics*

Different performance evaluation metrics have been proposed to evaluate the effectiveness of error detection and correction techniques using inverse logic. The metrics such as error detection rate, error correction rate, circuit complexity (e.g., number of ports and auxiliary bits), power consumption, and area utilization are commonly used to evaluate the performance of error detection and correction techniques. These metrics provide information about the trade-off between error range, circuit complexity, and resource usage. Using these performance metrics, the researchers can compare different approaches, identify areas for improvement, and guide the development of error correction and detection techniques with more efficient and reliable errors by reversible logic. The most advanced methods to detect and correct errors using reversible logic have focused on improving error detection techniques, improving error correction, exploring integration with other technologies, and emerging and developing performance measures. These advances pave the way for more efficient and reliable reversible logic-based error detection and correction systems, improving the reliability and accuracy of electronic systems.

9 *Integration into Existing Computing Architectures*

Integrating existing computer architectures is essential to make the logic-based reversible debugging techniques possible for real-world applications. For seamless compatibility, this involves the proper interfaces and protocol development. Additionally, the adaptability and scalability of reversible logic techniques need to be considered across different computing platforms.

9.1 *Future Directions and Research Opportunities*

The error detection and correction field using reversible logic offers interesting avenues for future research and development. As reversible logic continues to grow, researchers can explore the following directions to advance the art in the field.

9.2 *New Error Detection and Correction Schemes*

The further exploration of new error detection and correction schemes using reversible logic is essential. The researchers were able to investigate the development of codes specifically suitable for reversible logic circuits, taking into account the unique properties and constraints of reversible gates. This includes discovering

new coding schemes, error detection algorithms, and error correction techniques, delivering improved performance, reduced complexity, and fault tolerance. In addition, integrating the concepts of quantum error correction and fault-tolerant computation into reversible logic opens new avenues for error detection and correction. The research in this area contributes to developing more efficient and reliable error detection and correction techniques for reversible circuits.

9.3 Integration with Emerging Technologies

Integrating reversible logic-based error detection and correction techniques with emerging technologies has great potential. The researchers can explore how reversible reasoning can be applied alongside quantum computing, neural simulation computing, and other emerging computing models to improve error resilience and overall system performance. In quantum computing, the reversible logic can design fault-tolerant quantum circuits with enhanced error detection and correction. Studying the interaction between reversible logic and quantum error correction codes leads to advances in fault-tolerant quantum computing. Furthermore, exploring reversible logic for error detection and correction techniques in nano-scale devices such as nano-computing and molecular computing contributes to developing electrical systems.

9.4 Power Optimization Techniques

Energy efficiency is a significant benefit of reversible logic, and the research focuses on developing power optimization techniques that are particularly suitable for error detection and correction circuits using reversible logic. This includes exploring new methods to minimize power consumption during active error detection and correction, optimizing the use of extra bits, and investigating low-power design approaches for reversible circuits.

9.5 Error Propagation Analysis and Mitigation

Understanding and minimizing the error propagation in inverting logic circuits is essential for improving error detection and correction. Future research may focus on developing effective techniques for analyzing fault propagation paths, identifying potential fault hot spots, and developing strategies to minimize the impact of error propagation. The accuracy and reliability of inverting logic circuits can be significantly improved by reducing error propagation.

10 Conclusion

The ability to detect and correct errors and restore data integrity explored reversible logic's potential. The error detection and correction using reversible logic offers promising opportunities for accurate and reliable calculations in electronic systems. With its energy-saving and information-preserving properties, the reversible logic provides a unique foundation for designing error detection and correction mechanisms that reduce power consumption and preserve data integrity.

This research paper has provided a comprehensive review of reversible logic for error detection and correction techniques. We discussed the fundamentals of reversible logic, including the definition and properties of reversible logic gates, reversible logic synthesis techniques, and their applications in various fields. We have explored various error detection techniques, such as parity-based and syndrome-based methods and fault-tolerant reversible circuits. In addition, we looked at error correction techniques, including reversible Hamming codes, reversible Reed Solomon codes, and reversible LDPC codes.

Furthermore, we have highlighted the challenges and limitations in this area, such as the trade-off between error detection/correction and circuit complexity, error propagation and usability considerations, and fault tolerance. We discussed the most advanced methods, including advanced error detection techniques, debugging techniques, integration of emerging technologies, and performance metrics.

To further develop the field of error detection and correction with reversible logic, we have identified several directions and future research opportunities. These include discovering new error detection and correction schemes, integrating reversible logic into emerging technologies, developing power optimization techniques, analyzing and minimizing error propagation, and detecting and correcting errors in inverted quantum circuits. We can improve the reliability, efficiency, and applicability of reversible logic error detection and correction techniques by addressing these research directions. This will contribute to developing more powerful and energy-efficient electronic systems, paving the way for future advances in computing and beyond.

References

1. Kadbe, P.K., Waje, M.G.: Error detection in fault-tolerant reversible circuit using Fredkin Gates. In: Kumar, A., Mozar, S. (eds.) ICCCE 2021. Lecture Notes in Electrical Engineering, vol. 828. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-7985-8_58
2. James, R.K., Shahana, T.K., Jacob, K.P., Sasi S.: Fault-tolerant error coding and detection using reversible gates. In: TENCON 2007 IEEE Region 10 Conference, pp. 1–4 (2007)
3. Raj, V., Janakiraman, S., Rajagopalan, S., Amirtharajan, R.: Security analysis of reversible logic cryptography design with LFSR key on 32-bit microcontroller. Microprocess. Microsyst. **84**, 104265. ISSN 0141-9331. <https://doi.org/10.1016/j.micpro.2021.104265>. (2021)

4. Jassem, Y., Abdullah, A.: Enhancement of Quantum key distribution protocol for data security in a cloud environment. *ICIC Int.* **11**, 279–288 (2020). <https://doi.org/10.24507/icicelb.11.03.279>
5. Shamir, R., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM* **21**(2), 120–126 (1978). <https://doi.org/10.1145/359340.359342>
6. Mark A.W., Ryan K.L.: Chapter 5-a guide to homomorphic encryption. In: Ko, R., Choo, K.-K.R. (eds.) *The Cloud Security Ecosystem*, Syngress, pp. 101–127 (2015). ISBN 9780128015957, <https://doi.org/10.1016/B978-0-12-801595-7.00005-7>
7. Kaur, S.V., Wasson, V.: Enhancement in homomorphic encryption scheme for cloud data security. In: 2015 9th International Conference on Next Generation Mobile Applications, Services and Technologies, Cambridge, UK, pp. 54–59 (2015). <https://doi.org/10.1109/NGMAST.2015.38>
8. Zhang, Z., Wu, L., Gao, S., et al.: Robust reversible watermarking algorithm based on RIWT and compressed sensing. *Arab. J. Sci. Eng.* **43**(979–992), 2018 (2018). <https://doi.org/10.1007/s13369-017-2898-z>
9. Debnath, B., Jadav, C.D., De, D.: Design of reversible gates-based image steganography using quantum dot cellular automata for secure nano-communications. *Int. J. Recent. Technol. Eng.* **8**(4), 10408–10420 (2019). <https://doi.org/10.35940/ijrte.d8400.118419>
10. Edidi, S.A., Marada, R., Khan, T.A.: Improving data integrity with reversible logic-based error detection and correction module on AHB-APB bridge. In: *International Conference on Recent Advances in Electrical, Electronics, Ubiquitous Communication, and Computational Intelligence (RAEEUCCI)*, Chennai, India, pp. 1–9 (2023). <https://doi.org/10.1109/RAEEUC CI57140.2023.10134491>
11. Burgholzer, L., Wille, R., Kueng, R.: Characteristics of reversible circuits for error detection. *Array* **14**, 100165 (2022). ISSN 2590-0056, <https://doi.org/10.1016/j.array.2022.100165>
12. Gerrar, N.K., Zhao, S., Kong, L.: Error correction in data storage systems using polar codes. *IET Commun.* **15**(14), 1859–1868 (2021). <https://doi.org/10.1049/cmu2.12197>
13. Al-Mualla, M.A., Canagarajah, C.N., Bull, D.R.: Introduction to error resilience. In: Al-Mualla, M.E., Canagarajah, C.N., Bull, D.R. (eds.) *Signal Processing and its Applications, Video Coding for Mobile Communications*, p. 203. Academic Press (2002). ISBN 9780120530793, <https://doi.org/10.1016/B978-0-12-053079-3.50023-8>
14. Montanaro, A.: Quantum algorithms: an overview. *Npj Quantum Inf.* **2**, 15023 (2016). <https://doi.org/10.1038/npjqi.2015.23>

Enhancing Cyber Security Through Reversible Logic



Premanand Kadbe , Shriram Markande , and Manisha Waje 

Abstract Due to cyber threats' increasing complexity and severity, reversible logic has gained significant attention as a promising computing paradigm with the potential for energy efficiency and low power consumption in recent years. This article examines the potential advantages of reversible logic in cyber security. On enhancing cyber security solutions, we explored its theoretical foundations, implementation methods, and potential impact. Additionally, we discussed the integration of reversible logic with existing security mechanisms. The comprehensive review findings accentuate the value of reversible logic as a secure and efficient tool for developing cybersecurity solutions. This research article intends to impart present knowledge by comprehensively reviewing reversible logic and its potential applications in cyber security. This article sheds light on the promising avenues for utilizing reversible logic to develop robust and efficient cybersecurity solutions by exploring the theoretical foundations, implementation methodologies, integration strategies, advantages, and challenges. Finally, the aggregation presented in this article can inspire further research in this field.

Keywords Reversible logic · Cyber security · Cryptography

P. Kadbe (✉) · M. Waje

Department of Electronics and Telecommunication Engineering, G.H. Raisoni College of Engineering and Management, Wagholi, Pune, India
e-mail: premanand.kadbe.phdetc@ghrcem.raisoni.net

S. Markande

Department of Electronics and Telecommunication Engineering, Sinhgad Institute of Technology and Science (SITS), Pune, India

1 Introduction

The speedy growth of technology and the internet brings many benefits along with new challenges, especially in the area of cyber security in today's world. Cyber threats, hacking, data breaches, and identity theft pose significant risks to individuals, organizations, and nations. They often cannot keep up with increasingly sophisticated cyber attacks, while traditional security measures are somewhat effective. The Reversible logic, a computing paradigm that allows computation without loss of information, is emerging as a potential solution for improving cyber security. The reversible logic gates ensure that all input information can be unambiguously recovered from the output, unlike the traditional irreversible gates, allowing precise control and verification of computations. The application of reversible logic in cyber security removes a few limitations of conventional security measures and promises innovative solutions for protecting sensitive information. The main intention of this research is to explore the potential for enhancing cyber security measures by examining the theoretical foundations, implementation methods, integration strategies, benefits, and challenges associated with reversible logic [1].

The article is organized as follows. Section 2 briefly overviews the theoretical foundation of reversible logic, including reversible gates, circuits, synthesis techniques, and computational models. Moreover, Sect. 3 explores different implementations of reversible logic, focusing on design techniques, optimization algorithms, and performance evaluation approaches. In contrast, Sect. 4 discusses various cyber security applications of reversible logic, including encryption, authentication mechanisms, intrusion detection, and secure data storage. Section 5 discusses integrating reversible logic into existing security mechanisms and describes synergies between reversible logic and traditional security measures. Section 6 analyzes the benefits and challenges of reversible logic-based cybersecurity solutions, including energy efficiency, scalability, and potential security vulnerabilities.

Section 7 presents the case studies and experimental results to showcase the practical application of reversible logic in real-world cyber security scenarios. Finally, Sect. 8 concludes the article by summarizing the key findings and emphasizing the significance of reversible logic in enhancing cyber security while providing recommendations for future research and practical implementations.

2 Reversible Logic: Theoretical Foundations

The reversible logic gates are primal building blocks that enable computations to be carried out in a manner that allows perfect information recovery from the output. Unlike irreversible gates, reversible gates ensure that each input aggregation maps to a unique output aggregation, enabling bidirectional computations. Common examples of reversible gates include the Toffoli gate, Fredkin gate, and Peres gate [2].

The reversible circuits are composed of different interconnected reversible logic gates designed to perform specific computations, maintaining reversibility. These circuits are characterized by the absence of information loss and the ability to regress the system's state back to its initial state. Reversible circuits play an important role in achieving efficient reversible computing and have applications in various domains, including cryptography, error correction, and quantum computing.

2.1 Reversible Logic Synthesis Techniques

Reversible logic synthesis is the process of transforming an irreversible logic specification into an equivalent reversible logic circuit. Various techniques have been developed to synthesize reversible logic, including Boolean logic-based approaches and quantum-inspired methods.

Boolean logic-based synthesis techniques mainly focus on manipulating Boolean functions to achieve reversibility. These methods involve the application of transformation rules and algorithms to decrease the number of gates and optimize the efficiency of the designed circuit. Examples of Boolean logic-based synthesis techniques include Shannon decomposition, Bennett decomposition, and Multiple Control Toffoli Network (MCTN) synthesis.

The Quantum-inspired synthesis techniques draw inspiration from quantum computing principles and algorithms. They leverage concepts like quantum gates, quantum circuits, and quantum cost models for reversible logic circuit synthesis. The Quantum-inspired synthesis techniques exploit the quantum superposition and entanglement properties in reversible circuit optimization for specific applications [3].

2.2 Quantum and Classical Reversible Computing Models

Reversible computing models can be categorized into quantum reversible computing and classical reversible computing. Quantum reversible computing employs the principles of quantum mechanics like superposition and entanglement to perform the computations. The Quantum gates like the Hadamard gate, CNOT gate, and T gate form the basis of quantum reversible circuits. Quantum reversible computing has the potential to provide significant computational advantages, particularly in specific cryptographic algorithms and optimization problems.

The classical reversible computing models focus on achieving reversibility within the classical computing architectures, and these models aim to minimize energy dissipation with reduced heat generation during computations. Techniques like adiabatic computing, quantum computing, and reversible CMOS circuits have been developed to realize classical reversible computing models [4].

2.3 Formalism and Mathematical Representations of Reversible Logic

Encryption refers to the science of spoofing a message so that only the intended recipient can decrypt the received message. Encryption is central to data security. It not only ensures the message's confidentiality but also helps provide message integrity, authentication, and digital signatures. The first message or document sent is called plaintext; the obfuscated version is ciphertext. Plaintext and ciphertext are binary strings of the same length. Obfuscating the original plaintext is encryption, whereas decryption is the original plaintext restoration process. Cryptography is categorized broadly as private key and public key cryptography. The encryption or public key should be known to the outside world, and a private decryption or private key should be kept secret. Commonly used private key encryption algorithms comprise Data Encryption Standard, Advanced Encryption Standard, Blowfish, RC4, and public key encryption such as AES and elliptical curve encryption. Plaintext messages can be hidden in two ways. Steganography obscures the message's existence, while encryption makes the message incomprehensible to foreigners through various text transformations.

Formalism and mathematical representation are necessary in analyzing and designing reversible logic circuits. Formal languages like Permutation Reversible (PBR) logic, Reversible Structured Vectors (RSV), and RevLib provide a standardized way to describe and model reversible circuits.

Mathematical representations like matrices and permutation functions are used for reversible logic gates and circuit representation behavior. The matrices enable a concise representation of reversible gates, whereas permutation functions describe the input–output relationships of the reversible circuits. These formalism and mathematical representations are used to analyze, synthesize, and optimize reversible logic circuits.

Hence, understanding the theoretical foundations of reversible logic is essential for effectively utilizing it in cybersecurity applications. The reversible logic gates and circuits, synthesis techniques, quantum, classical reversible computing models, and formalism provide the necessary tools and frameworks for designing and implementing secure and efficient reversible logic-based cyber security solutions.

3 Implementation Methodologies

Implementing reversible logic gates requires suitable technologies to realize reversible operations without introducing any information loss. Several technologies have been explored for reversible logic gate implementations, including CMOS (Complementary Metal Oxide Semiconductor), quantum computing platforms, and emerging technologies like adiabatic and optical computing [5].

3.1 Reversible Logic Gates Using Different Technologies

The CMOS-based reversible logic gates leverage well-established CMOS technology widely used in conventional computing systems. The reversible CMOS circuit utilizes specialized circuit designs and logic transformations to ensure reversibility. These circuits can be optimized for energy efficiency and integrated into existing CMOS-based systems with relatively low overhead.

Quantum computing platforms like qubits-based quantum systems provide a robust framework for implementing reversible logic gates. Quantum gates like CNOT and Toffoli are inherently reversible and can be used for reversible circuit construction. The quantum-inspired reversible logic can exploit the unique properties of quantum systems to achieve efficient reversible computations.

Emerging technologies such as adiabatic logic and optical computing offer potential avenues for reversible logic implementations. The adiabatic logic circuit uses the energy conservation principle to achieve reversibility and has shown promising low-power applications. Optical reversible computing leverages light-based signals and optical components for reversible logic gates and circuit realization, enabling high-speed and energy-efficient computations.

3.2 Design and Optimization Techniques for Reversible Circuits

Designing efficient reversible circuits requires techniques that minimize gate count, reduce circuit complexity, and optimize performance metrics. Several design and optimization techniques have already been proposed in the literature.

The Gate-level synthesis approaches aim to generate reversible circuits directly from a given logic specification. Techniques like Shannon decomposition, Peres decomposition, and Binary Decision Diagram (BDD) [6] based synthesis are commonly employed to decompose the logic functions and construct reversible circuits.

The technology-independent synthesis technique generates reversible circuits without relying on specific technology constraints. These techniques consider various factors to optimize the circuit structure, including gate count, garbage outputs, and ancillary inputs. Examples of technology-independent synthesis algorithms include ESOP-based synthesis, template-based synthesis, and Multiple Control Toffoli Network (MCTN) syntheses [7].

Optimization techniques such as gate count minimization, garbage output reduction, and delay optimization aim to improve the efficiency and performance of reversible circuits. Methods like gate count optimization using template matching, encoding techniques, and reversible logic synthesis with application-specific optimizations have been proposed to achieve these objectives.

3.3 Reversible Logic Synthesis Algorithms and Tools

Several synthesis algorithms and tools have been developed to facilitate the implementation of reversible logic. These tools automate the process of transforming irreversible logic specifications into equivalent reversible logic circuits.

Tools such as RevKit, RevLib, and RevComp provide comprehensive environments for reversible logic synthesis, optimization, and analysis. They offer functionalities like gate-level synthesis, technology mapping, optimization of circuits, and performance evaluation.

The reversible logic synthesis algorithms like Exact Synthesis and Heuristic Synthesis employ different strategies for reversible circuit generation. The Exact Synthesis algorithms guarantee optimal solutions but may suffer from scalability issues for larger circuits. The Heuristic Synthesis algorithms provide approximate solutions with improved scalability and are suitable for larger circuits [8].

3.4 Performance Evaluation of Reversible Logic Implementations

Benchmarking and performance evaluation are essential for assessing the effectiveness of reversible logic implementations. Various metrics evaluate reversible circuits' quality and efficiency, including gate count, circuit depth, garbage outputs, and power consumption.

Benchmark suites like RevLib provide standardized benchmarks for evaluating the performance of reversible logic implementations. These benchmarks include various circuit sizes and complexities, enabling fair comparisons among synthesis algorithms and respective deployments.

Performance evaluation methodologies such as simulation and formal verification are employed to validate the correctness and functionality of the reversible circuits. The simulation-based approaches involve testing circuits using various input vectors and verifying the required output behavior. The formal verification techniques, such as equivalence checking and model checking, mathematically ascertain the correctness of the reversible circuits.

4 Cyber Security Applications

As the digital landscape continues to expand, the importance of robust cybersecurity measures becomes increasingly evident. The implementation methodologies for reversible logic involve selecting appropriate technologies, designing and optimizing circuits, employing synthesis algorithms and tools, and evaluating the performance of

reversible logic implementations. These methodologies are vital in realizing efficient and practical reversible logic-based cyber security solutions.

4.1 Cryptography: Secure Key Generation, Encryption, and Decryption

Cryptography is the fundamental aspect of cyber security, and reversible logic can enhance cryptographic algorithms and protocols. Reversible logic can be leveraged to generate a secure key and perform encryption and decryption operations.

The reversible logic is characterized by operations that can be perfectly reversed, making it particularly valuable in cyber security applications where data integrity, confidentiality, and traceability are paramount. Reversible logic ensures that data remains unaltered during cryptographic operations, making it suitable for tasks that require data integrity, such as digital signatures. Encryption algorithms can employ reversible logic to protect sensitive information from unauthorized access. The bidirectional nature of reversible logic enables traceability, allowing actions to be tracked, monitored, and audited with precision.

Reversible logic is a foundation for the cryptographic techniques that safeguard the data in transit and at rest. Cryptography forms the foundation of cyber security, and the reversible logic brings unique capabilities to the field, enabling secure encryption, decryption, and digital signatures. The reversible logic can be used in symmetric and asymmetric encryption schemes, ensuring data confidentiality through secure key-based transformations. The reversible logic plays a vital role in creating and verifying the digital signatures, which are critical for authentication and non-repudiation.

Reversible encryption ensures that the data transmitted over networks remains confidential and unaltered, safeguarding sensitive information from eavesdroppers. Reversible cryptography protects data stored in databases or physical storage devices, mitigating the risk of unauthorized access or tampering.

Intrusion detection systems (IDS) use reversible logic to monitor and respond to cyber threats. Reversible logic algorithms recognize network traffic patterns and anomalies, helping identify potential intrusions. The bidirectional nature of reversible logic allows for real-time analysis and response to suspicious activity, enabling rapid threat mitigation. Reversible logic-powered IDS can detect and respond to cyber threats in real time, reducing the risk of data breaches. Reversible logic-based algorithms can identify unusual patterns in network traffic, flagging potential attacks or vulnerabilities. Cryptographic authentication protocols like TLS (Transport Layer Security) and SSH (Secure Shell) rely on reversible logic to secure network communication between clients and servers. In compliance with data protection regulations (e.g., GDPR), reversible cryptography helps organizations protect user privacy by encrypting personally identifiable information (PII) and other sensitive data.

The advent of quantum computing challenges classical encryption schemes, and research in post-quantum cryptography using reversible logic is ongoing. There is

a need to develop hardware implementations of reversible logic for faster and more efficient cryptographic operations in resource-constrained environments—the development of standardized reversible logic-based cryptographic algorithms to ensure interoperability and security.

The reversible logic is poised to revolutionize the field of cyber security, offering solutions for data protection, intrusion detection, and threat mitigation. Its unique properties, like bidirectional operations and information preservation, make it a valuable tool in safeguarding digital assets from cyber threats. As the cyber security landscape continues to evolve, reversible logic is expected to play an increasingly significant role in ensuring digital systems and data security and integrity.

4.1.1 Symmetric Encryption

In symmetric encryption, reversible logic is employed to encrypt and decrypt data using the same key. The encryption process is designed to be reversible, ensuring that the original data can be accurately reconstructed when decrypted.

4.1.2 Asymmetric Encryption

In asymmetric encryption (public key cryptography), reversible logic enables the creation and use of key pairs for secure communication. Public keys can be used for encryption, while private keys are required for decryption. Reversible transformations ensure that data remains confidential during transmission.

4.2 Reversible Logic in Intrusion Detection and Prevention

Intrusion detection and prevention systems are essential components of cyber security, and the reversible logic brings unique advantages to these systems, enabling real-time monitoring and response to cyber threats. Intrusion detection and intrusion prevention systems rely on reversible logic for their core principles.

Reversible logic-based algorithms recognize patterns and anomalies in network traffic, data packets, and system behavior. These algorithms are designed to identify the deviations from standard patterns which may indicate potential intrusions. The bidirectional nature of reversible logic is well-suited for real-time analysis. It allows IDS and IPS systems to continuously monitor the incoming and outgoing network traffic, enabling immediate responses to suspicious activities.

The reversible logic-based intrusion detection and prevention systems have numerous applications in enhancing cyber security. The reversible logic-powered IDS can detect and respond to cyber threats in real time. This rapid response capability reduces the risk of data breaches and system compromises. The reversible logic-based algorithms excel at identifying unusual patterns or behaviors in the network

traffic or system logs, which is crucial for flagging potential attacks, zero-day vulnerabilities, or unauthorized activities. The IDS and IPS systems can use reversible logic to create and update the pattern-based signatures for known threats. The signatures serve as reference points for detecting previously identified attack patterns. The bidirectional analysis capabilities of reversible logic enable deep packet inspection and traffic analysis, allowing IDS and IPS systems to examine data at a granular level and detect threats or policy violations. The reversible logic provides traceability, making it possible to log and audit actions taken by the IDS or IPS. This information can be invaluable for forensic investigations following a security incident.

The reversible logic-based intrusion detection and prevention systems enhance cyber security by providing real-time threat monitoring and rapid response capabilities. These systems play a crucial role in safeguarding the networked environments, ensuring potential threats are detected and mitigated before they cause significant damage. As cyber threats evolve, the application of reversible logic in intrusion detection and prevention remains a key component of comprehensive cyber security strategies. The secure key generation is crucial for cryptographic systems. Reversible logic can generate random and secure cryptographic keys by exploiting the property of reversibility. The reversible circuit generates keys with high entropy. It ensures that the keys can be uniquely recovered from the outputs—encryption algorithms like symmetric and asymmetric encryption schemes benefit from reversible logic. The reversible circuits implement efficient and secure encryption algorithms by optimizing the gate count and minimizing the power consumption. The reversible logic enables the design of lightweight and energy-efficient encryption algorithms suitable for resource-constrained devices.

Similarly, the reversible logic is utilized in the decryption process. The reversible circuits provide an efficient and secure decryption algorithm that recovers the original plain text from encrypted data. The reversibility property ensures that the decryption process is error-free and maintains the security of the cryptographic system [9].

4.2.1 Authentication and Access Control Mechanisms

The authentication and access control mechanisms are crucial in ensuring the systems' and data's security and integrity. The reversible logic can be integrated into authentication protocols to enhance efficiency and security.

Reversible logic can enable the design of authentication schemes resistant to replay attacks, where an attacker intercepts and replays previously captured authentication messages. The authentication protocols can ensure that each authentication message is uniquely processed, preventing unauthorized access by reversible circuits.

The access control mechanisms like access control lists and role-based access control can also benefit from reversible logic. Reversible circuits can implement access control policies more efficiently and securely. The reversible logic-based access control mechanisms provide fine-grained access control, minimizing gate count and reducing the overhead associated with access control enforcement [10].

4.2.2 Intrusion Detection and Prevention Systems

The intrusion detection and prevention system is vital in safeguarding systems and networks from unauthorized access and malicious activities. The reversible logic can contribute to designing and implementing efficient and effective IDPS solutions.

The reversible logic can be integrated into the IDPS algorithm to enhance anomaly detection capabilities. The IDPS processes and analyzes the network traffic data with reduced power consumption and improved accuracy by employing reversible circuits. The reversible logic-based IDPS detects complex attack patterns, reduces false positives, and provides real-time intrusion detection.

Moreover, reversible logic can be utilized to develop intrusion prevention mechanisms. The reversible circuits can efficiently implement intrusion prevention algorithms that respond to the detected threats by blocking or mitigating malicious activities. The reversible logic-based IDPS enhances system resilience, minimizes resource consumption, and mitigates the impact of cyber attacks [11].

4.2.3 Secure Data Storage and Retrieval Techniques

Secure data storage and retrieval are the capacious facets of cyber security. The reversible logic is employed to enhance the security and efficiency of the data storage and retrieval mechanisms.

The reversible circuits are utilized for secure data encryption and storage. The sensitive information is encrypted and stored securely by integrating reversible logic into the data storage systems. The reversible logic-based data storage solutions ensure data integrity, confidentiality, and efficient retrieval.

The data retrieval mechanisms also benefit from the reversible logic. The reversible circuits can provide an efficient and secure method for retrieving stored data. The reversible logic-based retrieval techniques can minimize the gate count, reduce latency, and enable safe access to stored information.

5 Integration with Existing Security Mechanisms

In the ever-evolving cybersecurity landscape, integrating reversible logic with conventional cryptographic algorithms represents a promising approach to enhance data protection, security, and efficiency. The reversible logic can be integrated with traditional cryptographic algorithms to improve security, efficiency, and resilience. The cryptographic algorithms benefit from the reversibility property and exploit its advantages by incorporating reversible logic components.

5.1 Techniques for Integrating Reversible Logic with Conventional Cryptography

Reversible logic can be used to optimize the key generation process and encryption and decryption operations for symmetric encryption algorithms. The reversible circuits can improve the efficiency of key scheduling and expansion procedures, resulting in more secure and faster encryption/decryption processes [12]. The reversible logic can enhance key generation, distribution, and exchange protocols in asymmetric encryption algorithms. The reversible circuits can enable efficient and secure generation of public–private key pairs and facilitate secure key exchange between the communicating parties.

The hash functions essential for data integrity verification can also be improved with reversible logic. The reversible circuits can enhance the performance of hash function computations, reduce power consumption, and provide secure hashing operations. The reversible key management techniques focus on securely generating, storing, and exchanging cryptographic keys. These methods ensure that cryptographic keys are handled efficiently and securely, reducing the risk of key compromise. The Quantum Key Distribution protocols use reversible logic for secure key exchange, ensuring that keys remain confidential and tamper-evident.

5.2 Reversible Encryption Algorithms

Reversible data hiding techniques leverage the reversible logic to embed data within media files like images, audio, or video in a way that allows the original content to be reconstructed perfectly after extracting the hidden data. This is very useful for watermarking, steganography, and covert communication. In the least Significant Bit Substitution, the reversible logic is applied to replace the least significant bits of data with some hidden information. The subtle changes do not affect the overall quality or functionality of the media [13]. LSB substitution is commonly used in image and audio watermarking where additional information, such as copyright data, is embedded without degrading the visual or auditory quality. These techniques showcase the versatility and utility of reversible logic in enhancing various aspects of conventional cryptography. They ensure that the encryption, decryption, and data-hiding operations are secure and perfectly reversible, enabling the original data to be retrieved without any loss of information. As the field of reversible logic continues to evolve, these integration methods are poised to play a pivotal role in improving the security and efficiency of cryptographic systems.

5.3 *Complementary Roles of Reversible Logic and Existing Security Measures*

The reversible logic can complement the existing security measures by providing additional layers of protection and addressing specific security challenges. Integrating reversible logic with existing security measures is a powerful approach to bolster data protection, privacy, and resilience. Reversible logic is pivotal in preserving data integrity by ensuring that cryptographic operations do not result in data loss or corruption. Integrating with existing security measures guarantees that the data remains intact, enhancing the overall security of data storage, transmission, and processing.

5.3.1 Enhanced Encryption

Encryption is a cornerstone of cyber security, protecting the data from unauthorized access and ensuring confidentiality. Bidirectional and efficient properties of reversible logic enhance the existing encryption methods. Reversible logic optimizes the encryption processes, allowing for secure and efficient bidirectional transformations. This results in more substantial encryption schemes that resist attacks while minimizing computational overheads. Authentication mechanisms verify the identity of users and devices, preventing unauthorized access and ensuring trust in digital interactions. The reversible logic strengthens authentication by providing a secure foundation for cryptographic key exchange and user identity verification. By integrating reversible logic, the authentication protocols become more resilient to various attacks, including impersonation and man-in-the-middle attacks [14].

The complementary role of reversible logic and existing security measures enhances secure data storage and transmission, safeguarding sensitive information from unauthorized access and tampering. The reversible logic bolsters the resilience of cryptographic systems by ensuring the preservation of data integrity, optimizing encryption processes, and strengthening the authentication mechanisms.

5.3.2 Data Integrity in Cyber Security

Data integrity is a foundational concept in cyber security. It refers to the assurance that the data remains unchanged and uncorrupted throughout its life cycle. Data integrity is essential to maintain trust in digital interactions, preventing unauthorized alterations, and safeguarding sensitive information. The data should resist unauthorized modifications, ensuring its original state can be preserved. Mechanisms should be implemented to identify and rectify accidental or deliberate data corruption. The data integrity should be maintained from creation through storage, transmission, and processing [15].

The reversible logic ensures that the data remains unaltered during cryptographic operations. This is particularly valuable when data is encrypted or decrypted as it guarantees that original data can be reconstructed without any loss or corruption. By preserving the data integrity, the reversible logic helps detect unauthorized modifications or tampering attempts. The system can raise alarms or initiate corrective actions if data integrity is compromised. The reversible logic maintains the data integrity across the entire data life cycle, ensuring that data remains unchanged during storage, transmission, and processing. Integrating reversible logic with existing security measures strengthens data integrity protection, assuring stakeholders that their digital assets are secure and unaltered. This synergy enhances the overall cyber security framework, bolstering trust and reliability in digital interactions.

The intrusion detection and prevention systems can benefit from reversible logic by integrating it with an existing anomaly detection algorithm [16]. The reversible circuits can enhance anomaly detection accuracy by analyzing network traffic patterns in a reversible and energy-efficient manner; by combining reversible logic-based IDPS with the traditional signature-based detection methods, comprehensive and robust intrusion detection capabilities can be achieved.

Reversible logic can also be integrated into an access control mechanism to augment effectiveness. Fine-grained access control can be achieved, ensuring that only authorized entities can access sensitive resources by incorporating reversible circuits into access control systems. Reversible logic-based access control mechanisms can complement existing authentication and authorization systems, providing an additional layer of security.

Furthermore, the reversible logic can enhance the resilience of cryptographic protocols against side-channel attacks. The side channel attacks exploit the information leakage via the physical implementation of cryptographic algorithms. Hence, the vulnerability of cryptographic systems to such attacks can be reduced by designing reversible circuits that minimize side-channel leakage.

5.4 Addressing Security Vulnerabilities and Enhancing Resilience Through Reversible Logic

The reversible logic can address specific security vulnerabilities and enhance the resilience of systems against potential attacks. In an era marked by increasing cyber security threats and vulnerabilities, this section explores how reversible logic can address security vulnerabilities and enhance resilience across various information technology and cyber security domains. A constantly evolving threat landscape characterizes the digital landscape. Cyber adversaries employ sophisticated techniques to exploit vulnerabilities and compromise systems, making it essential to address security weaknesses proactively [17].

5.4.1 Common Security Vulnerabilities

Security vulnerabilities include software, network, social engineering, and cryptographic weaknesses. Software Vulnerabilities include weaknesses in software code that attackers, such as buffer overflows or injection attacks, can exploit. The network vulnerabilities show flaws in network configurations that expose systems to unauthorized access or data interception. In social engineering, human psychology is manipulated to deceive individuals into disclosing sensitive information or performing actions that compromise security. The cryptographic weaknesses explore vulnerabilities in encryption algorithms that could be exploited to decrypt encrypted data.

5.4.2 Role of Reversible Logic in Addressing Vulnerabilities

The ability of reversible logic to preserve data integrity addresses many common security vulnerabilities. Like mitigating software vulnerabilities by ensuring that the data remains unaltered during cryptographic operation, the reversible logic can protect against data corruption resulting from software vulnerabilities. The reversible logic also enhances encryption processes, safeguarding data in transit and mitigating network vulnerabilities. Reversible logic contributes to addressing vulnerabilities by optimizing encryption [18]. Strong encryption protects data even if a software vulnerability is exploited, rendering stolen data unreadable. Enhanced encryption reduces the risk of unauthorized access to data transmitted over networks addressing network vulnerabilities.

Cyber resilience involves proactive measures to minimize vulnerabilities and strategies for effective response and recovery. Reversible logic contributes to cyber resilience in several ways, such as reducing attack surface, efficient recovery, and adaptive security. By mitigating vulnerabilities, the reversible logic minimizes the attack surface available to cyber adversaries, making it more difficult for them to exploit weaknesses. The optimized encryption and data integrity preservation of reversible logic simplifies data recovery processes after security incidents, minimizing downtime. The bidirectional nature of reversible logic allows for adaptive security measures, enabling organizations to respond dynamically to emerging threats.

Reversible logic protects against data corruption caused by software vulnerabilities, ensuring data integrity. The reversible logic can be employed to ensure the integrity and authenticity of software updates, preventing malicious updates from compromising systems. The reversible logic strengthens network encryption, protecting data in transit and mitigating network vulnerabilities. Adaptive security measures enabled by reversible logic help networks respond effectively to emerging threats.

The research into post-quantum cryptography using reversible logic to address vulnerabilities is exposed by quantum computing. Developing industry standards for integrating reversible logic into existing security measures ensures interoperability

and security. Raising awareness about reversible logic's benefits and applications addresses vulnerabilities and enhances resilience.

Reversible logic offers innovative solutions to address security vulnerabilities and enhance cyber resilience. By preserving data integrity, optimizing encryption, and enabling adaptive security measures, the reversible logic contributes to developing robust cybersecurity strategies. As cyber threats continue to evolve, the integration of reversible logic is poised to play a pivotal role in bolstering the security and resilience of digital systems and organizations.

5.4.3 Addressing Security Vulnerabilities and Enhancing Resilience Through Reversible Logic

Understanding security vulnerabilities is crucial to safeguarding information and systems from cyber threats in an increasingly connected digital world. This section explores organizations' evolving threat landscape and common security vulnerabilities. The digital landscape is constantly in flux, with cyber adversaries developing new techniques and tools to exploit vulnerabilities. Understanding the dynamics of this landscape is essential for adequate cyber security. Cybercriminals and state-sponsored actors employ increasingly sophisticated methods to breach security defenses. Organizations of all sizes and industries are the potential targets, and attackers often tailor their strategies to specific targets. Adopting emerging technologies such as IoT and cloud computing introduces new attack vectors and vulnerabilities.

6 Advantages and Challenges

Various advantages and challenges related to cyber security enhancement through reversible logic are discussed below.

6.1 Energy Efficiency and Low Power Consumption Benefits

One of the momentous advantages of reversible logic is its energy efficiency and low power consumption characteristics. The reversible circuits inherently minimize information loss, reducing heat dissipation and energy consumption. This energy efficiency is particularly beneficial for resource-constrained devices and systems with limited power budgets, such as IoT devices and mobile platforms. Hence, energy consumption can be optimized by integrating reversible logic into cyber security solutions, leading to longer battery life, reduced operational costs, and a smaller environmental footprint.

6.2 Scalability and Fault-Tolerant Features

The reversible logic offers scalability and fault-tolerant capabilities that can benefit cyber security applications. The reversible circuits can be designed and optimized to handle large-scale computations efficiently. As the reversible logic circuits exhibit no information loss, they can be easily expanded or replicated without introducing errors, making them suitable for scaling up cryptographic systems and security mechanisms. Furthermore, reversible logic-based systems can incorporate fault detection and correction techniques, ensuring robustness against transient faults and hardware failures.

6.3 Security Concerns and Potential Attack Vectors

While the reversible logic offers advantages, it also introduces new security concerns and potential attack vectors. The reversible circuits can be vulnerable to various attacks like side-channel attacks, timing attacks, and fault attacks. The side channel attacks exploit information leakage from the physical implementation of reversible circuits, while timing attacks can leverage the unique properties of reversible logic gates to extract sensitive information. The fault attacks can target reversible circuits to induce errors, compromising security. It is also essential to carefully analyze the security implications of reversible logic-based solutions and employ countermeasures to mitigate these vulnerabilities.

6.4 Overheads and Trade-Offs

Implementing reversible logic-based cyber security solutions may involve certain overheads and tradeoffs. The reversible circuits typically require additional gates and ancillary inputs to ensure reversibility, which can increase the gate count and circuit complexity. This may result in increased hardware costs, latency, and resource utilization. Additionally, the reversible logic synthesis techniques and optimization algorithms may introduce computational overheads and design complexities. Tradeoffs between security, efficiency, and resource utilization must be carefully considered while designing and implementing reversible logic-based cybersecurity solutions.

6.5 Availability of Design Tools and Expertise

The availability of design tools and expertise in reversible logic synthesis and implementation is still limited compared to the traditional security methodologies. Designing efficient and secure reversible logic-based cybersecurity solutions requires specialized knowledge and continuously evolving tools. The development of user-friendly design tools and the availability of skilled professionals in reversible logic design are crucial to accelerate the adoption and practical implementation of reversible logic in cyber security.

Hence, the reversible logic offers energy efficiency, scalability, fault-tolerant features, and the potential to enhance cyber security. However, potential attack vectors and overheads must be carefully addressed in light of security concerns. By considering these advantages and challenges, researchers and practitioners can harness the potential of reversible logic to develop secure, efficient, and resilient cybersecurity solutions.

7 Case Studies and Experimental Results

Several case studies have showcased the practical applications of reversible logic in real-world cybersecurity scenarios. These case studies highlight the benefits and effectiveness of reversible logic in addressing specific security challenges.

7.1 Case Studies Demonstrating the Application of Reversible Logic in Real-World Cyber Security Scenarios

One case study focuses on the integration of reversible logic into cryptographic algorithms. The study demonstrates the implementation of a reversible logic-based encryption algorithm that offers improved security and energy efficiency compared to traditional algorithms. The case study includes performance evaluations like encryption/decryption speed, power consumption, and security analysis.

Another case study explores the use of reversible logic in intrusion detection systems. The study presents a reversible logic-based anomaly detection algorithm that detects network intrusions with higher accuracy and reduced power consumption. Experimental results demonstrate the effectiveness of the reversible logic-based approach in detecting known and unknown intrusion patterns.

Furthermore, a case study examines the integration of reversible logic into access control mechanisms. The study presents a reversible logic-based access control model that enhances fine-grained access control and reduces gate count in access control systems. The case study includes performance evaluations like access latency, scalability, and security analysis.

7.2 *Experimental Evaluations and Comparative Analyses of Reversible Logic-Based Security Solutions*

The experimental evaluations and comparative analysis are conducted to assess the performance and effectiveness of reversible logic-based security solutions compared to traditional security approaches.

For instance, an experimental evaluation compares reversible logic-based encryption algorithms' performance with conventional encryption algorithms. The review includes metrics like encryption or decryption speed, power consumption, and security analysis. The results demonstrate the advantages of reversible logic-based encryption regarding security and energy efficiency.

Another comparative analysis focuses on the effectiveness of reversible logic-based intrusion detection systems compared to the traditional approaches. The study includes metrics like detection accuracy, false positive rate, and resource utilization. The result showcases the superiority of reversible logic-based intrusion detection in terms of accuracy and energy efficiency.

Additionally, a comparative analysis examines the performance of reversible logic-based access control mechanisms compared to traditional access control systems. The study includes metrics like access latency, gate count, and security analysis. The results demonstrate the benefits of reversible logic-based access control in terms of efficiency and fine-grained access control.

These case studies, experimental evaluations, and comparative analyses contribute to the body of knowledge by providing empirical evidence of the practical application, performance, and advantages of reversible logic in cyber security. They offer insights into reversible logic-based security solutions' feasibility, effectiveness, and potential in real-world scenarios.

8 Conclusion and Future Scope

The reversible logic provides unique advantages, including energy efficiency, scalability, fault-tolerant features, and the potential to address security vulnerabilities. The cryptographic algorithms can be more secure and efficient by leveraging reversible logic. The access control mechanisms can achieve fine-grained control, and intrusion detection and prevention systems can enhance their accuracy and energy efficiency. The Reversible logic also offers benefits in secure data storage and retrieval.

However, the integration of reversible logic in cyber security also presents challenges. Security concerns, potential attack vectors, overheads, and the availability of design tools and expertise must be carefully addressed. The vigilance is required to mitigate vulnerabilities, analyze security implications, and develop countermeasures against potential threats.

To advance the reversible logic-based cyber security field, future research should explore emerging trends like quantum reversible logic, machine learning integration,

hardware implementation, and reversible logic in blockchain technology. Additionally, addressing open challenges related to security analysis, scalability, design tools, standardization, and practical deployments will contribute to the practical adoption of reversible logic in cybersecurity applications.

By combining theoretical advancements, experimental evaluations, and case studies, researchers and practitioners can harness the potential of reversible logic to develop secure, efficient, and resilient cybersecurity solutions. The collaboration among academia, industry, and government organizations is essential for interdisciplinary research and practical implementation of reversible logic-based cyber security measures.

In conclusion, reversible logic offers a valuable avenue to enhance cyber security, and its integration holds excellent potential for developing robust and efficient security solutions in an ever-evolving digital landscape.

The field of reversible logic and its application in cyber security continues to evolve, presenting several emerging trends and directions for future research.

With the furtherance of quantum computing, exploring the integration of reversible logic with quantum computing platforms and quantum algorithms holds promise for enhancing cyber security. Research on quantum reversible logic can contribute to developing secure quantum communication protocols, quantum-resistant encryption algorithms, and quantum error correction techniques.

Integrating machine learning techniques with reversible logic can open new avenues for improving cyber security. Research can focus on developing reversible logic-based machine-learning algorithms for anomaly, malware, and intrusion detection. Exploring the application of reversible logic in privacy-preserving machine learning can also be a fruitful area of investigation.

Further advancements in hardware technologies, such as nanoscale devices, quantum dot cellular automata, and emerging nonvolatile memory technologies, can enable more efficient and compact hardware implementations of reversible logic. Research focuses on developing novel reversible logic-based hardware architectures for secure and energy-efficient computing systems.

The integration of reversible logic in blockchain technology enhances the security and scalability of blockchain networks. Research explores reversible logic for efficient consensus protocols, secure transaction processing, and data privacy preservation in blockchain systems.

Open challenges and research gaps still need to be addressed, while the reversible logic shows promise in enhancing cyber security.

Further research is needed to analyze reversible logic-based security solutions' implications and vulnerabilities. Hence, developing rigorous security analysis frameworks, addressing potential attack vectors, and investigating countermeasures against security threats are essential focus areas.

As reversible logic scales up to larger circuits and complex systems, scalability becomes a significant challenge. Research explores techniques for optimizing reversible logic synthesis, reducing gate count, and improving the efficiency of large-scale reversible circuits.

Developing user-friendly design tools and methodologies for reversible logic-based cybersecurity solutions is crucial for broader adoption. Research focuses on creating efficient synthesis algorithms, optimization techniques, and simulation tools that facilitate the design and implementation of reversible logic-based security systems.

Establishing standardized benchmarks and performance metrics for evaluating reversible logic-based cybersecurity solutions is essential for fair comparisons and benchmarking. The research also contributes to developing standardized benchmarks, performance evaluation methodologies, and benchmark suites specific to reversible logic-based security applications.

References

1. Karunamurthi, S., Natarajan, V.K.: VLSI implementation of reversible logic gates cryptography with LFSR key. *Microprocess. Microsyst.* **69**. ISSN 01419331. <https://www.sciencedirect.com/science/article/pii/S014193311930122X> (2019)
2. Samrin, S.S., Patil, R., Itagi, S., Chetti, S.C., Tasneem, A.: Design of logic gates using reversible gates with reduced quantum cost. *Glob. Transit. Proc.* **3**(1). ISSN 2666285X. <https://www.sciencedirect.com/science/article/pii/S2666285X22000474> (2022)
3. Sasamal, T.N., Singh, A.K., Mohan, A.: Reversible logic circuit synthesis and optimization using adaptive genetic algorithm. *Procedia Comput. Sci.* **70**. ISSN 18770509. <https://www.sciencedirect.com/science/article/pii/S1877050915032184> (2015). <https://doi.org/10.1016/j.procs.2015.10.054>
4. Frank, M.P., Shukla, K.: Quantum foundations of classical reversible computing. *Phys. Inf. Phys. Found. Comput. Entropy* **23**, 701 (2021). <https://doi.org/10.3390/e23060701>
5. Saravanan, M., Manic, K.S., Umasuresh, Aravind, C.V., Wiselin, J.: Design of reversible logic gates for digital applications. *Asian J. Electr. Sci.* **2**(1) (2013). ISSN 2249-6297
6. Bryant, R.E.: Binary decision diagrams. In: Clarke, E., Henzinger, T., Veith, H., Bloem, R. (eds.) *Handbook of Model Checking*. Springer, Cham (2018). https://doi.org/10.1007/9783319105758_7
7. Große, D., Wille, R., Dueck, G.W., Drechsler, R.: Exact multiplecontrol Toffoli network synthesis with SAT techniques. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **28**(5) (2009). <https://doi.org/10.1109/TCAD.2009.2017215>
8. Sasamal, T., N., Singh, A.K., Mohan, A.: reversible logic circuit synthesis and optimization using adaptive genetic algorithm. *Procedia Comput. Sci.* **70** (2015). ISSN 18770509. <https://www.sciencedirect.com/science/article/pii/S1877050915032184>, <https://doi.org/10.1016/j.procs.2015.10.054>
9. Chandran, G., Mary, H.: VLSI implementation of image encryption and decryption using reversible logic gates. In: *International Conference on Power Electronics and Renewable Energy Applications (PEREA)*, Kannur, India (2020). <https://doi.org/10.1109/PEREA51218.2020.9339781>
10. Das, J.C., Debashis, D.: User authentication based on quantum dot cellular automata using reversible logic for secure nanocommunication. *Arab. J. Sci. Eng.* **41** (2016). <https://doi.org/10.1007/s133690151870z>
11. Kalinin, M., Krundyshev, V: Security intrusion detection using quantum machine learning techniques. *J. Comput. Virol. Hack. Tech.* **19**, 125–136 (2023). <https://doi.org/10.1007/s11416022004350>
12. Karunamurthi, S., Natarajan, V.K.: VLSI implementation of reversible logic gates cryptography with LFSR key. *Microprocess. Microsyst.* **69** (2019). ISSN 01419331, <https://doi.org/10.1016/j.micpro.2019.05.015>

13. Carrillo, P., Kalva, H., Magliveras, S.: Compression independent reversible encryption for privacy in video surveillance. *EURASIP J. Inf. Secur.* (2010). <https://doi.org/10.1155/2009/429581>
14. Manikandan, V.M., Bini, A.A.: An improved reversible data hiding through encryption scheme with block prechecking. *Procedia Comput. Sci.* **171** (2020). ISSN 1877-0509
15. Hasan, M.Z., Hussain, M.Z., Mubarak, Z., Siddiqui, A.A., Qureshi, A.M., Ismail, I.: Data security and integrity in cloud computing. In: *International Conference for Advancement in Technology (ICONAT)*, Goa, India (2023). <https://doi.org/10.1109/ICONAT57137.2023.10080440>
16. Azeez, N.A., Bada, T.M., Misra, S., Adewumi, A., Van der Vyver, C., Ahuja, R.: Intrusion detection and prevention systems: an updated review. In: Sharma, N., Chakrabarti, A., Balas, V. (eds.) *Data Management, Analytics and Innovation. Advances in Intelligent Systems and Computing*, vol. 1042. Springer, Singapore (2020). https://doi.org/10.1007/978-981-32-9949-8_48
17. Saki, A.A., Alam, M.A., Phalak, K., Suresh, A., Topaloglu, R.O., Ghosh, S.: A survey and tutorial on security and resilience of quantum computing. In: *IEEE European Test Symposium ETS* (2021)
18. Valinataj, M., Mirshekar, M., Jazayeri, H.: Novel low-cost and fault-tolerant reversible logic adders. *Comput. & Electr. Eng.* **53** (2013). ISSN 0045-7906. <https://doi.org/10.1016/j.compelteceng.2016.06.008>

Beyond Passwords: Enhancing Security with Continuous Behavioral Biometrics and Passive Authentication



Pankaj Chandre, Suvarna Joshi, Rahul Rathod, Jyoti Nandimath, Bhagyashree Shendkar, and Yuvraj Nikam

Abstract In the digital age, traditional password-based authentication systems are increasingly vulnerable to sophisticated cyberattacks, such as phishing, credential theft, and social engineering. Security paradigms are shifting towards continuous, non-intrusive methods that adapt to user behavior and context to address these limitations. This paper explores an advanced authentication framework that integrates continuous behavioral biometrics with passive authentication techniques to strengthen security beyond traditional methods such as passwords. The system captures and analyzes user behavior (e.g., typing patterns, gestures) alongside passive signals (e.g., location, device, time) to continuously validate user identity. The authentication system dynamically adapts to detect anomalies and mitigate fraud in real time by utilizing data from both behavioral and contextual sources. This adaptive, multi-layered approach enhances security, particularly in sensitive financial applications, by providing continuous monitoring and real-time fraud detection.

P. Chandre (✉) · S. Joshi · B. Shendkar · Y. Nikam
Department of Computer Science and Engineering, MIT School of Computing, MIT Art Design and Technology University, Pune, India
e-mail: pankajchandre30@gmail.com

S. Joshi
e-mail: suvarna.joshi@mituniversity.edu.in

B. Shendkar
e-mail: bhagyashree.shendkar@mituniversity.edu.in

Y. Nikam
e-mail: yuvraj.nikam@mituniversity.edu.in

R. Rathod · J. Nandimath
Department of Information Technology, MIT School of Computing, MIT Art Design and Technology University, Pune, India
e-mail: rahul.rathod@mituniversity.edu.in

J. Nandimath
e-mail: jyoti.nandimath@mituniversity.edu.in

Keywords Behavioral biometrics · Passive authentication · Anomaly detection · Continuous monitoring · Fraud detection · Real-time security

1 Introduction

Traditional password-based authentication has long been the cornerstone of digital security. However, it is increasingly proving inadequate in the face of modern cybersecurity threats. Passwords are often easy to guess, reused across multiple accounts, or stored insecurely [1]. Users frequently choose weak passwords that are vulnerable to brute-force attacks or phishing schemes, and even strong passwords can be compromised through sophisticated methods like keylogging, database breaches, or social engineering. As attackers evolve, passwords have become a single point of failure, placing sensitive data and systems at risk. Data breaches have increased in scope and frequency in recent years, with leaked credentials frequently acting as an attacker's point of access. According to industry studies, many breaches are caused by weak or stolen passwords [2]. These hacks affect consumers' and service providers' trust in addition to causing financial harm. Furthermore, because platforms and devices are becoming increasingly interconnected, a single password breach can result in extensive illegal access and expose private and company data. A potential remedy for the shortcomings of password-based systems is behavioral biometrics. Behavioral biometrics analyzes patterns in human behavior, including typing speed, mouse movements, touchscreen gestures, and gait, for authentication, in contrast to traditional authentication approaches that rely on static credentials. This dynamic and ongoing approach provides users real-time behavior tracking during a session [3]. Behavioral biometrics-powered continuous authentication provides a more reliable security layer by passively confirming the user's identity without interfering with their experience. This method is thought to be the secure, frictionless authentication of the future.

This survey aims to explore the latest developments in behavioral biometrics, emphasizing passive and continuous authentication methods. Identity verification techniques have changed significantly over the past ten years, moving away from static, one-time logins and toward ongoing user behavior tracking. This survey aims to give a thorough overview of the significant techniques and technologies being created and used in various sectors, including online retail, healthcare, and finance [4, 5]. This paper will shed light on recent advancements and show how the security landscape is changing due to these new methods. Although there are many advantages to continuous and passive behavioral authentication, such as increased security and user convenience, its implementation has certain obstacles. This survey will objectively assess these strategies' efficacy in practical applications, gauging how well they reduce security breaches and enhance user experience. It will also cover the possible drawbacks, including scaling problems, false positives, and privacy

concerns. This study will evaluate the advantages and disadvantages of both continuous and passive authentication through a thorough analysis of previous research, offering a fair assessment of its place in future authentication systems.

2 Background and Literature Review

A. Evolution of Authentication Mechanisms

Authentication mechanisms have evolved significantly over time, adapting to the growing complexity of cybersecurity threats. Initially, passwords were the primary method of securing access to systems. However, as attackers developed methods to steal or crack passwords, reliance on this single-factor authentication became insufficient. The implementation of two-factor authentication (2FA), which mandates that users authenticate their identity using two separate factors—something they own (like a token or phone) and something they know (like a password—was the following significant change[6]. Although this method increased security, it caused additional friction for users, discouraging some of them from using it. Another advancement was the introduction of biometric authentication, which includes voice, facial, and fingerprint recognition. Passwords could be replaced by a more user-friendly and safe option: biometrics. On the other hand, physical biometrics are static and subject to theft or spoofing. This opened the door for the most recent development in authentication techniques, behavioral biometrics, which examines ongoing patterns in user activity.

B. Behavioral Biometrics: Definition and Scope

Behavioral biometrics refers to the identification and authentication of individuals based on their unique patterns of behaviour [7]. Unlike physical biometrics, which rely on physical traits (fingerprints, iris scans), behavioral biometrics capture how users interact with devices. Common examples include:

- Typing dynamics: The rhythm, speed, and pressure applied while typing.
- Mouse movements: How users move, click, and interact with a mouse or touchpad.
- Touchscreen gestures: The patterns of swipes, taps, and pinches on a smartphone or tablet.
- Gait analysis: How a person walks, detected through mobile device sensors.
- Device handling: How a user holds and tilts their device while using it.

Behavioral biometrics are dynamic, continuously gathering data on how users interact with devices, in contrast to physical biometrics rely on static, physical traits that remain constant over time (such as fingerprints and iris patterns). Because an attacker would have to duplicate one characteristic and a whole behavioral profile, they are less vulnerable to theft or spoofing. Moreover, behavioral biometrics provide greater resilience over extended usage by adjusting to minute modifications in a

user's behavior. They can also be gathered passively during routine user engagement, enhancing security without interfering with user activity.

C. Continuous Authentication versus One-Time Authentication

Traditional authentication methods, such as password entry or biometric scanning, involve one-time authentication—users verify their identity once when logging in. Then, they are granted continuous access until they log out, or the session ends. This method does, however, come with security dangers because an intruder may manage to stay hidden the entire time they are in. Continuous authentication, on the other hand, keeps an eye on users during their session by examining behavioral patterns, including typing speed, mouse movement, or device handling, to make sure the person utilizing the system is still the authorized user [8, 9]. By constantly assessing identity, this technique lowers the possibility of unauthorized access or session hijacking if an attacker takes over mid-session. With continuous authentication, the system never fully trusts any session without continual verification—a change from a static to a dynamic paradigm.

D. Passive Authentication

Passive authentication refers to the process of verifying a user's identity without requiring explicit action on their part. Instead of entering a password or scanning a fingerprint, users are authenticated based on patterns collected in the background, such as how they type, scroll, or interact with a device [10–12]. This passive detection occurs continuously throughout the session, ensuring the legitimate user is consistently validated without additional prompts. The enhanced user experience is the main advantage of passive authentication. Passive systems facilitate seamless engagement and uninterrupted productivity by eliminating the need for frequent authentication requests. Simultaneously, security is preserved since the system keeps an eye out for any irregularities in behavior that could point to fraud or illegal access. Thus, passive authentication is an appealing alternative to contemporary authentication systems since it combines strong security with user comfort.

Buttons, and navigating the screen. By examining mouse trajectories, click velocity, scroll behavior, and pointer movement velocity; systems can create a behavioral profile that allows user differentiation [13, 14]. Because these patterns rely on habits, motor abilities, and even hardware, they are challenging to duplicate. Particularly useful in differentiating between authentic users and imposters—even if the latter have stolen login credentials—are mouse and touchpad biometrics. The technology is widely used in continuous authentication, especially in environments like online banking, where unauthorized access needs to be detected in real-time.

E. Touchscreen and Swiping Behavior

Touchscreen interactions on mobile devices offer yet another rich source of biometric data about behavior. Users display distinct patterns on touchscreens in pinch, zoom, swipe, and tap [15]. Finger pressure, swipe speed, and the angle at which the device is held are some variables that affect these activities. Because touchscreen behavior enables passive user monitoring without frequent credential input, it is beneficial

for continuous authentication [16]. By examining these motions, mobile devices can confirm the user's identification throughout the session. One way to be sure an unauthorized user has not taken over is to track a person's usual swipe gesture for unlocking their phone or interacting with apps. This type of authentication is prevalent in mobile banking and healthcare apps, where sensitive data requires robust security but minimal friction for the user.

F. Gait Analysis

Gait analysis is a behavioral biometric that recognizes persons based on their gait patterns using sensors found on mobile devices, such as gyroscopes and accelerometers. Everybody has a different gait, which is impacted by walking rhythm, stride length, and body composition. Systems can continuously verify people while they walk or move with their smartphones by collecting data from various sensors [17]. This approach is constructive with mobile devices, where the user frequently holds the phone in motion. When an unauthorized user manipulates a device, such as after a phone has been stolen, gait analysis can identify this. Research shows that gait-based authentication can achieve high accuracy rates, especially when combined with other behavioral data like typing or touchscreen interaction. It offers an additional security layer for mobile apps, smartwatches, and wearable devices.

G. Device Handling (Grip, Tilt, Motion)

Device handling biometrics examine how users hold, tilt, and move their phones or tablets when interacting with them physically. Every person has a unique style of holding and manipulating their device, impacted by hand size, grip strength, and even posture. These variables can be watched over time to verify a user's identification [18]. For example, when unlocking a smartphone or interacting with apps, sensors can detect the angle at which the device is held, the force used to press buttons, and how the phone is tilted during use. If an attacker gains access to the device, changes in handling patterns can trigger security alerts or re-authentication requirements [19]. Device handling biometrics are especially useful for mobile authentication, where passive and unobtrusive methods are preferred to enhance security without disrupting the user experience. This technique is increasingly being integrated into smartphones, wearable devices, and smart home systems for continuous protection.

3 Passive Behavioral Authentication

Figure 1 illustrates a continuous behavioral biometrics and passive authentication system that significantly enhances security beyond traditional password mechanisms. The User Device serves as the entry point, collecting two types of data: behavioral data (e.g., typing patterns, gestures) and passive data (e.g., location, time, and device information). This information is sent to two primary engines. The Behavioral Biometrics Engine processes the behavioral data, analyzes user patterns, and performs continuous monitoring to detect real-time anomalies. It compares

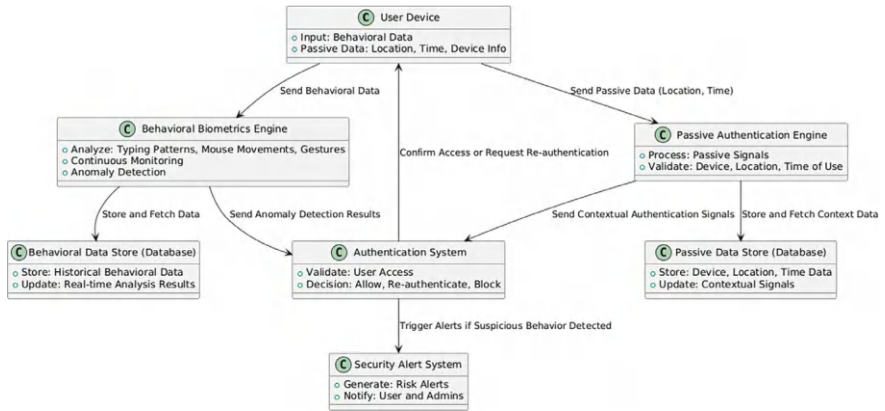


Fig. 1 A continuous behavioral biometrics and passive authentication system that significantly enhances security beyond traditional password mechanisms

current behaviors with historical data stored in the Behavioral Data Store, which is continuously updated with real-time results.

Simultaneously, the Passive Authentication Engine processes contextual signals like the user’s device, location, and use time, validating these signals against stored passive data in the Passive Data Store. The behavioral and passive authentication results are sent to the Authentication System, where they are aggregated and analyzed to decide—whether to allow access, prompt for re-authentication, or block the user based on any detected anomalies or suspicious behavior.

The Security Alert System also plays a critical role by generating risk alerts if any irregularities are identified during the authentication process. It promptly notifies the user and system administrators, ensuring swift responses to potential threats. This system provides adaptive security by continuously monitoring user behavior and passive signals, allowing for real-time detection of fraud or unauthorized access, thus offering a more dynamic and robust alternative to static password-based systems.

A. How Passive Detection Works

Without requiring direct input or actions from the user, passive behavioral authentication works by continuously observing a user’s interactions and behavior patterns in the background. Passive detection works seamlessly while users engage with their devices, unlike typical authentication systems requiring users to input passwords, passcodes, or biometric data actively [20, 21]. Through monitoring multiple behavioral biometrics, including mouse movements, touchscreen gestures, typing patterns, and device handling, passive authentication systems generate a behavioral profile specific to every user. To verify that the user is authentic, these systems continuously match the user’s activity to the pre-established profile. Because of this discreet, real-time monitoring, passive detection is quite helpful for continuous authentication without interfering with the user experience.

B. Applications in Modern Systems

Passive behavioral authentication is growing in sectors where user comfort and security are vital considerations. Passive detection is a technique used in banking to track user behavior on online banking platforms. If the system notices anomalous behavior, including distinct typing patterns or strange mouse movements, it may flag the session for additional verification or re-authentication [22]. This lessens the likelihood of fraud without interfering with authorized users' banking experiences. Passive authentication aids in the security of e-commerce transactions by continuously confirming the customer's identity while they explore, add products to their cart and finish the transaction. Reducing the need for additional authentication enhances the user experience and lowers the likelihood of fraudulent purchases. In corporate security, passive behavioral biometrics protect sensitive data within organizations. Employee behavior is monitored as they access systems, handle confidential documents, or perform critical operations. Any anomalies—such as unusual activity during a session—can trigger a security response, ensuring that only authorized personnel can access sensitive information.

C. Impact on User Experience

One of the key advantages of passive behavioral authentication is its ability to reduce friction in the user experience. Traditional authentication methods, such as passwords, PINs, or biometric scans, require users to perform active steps to verify their identity. This can disrupt workflow and lead to frustration, especially when users must re-authenticate multiple times during a session. Passive authentication, on the other hand, operates in the background silently and constantly verifies the user without demanding any further input [23]. Because there are fewer disruptions overall, the experience is more seamless and fluid. Users must not constantly enter passwords or credentials to work, browse, or shop. Passive authentication appeals to consumers and service providers since it balances usability and safety by increasing convenience without sacrificing security.

D. Security Benefits

Passive behavioral authentication improves security by detecting abnormalities and possible breaches in real-time. Since it monitors a user's actions at all times during a session, it can quickly identify any variations from normal behavior that would point to unauthorized access [24]. For instance, if an attacker manages to access a device or account, their actions—like typing, navigating the mouse, or using the device—will probably differ from those of a genuine user [25]. The system can detect these variations immediately and then take appropriate action—locking the account, alerting the user, or requesting more verification, for example. An essential line of defense against risks like session hijacking—where an attacker gets access after the user logs in—is provided by this real-time monitoring. Traditional authentication techniques, such as passwords, provide no defense against mid-session threats because they only verify identification at the beginning of a session. This vulnerability is addressed by passive behavioral authentication, which provides proactive defense against insider

and external threats by continuously confirming the user. This makes it an effective tool for corporate systems, financial services, and healthcare, requiring maximum security with the least disturbance.

4 Comparison to Traditional Authentication

A. Security Improvements

Traditional password-based systems have long been the cornerstone of digital security but come with significant vulnerabilities. Passwords can be stolen, guessed, or compromised through phishing attacks, and once breached, an attacker can gain full access to a system. Additionally, many users reuse passwords across multiple platforms, amplifying the risk if one system is breached. Behavioral biometrics that are passive and continuous enhance security by offering dynamic, multi-layered authentication [26]. In contrast to passwords, behavioral biometrics rely on an individual's distinct behavior patterns, such as mouse clicks, typing rhythms, and device handling. Because they would need to duplicate not just one credential but the whole behavioral profile of the authentic user, this makes it considerably more difficult for attackers to spoof. Furthermore, even if an attacker manages to obtain access, their distinct interaction patterns will allow for detection thanks to continuous authentication, which continuously observes activity throughout a session. One significant benefit of real-time anomaly detection over standard one-time authentication is that, if suspicious activity is discovered, it can quickly set off alarms or require re-authentication, thereby avoiding illegal access in the middle of a session.

B. User Experience

One of the main criticisms of traditional authentication methods is the friction they introduce to the user experience. Passwords, PINs, and biometric scans require active user input, interrupting workflow or online interactions. Repeated logins and re-authentications can frustrate users, particularly in environments where high-security demands frequent credential checks [27]. Especially in passive and continuous forms, behavioral biometrics provide a much more smooth and convenient experience. These techniques authenticate users in the background, obviating the need for user intervention [28]. For instance, passive monitoring is carried out on typing patterns, mouse movements, and touchscreen motions when people perform their jobs. As a result, fewer login attempts or extra authentication procedures are required, which minimizes disruptions while preserving a high level of security.

Continuous behavioral authentication significantly increases user happiness by lowering barriers and improving usability, especially in settings where security usually causes friction (e.g., online banking and corporate networks). As a result, a less intrusive and more secure system is created, providing a seamless user experience that is difficult to do with conventional techniques.

C. Case Studies

Several industries have successfully implemented continuous and passive behavioral biometrics, with notable improvements in security and user experience. Here are a few examples:

- **Banking and Financial Services:** The internet platforms of major banks have been equipped with passive behavioral biometrics to identify fraudulent transactions and unapproved account access. These systems track user behavior on banking apps, including typing, swiping, and navigating, to identify patterns that might point to identity theft or account takeover[29]. Since most consumers are unaware that they are being continuously validated, case studies have demonstrated a significant decrease in fraud with little to no impact on the user experience.
- **E-commerce:** Behavioral biometrics are employed in online retail to protect transactions without interfering with the customer's shopping experience[30]. E-commerce platforms, for example, monitor customers' mouse movements, scrolling patterns, and keyboard patterns as they shop, add products to their carts, and complete the checkout process. Because of this, users can access accounts with varied interaction patterns, and the system can detect potentially fraudulent activity without requiring them to go through additional authentication processes. Retailers have reported higher consumer satisfaction and decreased fraud rates.
- **Corporate Security:** Behavioral biometrics are utilized in the workplace to safeguard confidential company information and systems. Employees' typing, mouse movements, and interactions with internal platforms are continuously used to authenticate them[31]. In one case study, a sizable organization used continuous authentication to monitor employee sessions, thereby reducing insider threats and unlawful access. The solution improved security without interfering with workflow by enabling the identification and thwarting of possible breaches in real-time.

These case studies demonstrate that continuous and passive behavioral biometrics are not just theoretical solutions—they have been proven effective in real-world scenarios, offering robust security with minimal impact on user experience.

5 Challenges and Limitations

A. Data Privacy and Ethical Concerns

There are serious ethical and data privacy issues when collecting behavioral data for authentication reasons. By its very nature, behavioral biometrics is the ongoing gathering and examination of users' distinct interaction patterns, such as the rhythm of their typing, the movements of their mouse, or the way they grip their device [32]. Although this data is extremely valuable for security, it also raises concerns about the appropriate level of monitoring and the handling, storage, and use of such sensitive data. Even for security reasons, being continuously watched over may seem

intrusive to many people. It is imperative to guarantee that data is gathered with informed consent, preserved securely, and utilized exclusively for authentication. Striking a balance between enhancing security and protecting user privacy is one of the most pressing ethical challenges in implementing passive and continuous behavioral authentication systems. This often involves adhering to stringent data protection regulations, such as GDPR, and implementing transparent data usage policies.

B. Accuracy and False Positives

While behavioral biometrics have demonstrated considerable potential, accuracy remains a challenge, especially in real-world, diverse environments. Erroneous detections, or false positives, occur when valid users are labeled imposters due to behavioral patterns being altered by factors such as user weariness, stress, or device modifications [33]. When users are forced to continually re-authenticate or verify their identity, these errors might negatively impact their experience. However, a security risk is associated with false negatives, which occur when an attacker imitates a user's behavior just enough to evade detection. A critical area of ongoing study is lowering false positives and raising the overall accuracy of continuous authentication systems. Improvements are reducing errors in machine learning models, which can recognize and adjust to minute behavioral changes. However, the systems still encounter difficulties in contexts where behavior might vary considerably.

C. Scalability

Another significant limitation is the scalability of behavioral biometric systems, especially when deployed across diverse platforms and user populations [34]. Since every user's behavioral profile differs, large-scale applications like corporate networks and banking platforms with millions of users require systems to process and store enormous volumes of data in real time [35]. Furthermore, variations in usage situations (e.g., office vs. mobile) and technology (e.g., different computers or cell-phones) can impact the precision of biometric readings, making it challenging to deploy a universal solution that functions flawlessly for everyone. Developing scalable, platform-neutral systems that can manage this degree of complexity without sacrificing accuracy and performance is a significant obstacle to the advancement of behavioral biometrics in the future. Solutions to effectively handle and analyze large datasets may include integrating distributed computing and cloud-based infrastructures.

D. Adversarial Attacks

Like any security system, behavioral biometric systems are not immune to adversarial attacks. Attackers might try to imitate or modify a user's behavioral habits to trick continuous authentication systems. An attacker might attempt to mimic the victim's actions, for instance, if they have access to video footage of them typing or using a gadget. We call this kind of attack behavioral spoofing. Furthermore, to get around the system, more cunning opponents might employ machine learning to examine and replicate a user's biometric data [36]. Developers are developing anti-spoofing

techniques and systems that recognize unusual efforts to alter behavioral patterns to counter such attacks. However, since there is an ongoing arms race between attackers and defenders, behavioral biometric systems must continuously improve to counter new adversarial threat types and stay effective.

6 Conclusion

Integrating continuous behavioral biometrics and passive authentication offers a robust alternative to traditional password-based security systems. As the system architecture depicts, this approach combines real-time user behavior monitoring (e.g., typing patterns, gestures) with contextual signals (e.g., location, time, device information) to dynamically validate user identities. The Behavioral Biometrics Engine and Passive Authentication Engine work in tandem, analyzing both active and passive data streams to detect anomalies, which the Authentication System further assesses. A Security Alert System alerts users and administrators if suspicious behavior is detected, enabling timely responses to potential threats. This multi-layered system enhances security by continuously adapting to changes in user behavior and context, offering real-time fraud detection and reducing the reliance on static credentials like passwords. By leveraging behavioral and contextual data, this adaptive authentication model provides enhanced protection, particularly for sensitive financial applications, where security breaches can be costly.

References

1. Stragapede, G., Vera-Rodriguez, R., Tolosana, R., Morales, A., Acien, A., Le Lan, G.: Mobile behavioral biometrics for passive authentication. *Pattern Recognit. Lett.* **157**, 35–41 (2022). <https://doi.org/10.1016/j.patrec.2022.03.014>
2. Abuhamad, M., Abusnaina, A., Nyang, D., Mohaisen, D.: Sensor-Based Continuous Authentication of Smartphones' Users Using Behavioral Biometrics : A Contemporary Survey, pp. 1–19
3. Ann, P., Preetha, T.K.: A broad review on non-intrusive active user authentication in biometrics. *J. Ambient Intell. Humans. Comput.* **14**(1), 339–360 (2023). <https://doi.org/10.1007/s12652-021-03301-x>
4. Finnegan, O.L., et al.: The utility of behavioral biometrics in user authentication and demographic characteristic detection : a scoping review. *Syst. Rev.* 1–17 (2024). <https://doi.org/10.1186/s13643-024-02451-1>
5. Gaddekar, B., Hiwarkar, T.: A critical evaluation of business improvement through machine learning: challenges, opportunities, and best practices. *Int. J. Recent Innov. Trends Comput. Commun.* **11**(10s), 264–276 (2023). <https://doi.org/10.17762/ijritcc.v11i10s.7627>
6. Gunuganti, A.: Behavioral Biometrics for Continuous Authentication, vol. 1, no. 3, pp. 1–5 (2023)
7. Peng, G., et al.: Continuous authentication with touch behavioral biometrics and voice on wearable glasses. *IEEE Trans. Hum.-Mach. Syst.* **47**(3), 404–416 (2017). <https://doi.org/10.1109/THMS.2016.2623562>

8. Survey, A.: Security, Privacy, and Usability in Continuous Authentication: A Survey, pp. 1–26 (2021)
9. Pathak, G.R., Patil, S.H.: Mathematical model of security framework for routing layer protocol in wireless sensor networks. *Phys. Procedia* 78(December 2015), 579–586 (2016). <https://doi.org/10.1016/j.procs.2016.02.121>
10. Fraz, A., Sigurd, B., Bian, E.: Privacy-preserving continuous authentication using behavioral biometrics. *Int. J. Inf. Secur.* **22**(6), 1833–1847 (2023). <https://doi.org/10.1007/s10207-023-00721-y>
11. Wang, C.: Behavioral Authentication for Security and Safety, vol. 3 (2024)
12. Gadekar, B.P., Hiwarkar, T.: A conceptual modeling framework to measure the effectiveness using ML in business analytics. *Int. J. Adv. Res. Sci. Commun. Technol.* 2(1), 399–406 (2022). <https://doi.org/10.48175/ijarsct-7703>
13. Junquera-Sánchez, J., Cilleruelo, C., De-Marcos, L.: Access Control beyond Authentication, vol. 2021 (2021). <https://doi.org/10.1155/2021/8146553>
14. Pathak, G.R., Premi, M.S.G., Patil, S.H.: LSSCW: a lightweight security scheme for cluster-based Wireless Sensor Network. *Int. J. Adv. Comput. Sci. Appl.* 10(10), 448–460 (2019). <https://doi.org/10.14569/ijacsa.2019.0101062>
15. Chandre, P.R., Mahalle, P.N., Shinde, G.R.: Machine learning based novel approach for intrusion detection and prevention system: a tool based verification. In: 2018 IEEE Global Conference on Wireless Computing and Networking (GCWCN), pp. 135–140 (2018). <https://doi.org/10.1109/GCWCN.2018.8668618>
16. Abuhamad, M., Abusnaina, A., Member, G.S., Nyang, D., Mohaisen, D., Member, S.: Sensor-based continuous authentication of smartphones' users using behavioral biometrics : a contemporary survey. *IEEE Internet Things J.* **8**(1), 65–84 (2021). <https://doi.org/10.1109/JIOT.2020.3020076>
17. Peng, G., et al.: Continuous authentication with touch behavioral biometrics and voice on wearable glasses. *IEEE Trans. Hum.-Mach. Syst.* **47**(3), 404–416 (2017)
18. Hernández-Álvarez, L., De Fuentes, J.M., González-Manzano, L.: Privacy-Preserving Sensor-Based Continuous Authentication, pp. 1–23 2(021)
19. Kotwal, J., Kashyap, D.R., Pathan, D.S.: Agricultural plant diseases identification: from traditional approach to deep learning. *Mater. Today Proc.* **80**(xxxx), 344–356 (2023). <https://doi.org/10.1016/j.matpr.2023.02.370>
20. Li, S., Iqbal, M., Saxena, N.: Future industry internet of things with zero-trust security. *Inf. Syst. Front.* (2022). <https://doi.org/10.1007/s10796-021-10199-5>
21. Jawale, A., Warole, P., Bhandare, S., Bhat, K., Chandra, R.: Jeevn-Net: brain tumor segmentation using cascaded U-net & overall survival prediction. *Int. Res. J. Eng. Technol.* 56–62 (2020)
22. Vorokhob, M., Kyrychok, R., Yaskevych, V., Dobryshyn, Y., Sydorenko, S.: Modern perspectives of applying the concept of zero trust in building a corporate information security policy. *Cybersecur. Educ. Sci. Tech.* **1**(21), 223–233 (2023). <https://doi.org/10.28925/2663-4023.2023.21.223233>
23. Weinberg, A.I., Cohen, K.: Zero Trust Implementation in the Emerging Technologies Era: Survey, no. 2021 (2024). <http://arxiv.org/abs/2401.09575>
24. Kore, V.S., Tidke, B.A., Chandra, P.: Survey of image retrieval techniques and algorithms for image-rich information networks. *Int. J. Comput. Appl.* **112**(6), 39–42 (2015). <https://www.ijcaonline.org/archives/volume112/number6/19674-1244/>, <https://research.ijcaonline.org/volume112/number6/pxc3901244.pdf>
25. Abdalla Mahmoud, A., Elisha Nyamasvisva, T., Valloo, S.: Zero Trust Security Implementation Considerations in Decentralised Network Resources for Institutions of Higher Learning Transmitter development for Oil Exploration in Offshore Environment View project, no. June (2022). <https://www.researchgate.net/publication/361595829>
26. Nyamasvisva, T.E., Abdalla, A., Arabi, M.: A comprehensive swot analysis for zero trust network security model. *Int. J. Infrastruct. Res. Manag.* **10**(1), 44–53 (2022). <https://iukl.edu.my/rmc/publications/ijirm/>

27. Damre, S.S., Shendkar, B.D., Kulkarni, N., Chandre, P.R., Deshmukh, S.: Smart healthcare wearable device for early disease detection using machine learning. *Int. J. Intell. Syst. Appl. Eng.* **12**(4s), 158–166 (2024)
28. Federici, F., Martintoni, D., Senni, V.: A Zero-Trust Architecture for Remote Access in Industrial IoT Infrastructures. *Electronics* **12**(3) (2023). <https://doi.org/10.3390/electronics12030566>
29. He, Y., Huang, D., Chen, L., Ni, Y., Ma, X.: A survey on zero trust architecture: challenges and future trends. *Wirel. Commun. Mob. Comput.* **2022** (2022). <https://doi.org/10.1155/2022/6476274>
30. Zhang, Y.: Privacy-preserving with zero trust computational intelligent hybrid technique to english education model. *Appl. Artif. Intell.* **37**(1) (2023). <https://doi.org/10.1080/08839514.2023.2219560>
31. Oh, S.R., Seo, Y.D., Lee, E., Kim, Y.G.: A comprehensive survey on security and privacy for electronic health data. *Int. J. Environ. Res. Public Health* **18**(18) (2021). <https://doi.org/10.3390/ijerph18189668>
32. Sarkar, S., Choudhary, G., Shandilya, S.K., Hussain, A., Kim, H.: Security of zero trust networks in cloud computing: a comparative review. *Sustainability* **14**(18) (2022). <https://doi.org/10.3390/su141811213>
33. Pinto, S., Santos, N.: Demystifying arm trust zone: a comprehensive survey. *ACM Comput. Surv.* **51**(6) (2019). <https://doi.org/10.1145/3291047>
34. Dhotre, D., Chandre, P.R., Khandare, A., Patil, M., Gawande, G.S.: The rise of crypto malware: leveraging machine learning techniques to understand the evolution, impact, and detection of cryptocurrency-related threats. *Int. J. Recent Innov. Trends Comput. Commun.* **11**(7), 215–222 (2023). <https://doi.org/10.17762/ijritcc.v11i7.7848>
35. Ahmid, M., Kazar, O.: A comprehensive review of the Internet of Things security. *J. Appl. Secure. Res.* **18**(3), 289–305 (2023). <https://doi.org/10.1080/19361610.2021.1962677>
36. Ghasemshirazi, S., Shirvani, G., Alipour, M.A.: Zero Trust : Applications, Challenges, and Opportunities (2023). <https://arxiv.org/abs/2309.03582>

AI for Fraud Prevention and Threat Intelligence

A Greedy Hybrid Ensemble Approach for Security Applications: Fraud, Intrusion, and Malware Detection



Monika Mangla, Nonita Sharma, Madhuchhanda Tripathy, Vaishali Mehta, and Manik Rakhra

Abstract This manuscript presents a thorough analysis of the ensemble model created using the Greedy Approach focusing specifically for security domain. The objective of this research is to achieve the optimal combination of the 5 base classifiers to achieve optimal performance metrics. The proposed greedy ensemble approach is simulated on the credit card dataset towards determining frauds, to validate its effectiveness and efficiency. After analysis of the obtained results, it is evident that it achieves F1 score of 0.83 which is substantially higher than 0.79 of the random approach. Thus, it can be safely concluded that the greedy approach can be utilized for devising ensemble modeling. Thus, the current research work can be considered as a contributing step towards practical security issues such as fraud detection, malware classification, and intrusion detection.

Keywords Ensemble · Greedy optimization · Bagging · Boosting · F1 score

M. Mangla (✉)

Dawarkadas J. Sanghvi College of Engineering, Mumbai, India
e-mail: manglamona@gmail.com

N. Sharma

Indira Gandhi Delhi Technical University for Women, New Delhi, India
e-mail: nsnonita@gmail.com

M. Tripathy

Intel Technology, Bengaluru, India
e-mail: madhuchhanda.tripathy@intel.com

V. Mehta

Maharishi Markandeshwar deemed to be University, Mullana, Ambala, India
e-mail: dr.vaishalimehta@mmumullana.org

M. Rakhra

Lovely Professional University, Phagwara, India
e-mail: manik.23538@lpu.co.in

1 Introduction

Ensemble learning is a well-known approach in machine learning (ML) that combines the output from many models to produce a common set of outputs depending on a set of parameters [1]. Results are substantially more accurate when ensemble techniques are used. While some ML models excel at modelling a certain characteristic of the data, others excel at modelling a different component of the data. It combines various models to improve the stability and prognostication power of the model. Each model's variances and biases are offset by the robustness of the models as a whole as shown in Fig. 1. This offers a composite prediction with a final accuracy that surpasses that of the individual models [1]. Considering the efficiency of ensemble modeling, it has been widely employed in various domains namely education, healthcare, agriculture and many more.

Ensemble methods can be divided into two groups: Sequential ensemble techniques and Parallel ensemble techniques. Sequential ensemble techniques involve training the base learners consecutively. One example of a sequential ensemble is Boosting where base models depend on the output of each other. The output of one base model is sequentially fed to the other. All base learners may not be given equal importance as incorrectly classified data points may be given more importance in training the next model. Gradient boosting and Adaptive Boosting (AdaBoost) are common boosting algorithms [2].

Parallel ensemble techniques involve training the base learners parallelly [3]. A common example of a parallel ensemble is Bagging. It trains multiple weak learners such as Decision Tree in parallel and outputs the aggregated result of all base models. The sample generated to train each weak learner may or may not be identified as samples are fetched with replacement technique. Hence, base models can be trained independently. Random Forest is the common bagging algorithm that overcomes the limitations of overfitting in its base learner.

In ML, optimization refers to changing hyper-parameters to enhance the functionality of the algorithm [4]. The trade-off between bias and variance is determined by the hyperparameters. The process of optimization involves several strategies. The cost function is minimized using the Gradient Descent technique, which lowers the error [5]. The objective is to achieve a local minimum where the cost function can

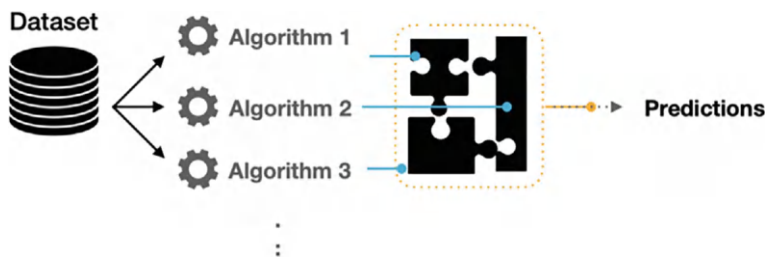


Fig. 1 Ensembling approach

no longer be decreased. It is done if global optimum cannot be achieved. This is done by determining the hyperparameter learning rate, leading to determination of step-size at each iteration. Choosing a learning rate which is neither too low nor too high, is a crucial step in this process [6]. During stochastic gradient descent, data points are chosen randomly during each iteration, making gradient descent a random component. The idea of evolution is inspired by genetic algorithms. Here, certain models that produce inferior results are discarded while good models are retained. To further improve the results, these models are altered with the help of other models. Models are referred to as variations of the earlier form when this process is repeated several times.

The objective of this research work is to create a new ensemble model using greedy optimization techniques in order to find an ensemble model yielding the optimum performance metrics such as accuracy, precision, recall, and F1-score. The main objective of the current research work is as follows:

- Comparison of various classifiers on the basis of performance metrics like accuracy, precision, recall and F1 score.
- Determining various simulations applied to execute the proposed greedy approach to develop an ensemble model using a combination of classifiers.
- Comparison of the proposed hybrid ensemble model to the base model.

This paper is organized as follows: Sect. 2 describes the related work, Sect. 3 explains the methodology, Sect. 4 is the analysis of research questions, Sect. 5 concludes research work.

2 Related Work

Several researchers have worked on the optimization of base learners. The tabular comparison of the state-of art work done in this area is given in Table 1 which clearly demonstrates the efficacy of ensemble modeling in the various domains highlighting its significance [7].

The comparative analysis clearly establishes the efficacy of ensemble modeling over base learners in various domains [15]. Carrying out the research further, authors in current research work presents the applicability of greedy optimization on ensemble modeling and shows the superiority of the proposed method by simulating on the credit card dataset. The prime objective or motive behind selection of credit card dataset is unprecedented growth in the domain of fraudulent transactions during the past decade. Hence, an efficient method in this domain is of paramount significance considering its impact on social and economical reputation of a nation at international level.

Table 1 Comparative evaluation of the state-of-art in ensemble modelling

Citation	Year	Dataset	Methods	Results	Main contribution	Research gap
[8]	2024	<ul style="list-style-type: none"> Popular regression benchmark datasets from KEEL Various sizes of regression datasets evaluated 	<ul style="list-style-type: none"> Greedy optimization strategy Boosting negative correlation learning framework 	<ul style="list-style-type: none"> Improved regression accuracy Enhanced generalization performance 	Proposed Greedy Deep Stochastic Configuration Networks Ensemble	Not defined
[9]	2024	<ul style="list-style-type: none"> Kaggle Dataset 	<ul style="list-style-type: none"> Ensemble Technique: Stacking 	<ul style="list-style-type: none"> 99.6% accuracy and MAE of 0.003 	To advance the preventive maintenance system	Higher computational complexity
[10]	2023	<ul style="list-style-type: none"> Breast Cancer Dataset 	<ul style="list-style-type: none"> Snapshot Ensemble Technique 	<ul style="list-style-type: none"> Accuracy of 86.6%, higher than the state-of-art 	t-distributed stochastic neighbor embedding employed for dimensionality reduction	Tradeoff between accuracy and cost
[11]	2022	<ul style="list-style-type: none"> Vinho Verde wine dataset 	<ul style="list-style-type: none"> Classification techniques: Random Forest, Decision Tree, KNN, ANN 	<ul style="list-style-type: none"> Dataset reduction without impact on performance Random Forest outperforms all classifiers 	Reduced dataset size from 13 to 9 attributes	Impact of preprocessing on performance
[12]	2022	<ul style="list-style-type: none"> Benchmark time series data sets 	<ul style="list-style-type: none"> Hybrid layered based greedy ensemble reduction (HLGER) 	<ul style="list-style-type: none"> Superior generalization performance 	Developed hybrid layered based greedy ensemble reduction architecture	Predictor deletion based on lowest accuracy and diversity
[13]	2022	<ul style="list-style-type: none"> 8 Datasets taken from PROMISE and ECLIPSE repository 	<ul style="list-style-type: none"> Sequential Ensemble Modelling 	<ul style="list-style-type: none"> Better prediction and lesser values of Error Metrics 	Developed sequential ensemble model for software fault prediction	Optimization of maintenance cost
[13]	2021	<ul style="list-style-type: none"> Not mentioned 	<ul style="list-style-type: none"> Branch and Bound Algorithm 	<ul style="list-style-type: none"> High Performance Lesser no. of rules 	Better Generalization Performance	May lead to overfitting
[14]	2018	<ul style="list-style-type: none"> Dataset of potentially fake wines synthesized from real samples 	<ul style="list-style-type: none"> Bayesian network classifiers Multilayer perceptron Sequential minimal optimization Ensemble Learning: Bagging and Boosting 	<ul style="list-style-type: none"> Ensemble learning improves BNC and MLP classifiers significantly 	Improved detection of high-value wine forgeries	Not mentioned

3 Methodology

The prime goal of this research is to construct a classification model, which will boost the accuracy and dependability of forecasts. Results of classification models are trained using 5 ML models namely Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Naive Bayes (NB), Decision Tree (DT) which are combined to create a hybrid ensemble model using a greedy approach on one of its performance metrics. This ensemble approach outperforms the base models in terms of performance, utilizing a step-by-step methodology. The dataset used for this research work is pertaining to Credit Card Fraud Detection Dataset taken from Kaggle (<https://www.kaggle.com/datasets/yashpaloswal/fraud-detection-credit-card>). The dataset includes credit card transactions performed by European cardholders in September 2013. There were 492 frauds out of 284,807 transactions in this dataset of transactions that took place over the course of 2 days. It only has input variables that are numbers. The original features and further background information about the data are concealed due to confidentiality concerns.

3.1 Preprocessing

During building an optimal hybrid ensemble model, the data needs to be preprocessed. This Credit Card Fraud dataset consists of 31 columns and 284,907 data rows. To implement the preprocessing, there are mainly 3 steps namely removal of incorrect values, missing values and the outlier. Although there are various approaches to manage missing values. The most popular method for removing rows and columns is to use the pandas 'dropna()' function. This technique is not needed in the considered dataset as considered dataset has no missing values. Second preprocessing step is dimensionality reduction which retains only the significant features removing the least significant features conserving the computational requirements. Here, the downloaded dataset has already gone under this process and the columns from V1 till V28 have been reduced to the current form. The last method in preprocessing is to convert the data from categorical values to numeric values. Since the dataset consists of only numeric data, this step can also be skipped.

3.2 Training

In this research work, authors have split the dataset into 85% for training and 15% for testing giving `test_size = 0.15`, `random_state = 0` and `stratify = y`.

3.3 Optimization Technique

During current research work, an incremental greedy approach is adopted to determine the best suited ensemble model on the chosen dataset. F1 score is the property on which the greedy approach will follow.

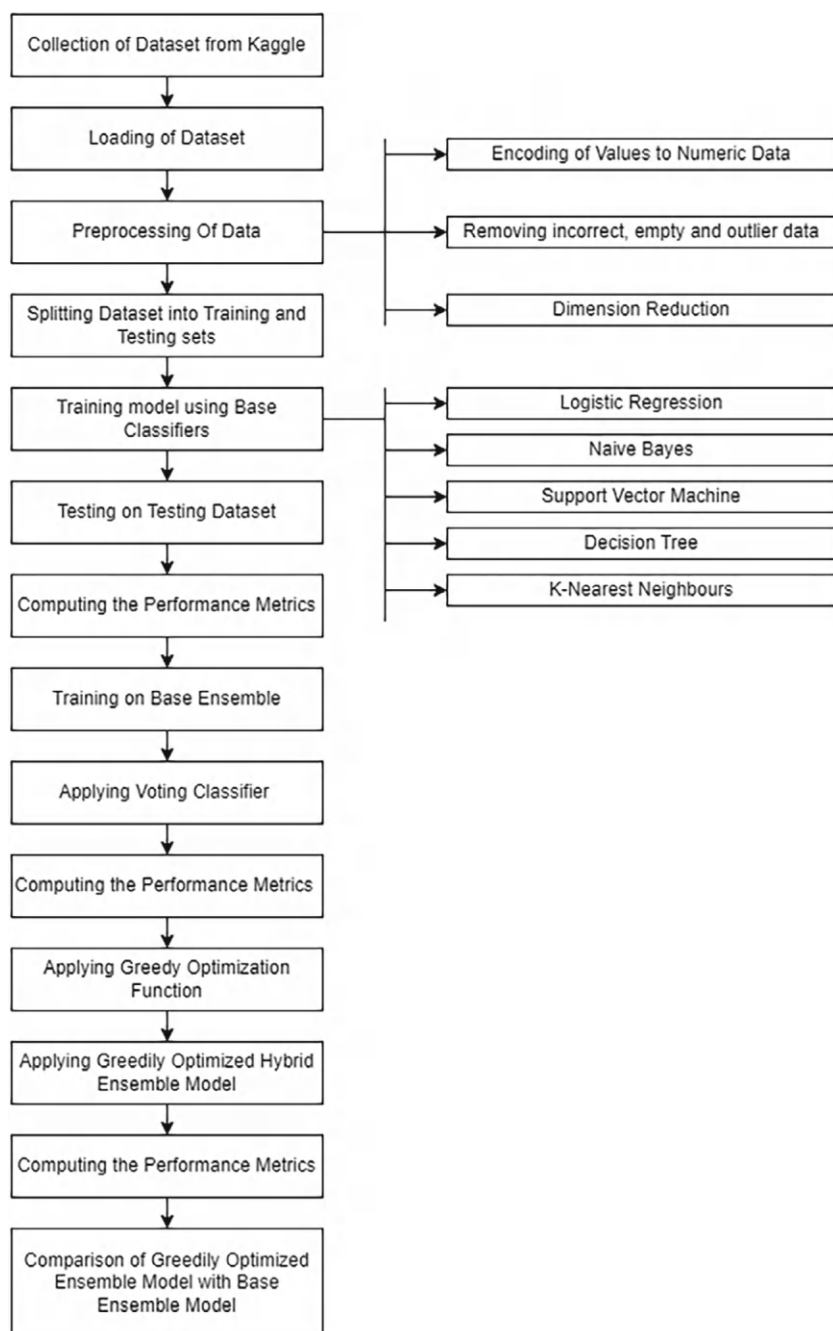
In the first step, results from base models are taken into consideration and they are sorted in decreasing order of accuracy. This is done because models having better performance is preferred over other models. As outputs of each step are taken into account before going to the next step, this can be called a variation of Boosting Technique. However, current research work has one restriction that iterations of a single base model for ensemble models is limited to 2. This means that a model can occur only in 3 ways: Not present at all, present once and present 2 times. This limits the chances of getting stuck in an infinite loop while selecting the models greedily.

The optimization technique used in creating the final ensemble model is Greedy Approach. Using this approach, each model is added to the ensemble model array sequentially and on the basis of the F1 score produced at each iteration, the picking of the classifier is done. This process is continued till the number of base learners in the sorted array exhaust. Finally, results from the generated ensemble are combined using a maximum vote approach.

3.4 Proposed Approach Algorithm

This section contains a basic algorithm of the adopted optimization technique. Certain parts of the algorithm are assumed to be calculated with the help of helper functions. The flowchart for proposed methodology is illustrated in Fig. 2.

Algorithm of Greedy Ensemble:
1. Create an empty array of size n, where n is the number of individual classifiers.
2. Insert the F1 score of each classifier in the array.
3. Sort the array in decreasing order.
4. Pick the first element of the array, and add to the estimator array as the base ensemble.
5. Compute the base model F1 Score.
6. Pick the next element from the array and add to the Estimator array.
7. Compute the F1 score.
8. If the F1 score is greater than previously obtained F1 score, add this model to the final ensemble. Else, Discard this iteration.
9. Repeat Step 6,7, and 8 for Twice the same model.
10. Repeat steps from Step 6 for each of the subsequent models in the array
Final Time Complexity : <i>O(nlogn)</i>

**Fig. 2** Proposed methodology

4 Results and Discussion

This section discusses the different research objectives as mentioned earlier. The dataset is split into 85% training samples and 15% testing samples.

(A) *Comparison of various classifiers on the basis of performance metrics like accuracy, precision, recall and F1 score.*

This section presents the performance metrics of various base classifiers. This comparative analysis is carried out to determine the order in which base classifiers must be considered during development of ensemble model. The performance metrics viz. Accuracy, Presion, Recall and F1 score are illustrated in Table 2 which clearly illustrates that KNN yields best performance and thus it will be considered first among all 5 base ML models. The same is pictorially illustrated in Fig. 3.

Table 2 Comparative analysis of base models

Classifier	Accuracy	Precision	Recall	F1 score
LR	0.99	0.84	0.71	0.77
NB	0.97	0.06	0.81	0.11
SVM	0.99	0.95	0.59	0.73
KNN	0.99	0.96	0.71	0.82
DT	0.99	0.80	0.71	0.75

Comparative Analysis

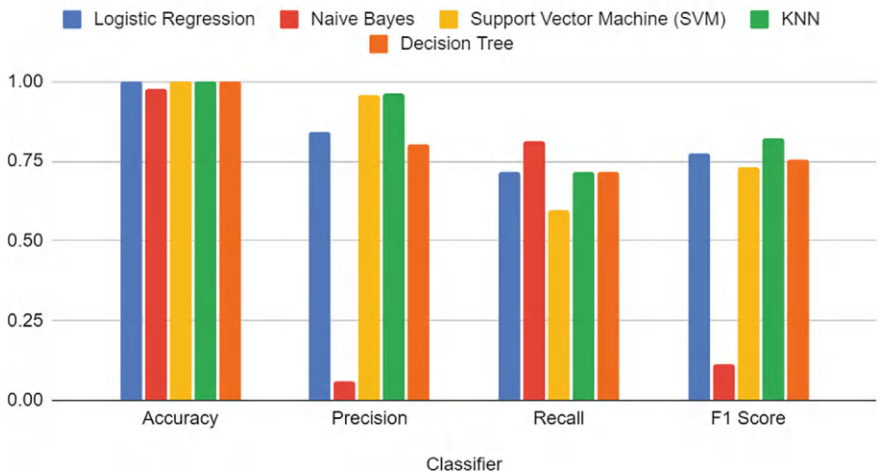


Fig. 3 Visualization of comparative analysis of base models

(B) *Determining various simulations applied to execute the proposed greedy approach to develop an ensemble model using a combination of classifiers?*

This research work uses the greedy algorithm to create a hybrid ensemble model which yields the maximum F1 score. Since, there are 5 classifiers, each will be tested twice for including them in the final model. Thus, there will be 10 iterations in total as illustrated in Table 3. Thus, the final model obtained from the above-mentioned approach is: 2KNN + 1 LR + 2 DT + 1 NB, yielding the F1 score as 0.833. Hence, this is the optimum value of the F1 score that can be reached using the above-mentioned hybrid ensemble model.

(C) *Comparison of the proposed hybrid ensemble model to the base model.*

This research work takes the base model computed by repeating the classifiers in random fashion. This model is created to draw a comparison between the optimal approach and other random approaches. Table 4 provides a statistical comparison between these two models on the basis of F1 score. The same is graphically illustrated in Fig. 4.

Now, from the results shown in this section, it is pretty clear that ensemble approach outperforms base ML models. Further, greedy ensemble model is compared with randomly created ensemble model which further strengthens the effectiveness of greedy approach. From the results, it can be assuredly claimed that greedy ensemble approach is an effective method and hence can be applied in various domains in real life.

Table 3 Iterations of the model

Iteration no.	Model used	Recall	F1 score
1	1KNN	0.82	Model accepted
2	2KNN	0.82	Model accepted
3	2KNN + 1LR	0.82	Model accepted
4	2KNN + 2LR	0.79	Model rejected
5	2KNN + 1LR + 1DT	0.81	Model rejected
6	2KNN + 1LR + 2DT	0.83	Model accepted
7	2KNN + 1LR + 2DT + 1SVM	0.82	Model rejected
8	2KNN + 1LR + 2DT + 2SVM	0.82	Model rejected
9	2KNN + 1LR + 2DT + 1NB	0.83	Model accepted
10	2KNN + 1LR + 2DT + 2NB	0.79	Model rejected

Table 4 Comparative analysis of greedy approach and random approach

Model used	Recall	F1 score
7 LR + 5 DT + 5 SVM + 5 KNN + 5 NB	Random approach	0.79
2 KNN + 1 LR + 2 DT + 1 NB	Greedy approach	0.83

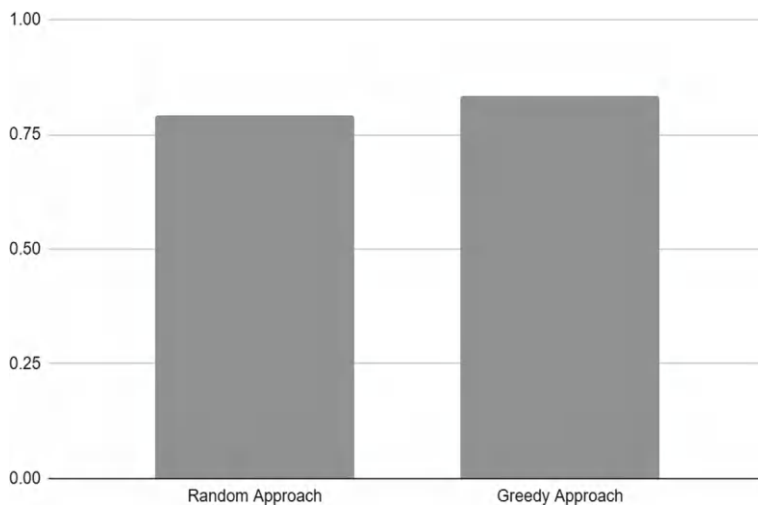


Fig. 4 Comparison of random approach versus greedy approach

5 Conclusion

This research work presents a thorough analysis of the ensemble model created using the Greedy Approach. The objective of this research was to identify the optimal combination of the 5 classifiers which can yield the maximum F1 Score. For the same, authors proposed a greedy ensemble approach which is simulated on credit card dataset. In order to validate the effectiveness of proposed method, it is evaluated on credit card dataset. The produced results are compared with base ML models and randomly chosen ensemble model. The obtained results demonstrate and validate the efficacy of proposed approach and hence it can yield an optimal solution to predict the fraudulent cases. The application of proposed approach may surely go beyond this scenario as classification tasks are frequently carried out in numerous domains namely agriculture, healthcare, education and many more.

References

1. Sharma, N., Dev, J., Mangla, M., Wadhwa, V. M., Mohanty, S. N., Kakkar, D.: A heterogeneous ensemble forecasting model for disease prediction. *New Gener. Comput.* 1–15 (2021)
2. Sharma, N., Mangla, M., Yadav, S., Goyal, N., Singh, A., Verma, S., Saber, T.: A sequential ensemble model for photovoltaic power forecasting. *Comput. Electr. Eng.* **96**, 107484 (2021)
3. Yeo, M., Fletcher, T., Shawe-Taylor, J.: Machine learning in fine wine price prediction. *J. Wine Econ.* **10**(2), 151–172 (2015)
4. https://www.researchgate.net/publication/350110244_Prediction_of_Wine_Quality_Using_Machine_Learning_Algorithms

5. Lee, S., Park, J., & Kang, K.: Assessing wine quality using a decision tree. In: 2015 IEEE International Symposium on Systems Engineering (ISSE) (2015). <https://doi.org/10.1109/syseng.2015.7302752>
6. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decis. Support. Syst.* **47**(4):547–553, (2009). (Elsevier)
7. Shanmuganathan, S., Sallis, P., Narayanan, A.: Data mining techniques for modelling seasonal climate effects on grapevine yield and wine quality. In: IEEE International Conference on Computational Intelligence, Communication Systems and Networks, pp. 82–89 (2010)
8. Zhang, C., Wang, Y., Zhang, D.: Greedy deep stochastic configuration networks ensemble with boosting negative correlation learning. *Inf. Sci.* (2024). <https://doi.org/10.1016/j.ins.2024.121140>
9. Ojha, E., Sharma, N., Mangla, M.: An ensemble framework for predictive maintenance of WaterPumps. In: 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, pp. 552–558 (2024). <https://doi.org/10.1109/IDCIoT59759.2024.10467905>
10. Sharma, N., Sharma, K.P., Mangla, M., Rani, R.: Breast cancer classification using snapshot ensemble deep learning model and t-distributed stochastic neighbor embedding. *Multimed. Tools Appl.* **82**(3), 4011–4029 (2023)
11. Parneeta, D., Suyash, Mani S., Lakshay C.: Detailed study of wine dataset and its optimization. *Int. J. Intell. Syst. Appl.* Undefined (2022). <https://doi.org/10.5815/ijisa.2022.05.04>
12. Lin, Q.: Metaheuristic Ensemble Pruning via Greedy-Based Optimization Selection. *Int. J. Appl. Metaheuristic Comput.* Undefined (2022). <https://doi.org/10.4018/ijamc.292501>
13. Mangla, M., Sharma, N., Mohanty, S.N.: A sequential ensemble model for software fault prediction. *Innov. Syst. Softw. Eng.* **18**(2), 301–308 (2022)
14. Ye, R., Dai, Q.: A novel greedy randomized dynamic ensemble selection algorithm. *Neural Process. Lett.* Undefined (2017). <https://doi.org/10.1007/S11063-017-9670-Y>
15. Portinale, L., Locatelli, M.: Investigating the Role of Ensemble Learning in High-Value Wine Identification (2018)

Optimizing Ensemble Models for Security Applications: A Comparative Study of Greedy and Dynamic Approaches



Monika Mangla, Nonita Sharma, Saumyaranjan Acharya, Vaishali Mehta, and Manik Rakhra

Abstract Current research work aims to explore three different ensemble optimization techniques namely Base Ensemble, Dynamic Ensemble Selection (DES), and Greedy Optimization. This analysis is carried out to determine the trade-off among predictive accuracy and the computational requirement in Ensemble modeling. This work also seeks to investigate optimization paradigms –Dynamic Ensemble Selection Performance (DESP), K-Nearest Oracles Eliminate (KNORA-E) and K-Nearest Oracles Union (KNORA-U) in the context of DES and growing and pruning strategies based on Greedy Optimization. Considered ensemble optimization techniques have been experimentally implemented on the credit card dataset and the results reveal that Greedy Optimization with growing and pruning strategies outperforms the other ones by achieving the highest accuracy of 0.72. Thus, results advocate the significance of greedy optimization and hence can be implemented in various applications including security-critical applications such as identification of threat mitigations that requires high accuracy and low computational power.

Keywords Ensemble model · Classification · Optimization · Greedy · Dynamic · Growing · Pruning · Accuracy

M. Mangla (✉)

Dawarkadas J. Sanghvi College of Engineering, Mumbai, India

e-mail: manglamona@gmail.com

N. Sharma

Indira Gandhi Delhi Technical University for Women, New Delhi, India

S. Acharya

SAP Labs, Bengaluru, India

V. Mehta

Maharishi Markandeshwar deemed to be University, Mullana, Ambala, India

e-mail: dr.vaishalimehta@mmumullana.org

M. Rakhra

Lovely Professional University, Phagwara, India

e-mail: manik.23538@lpu.co.in

1 Introduction

In the recent past, a growing interest has been seen in the applicability of ensemble models owing to their ability of creating an optimal productive model by combining various machine learning (ML) models [1]. Ensemble modeling aim to produce one optimal predictive model by combining several base models. Examples of simplest ensemble techniques are max voting, averaging and weighted averaging [2]. In max voting, prediction by each base model is considered as a vote and the class with maximum votes is the final prediction. In weighted averaging, models are given preference in terms of weights i.e., higher prioritized models are assigned higher weights and the one with lower preference is given lower weight [3]. These assigned weights are considered to evaluate the weighted average which is used for final prediction. The advanced ensemble techniques are stacking, bagging and boosting [4].

The optimization techniques for ensemble aim to apply various strategies to enhance performance and accuracy in ML. Several techniques such as momentum methods, aggregation methods, hybrid optimization and structural changes of neural networks are proposed by authors in Sharma et al. [5]. These techniques are primarily focusing on improvement of selection process, hyper-parameters tuning, and leveraging diversity among models aiming to achieve better generalization but at lesser computational costs. This can be achieved by modifying the algorithm so that it runs in less time with limited resources [6].

As discussed earlier, current research work focuses on exploring 3 optimization techniques in order to find the best trade-off among predictive accuracy and the computational requirement. In DES, a prediction is made by automatically selecting a portion of the ensemble members. Greedy optimization techniques improve accuracy of ensemble methods in ML by strategically selecting base learners and combining them to optimize predictive performance [7].

This research work also aims to overcome the limitation in investigating optimization paradigms in the context of DES and growing and pruning strategies based on Greedy Optimization. DES based models are the models which utilize the nearest neighbor technique to achieve dynamic classifier selection for every instance. In DESP, the classifiers are ranked according to a predefined schema. The approaches KNORA-E and KNORA-U use a model where the best performing classifiers are selected based on the k-nearest neighbors of the given instance. Greedy Optimization models are based on both adding and removing.

This work examines the advantages and disadvantages of both DES and Greedy Optimization approaches in order to determine the best way to enhance ensemble models. Current research work is organized into various sections. The objective of study is put in Sect. 1 while the related work carried out by different researchers is presented in Sect. 2. Proposed methodology is well elaborated in Sect. 3 while results are discussed in Sect. 4. The conclusion and future work is presented in Sect. 5.

2 Related Work

Several researchers have worked on the optimization of base learners. The tabular comparison of the state-of art work done in this area is given in Table 1. The presented table shows the efficacy of ensemble modeling in the various domains and highlight the significance of applying optimization in ensemble modeling [8]. The authors in this research work present the applicability of greedy optimization on ensemble modeling and shows the superiority of the proposed method by simulating it on the credit card dataset.

3 Methodology

The major goal of this research is to construct a classification model to boost the accuracy and dependability of forecasts. Suggested ensemble model uses 5 base ML models namely Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Naive Bayes (NB), and Decision Tree (DT) which are combined using a greedy approach. Using greedy approach, a model is added if it outperforms the current performance metrics. Following are 5-steps used in proposed methodology as illustrated in Fig. 1.

3.1 Data Collection

The dataset used for this research work is Credit Card Fraud Detection Dataset (<https://www.kaggle.com/datasets/yashpaloswal/fraud-detection-credit-card>).

3.2 Data Preprocessing

In order to clean data so that it produces accurate results using proposed model, missing value and noise needs to be eliminated from the data through data preprocessing techniques. In order to preprocess, libraries such as sklearn—train_test_split, StandardScaler and MinMaxScaler are imported with the reading dataset. Also, rows containing missing values are deleted. Finally, the data is split into training and testing sets.

Table 1 Comparative evaluation of the state-of-art in ensemble modelling

Citation	Year	Dataset	Methods	Results	Applications	Identified research gap
[9]	2024	• Reservoir Test Case	• 4 Momentum methods combined with ensemble gradient approximation	• Improved net present value • Fewer Simulations required	Petroleum Reservoir Management	Limited exploration of momentum strategies in Ensemble Modeling
[10]	2024	• Kaggle Dataset	• Optimized Weighted Averaging • Aggregation over minimum	• Better Accuracy • Improved Robustness	Functional Optimization Problem	Lesser Research on Optimization in Ensemble methods
[11]	2024	• Water Pump Dataset	• Deep Copy Stacking Technique	• Improved Accuracy • Faster • Lesser Computational Complexity	Industry Management	Trade-off between Accuracy and Complexity
[12]	2023	• Generalized Dataset	• Stacking Ensemble Technique	• Ensemble methods optimizes BN effectively • No impact of Hyperparameter Tuning	Feature Integration Strategies	Limited impact of hyperparameter tuning
[13]	2023	• Wind Speed Assimilation Dataset • Koteweg-de Vries-Burgers Model Dataset	• Newton Method • Conjugate-Gradient Method	• Newton Method converges faster	Data Assimilation Experiments	Not stated
[14]	2022	• Benchmark time series data sets	• Hybrid layered based greedy ensemble reduction (HLGER)	• Superior generalization performance	Developed hybrid layered based greedy ensemble reduction architecture	Predictor deletion based on lowest accuracy and diversity

(continued)

Table 1 (continued)

Citation	Year	Dataset	Methods	Results	Applications	Identified research gap
[15]	2022	<ul style="list-style-type: none">8 datasets taken from PROMISE and ECLIPSE repository	<ul style="list-style-type: none">Sequential Ensemble Modelling	<ul style="list-style-type: none">Better prediction and lesser values of Error Metrics	Developed sequential ensemble model for software fault prediction	Optimization of maintenance cost
[16]	2021	<ul style="list-style-type: none">Not Mentioned	<ul style="list-style-type: none">Branch and Bound Algorithm	<ul style="list-style-type: none">High PerformanceLesser no. of rules	Better Generalization Performance	May lead to overfitting
[17]	2018	<ul style="list-style-type: none">Dataset of potentially fake wines synthesized from real samples	<ul style="list-style-type: none">Bayesian network classifiersMultilayer perceptronSequential minimal optimizationEnsemble Learning: Bagging and Boosting	<ul style="list-style-type: none">Ensemble learning improves BNC and MLP classifiers significantly	Improved detection of high-value wine forgeries	Not mentioned

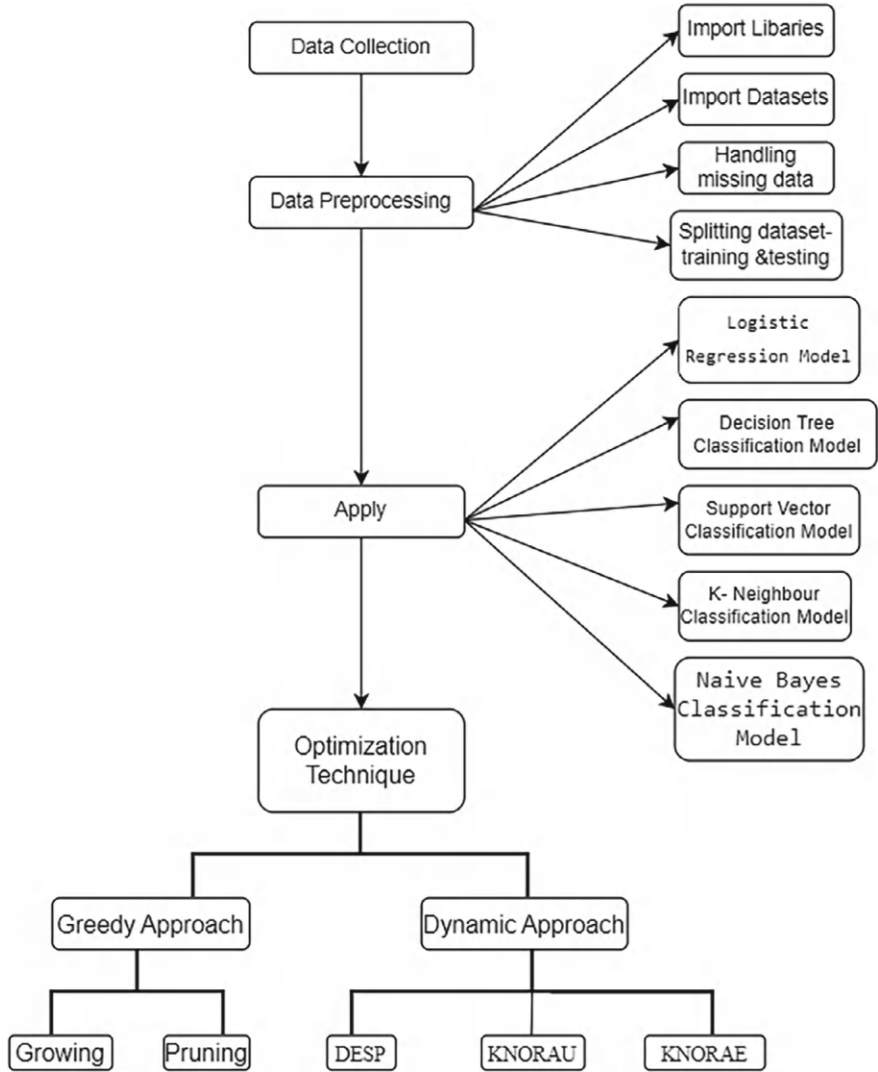


Fig. 1 Proposed Methodology

3.3 Ensemble Model

As mentioned earlier, the proposed work uses 5 base models and the ensemble model is created using greedy approach. Here, the performance metrics of 5 base models are compared and multiple copies of outperforming base model is included in the ensemble model as long as it improves the performance. When it demonstrates a

decline in the performance, the next best performing base model is included in the ensemble model.

3.4 Optimization Technique

To optimize the ensemble model, greedy ensemble and dynamic ensemble approached are tried. In the greedy model authors used 2 different methods namely growing and pruning. Here, growing starts with no model and model is added to the ensemble model based on results i.e., models are added in a greedy manner. It provides enhanced efficiency as a small number of models can provide escalated accuracy. On the contrary, pruning starts by including all the models and models are removed from ensemble model if it deteriorates its efficiency.

Thus, the list of all the models is maintained and from the list of the models, the model is added or removed, based on the performance. The dynamic ensemble methods used are DESP, KNORA-E, KNORA-U. It selects subsets of the model just-in-time. KNORA-E and KNORA-U uses the K-nearest neighbor approach to locate data closest to predicted value. KNORA-E reduces neighbor value to the smallest value while KNORA-U selects all the neighboring classifiers which have at least one correct value. DESP, uses FIRE-DES selection schema to select most competent classifiers.

3.5 Algorithms

This section presents the algorithm of the adopted optimization technique. The performance of each model in the ensemble is evaluated using the historical records stored. DESP is implemented in a k-nearest neighbors (KNN) fashion, whereby a set of similar training examples is selected for every new instance. Further on, it assesses models' performances using those analogous training examples. Models that performed well in that region are used to forecast the current event. This technique modifies the ensemble according to each instance by focusing the attention on models that have performed well for similar instances previously, which might lead to better predictive accuracy with different datasets.

Input:

1. A pool of classifiers $\mathcal{C} = \{c_1, c_2, \dots, c_k\}$
2. A new instance x to be classified
3. A set of training instances with known labels
4. A parameter k for the number of nearest neighbors

Algorithm of DESP:

Step 1: Identify the **k-nearest neighbors** of x from the training set using a distance metric.

Step 2: For each classifier c_i in \mathcal{C} , evaluate the performance of c_i on the k-nearest neighbors by calculating the number of correct predictions.

Step 3: Select a subset of classifiers $C' \subseteq C$ that achieved the best performance on the k-nearest neighbors.

Step 4: For the selected subset C' , obtain the predictions for the new instance x .

Step 5: Combine the predictions of the classifiers in C' to get the final prediction for x .

Algorithm of KNORA-E:

Step 1: Identify the k-nearest neighbors of x from the training set using a distance metric.

Step 2: For each classifier c_i in \mathcal{C} , Check if c_i correctly classifies all k-nearest neighbours, if yes add c_i to C'

Step 3: If no classifier meets the strict criterion; gradually relax the condition by reducing the number of neighbours that each classifier must correctly classify until at least one classifier is selected.

Step 4: For the selected subset C' , obtain the predictions for the new instance x .

Step 5: Combine the predictions of the classifiers in C' to get the final prediction for x .

Algorithm of KNORA-U:

Step 1: Identify the k-nearest neighbors of x from the training set using a distance metric.

Step 2: For each classifier c_i in \mathcal{C} , Check if c_i correctly classify at least one k-nearest neighbours, if yes add c_i to C'

Step 3: For the selected subset C' , obtain the predictions for the new instance x .

Step 4: Combine the predictions of the classifiers in C' to get the final prediction for x .

3.6 Applications of Security in Ensemble Optimization

This section outlines some domains where ensemble optimization can be employed in real-life particularly in security related applications.

1. **Fraud Detection in Financial Systems:** Ensemble models optimized with DESP, KNORA-E, KNORA-U, and Greedy Optimization are particularly effective at identifying fraudulent transactions in financial systems. These methods can be employed to scrutinize extensive transaction datasets and enhance detection precision through the integration of numerous classifiers.
2. **Intrusion Detection Systems (IDS):** These methods can improve Intrusion Detection Systems by combining classifiers to identify network intrusions with increased accuracy. Dynamic selection can adaptively identify the most effective models for recognizing novel patterns in cyberattacks.
3. **Malware Detection:** Optimization techniques can be utilized to categorize files as malicious or benign by employing datasets containing information derived from file metadata, network activity, and system logs.
4. **Spam Filtering:** Spam detection in emails or texts can utilize ensemble learning to discern patterns of phishing or malicious material, thereby minimizing false positives and enhancing security.
5. **Biometric Authentication Systems:** For systems employing fingerprints, facial recognition, or other biometric data, ensemble models enhance classification accuracy by integrating multiple base classifiers.

This widened list serves as the motivation behind undertaking the research work of optimizing an efficient ensemble model. Although the authors have validated the effectiveness of proposed work using credit card related dataset but the same can be employed in any domain.

4 Results and Discussion

This section discusses and present the visualization of the results. As mentioned earlier, the dataset is related to credit cards and have been collected from kaggle. The dataset is divided into 85% training samples and 15% testing samples.

4.1 Results of Ensemble Modelling

Authors have implemented 5 ML models—linear regression classification model, decision tree classification model, support vector classification model, k-nearest neighbor classification model and naive bayes classification model. Out of the 5 models, SVM classification model provides highest accuracy of 0.74 while naive bayes gives lowest accuracy of 0.40. Although, precision for both these models are comparable i.e., 0.67 and 0.7. The accuracy of proposed ensemble is 0.70. The performance metrics of base 5 models and ensemble model is presented in Table 2 and Fig. 2.

Table 2 Comparative analysis of base models

	LR	DT	SVC	KNN	NBC	Ensemble
Accuracy	0.67	0.73	0.74	0.69	0.40	0.70
Precision	0.65	0.65	0.67	0.65	0.7	0.7
Recall	0.68	0.69	0.72	0.7	0.41	0.41
F1-score	0.66	0.65	0.65	0.65	0.37	0.37

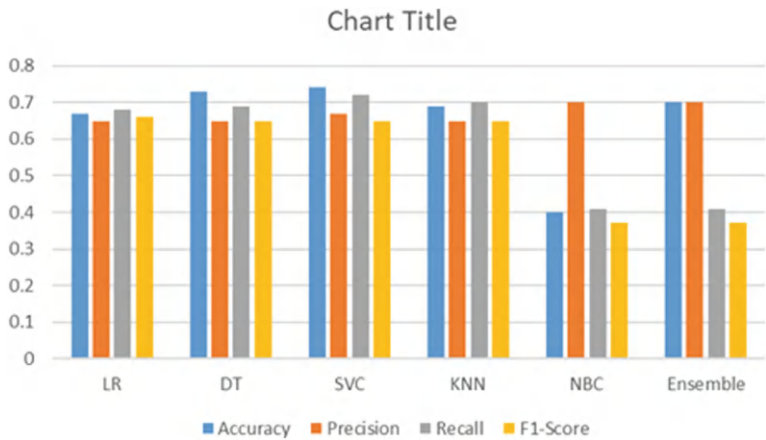


Fig. 2 Visualization of comparative analysis of base models

4.2 Simulation Results of Various Techniques

Authors optimized the ensemble model using greedy and dynamic approach. Greedy approach used two methods namely growing and pruning. The dynamic ensemble methods used are DESP, KNORA-E, KNORA-U. Out of the 2 optimized techniques, greedy approach provided the highest level of accuracy of 0.72. Apart from accuracy, greedy method also yielded highest precision, recall and F1-score as shown in Table 3. The same is graphically illustrated in Fig. 3. It is worth noting that although the Dynamic approach also obtained higher accuracy than the base ensemble, it is lower than the greedy ensemble approach. Out of 3 dynamic techniques, KNORA-U yielded the lowest accuracy of 0.705 in comparison to 0.71 given by other two methods.

4.3 Results of Proposed Model

As mentioned earlier, the performance yielded by Greedy algorithms shows highest precision, recall and F1-score. While the base ensemble model gives the accuracy of 0.703 as demonstrated in Table 4 and Fig. 4.

Table 3 Comparative analysis of Greedy and Dynamic Ensembling techniques

	Ensemble	Greedy		Dynamic		
		Pruning	Growing	DESP	KNORA-U	KNORA-E
Accuracy	0.70	0.72	0.72	0.71	0.70	0.71
Precision	0.7	0.66	0.65	0.5	0.48	0.5
Recall	0.41	0.71	0.67	0.53	0.43	0.53
F1-score	0.37	0.65	0.66	0.51	0.45	0.51

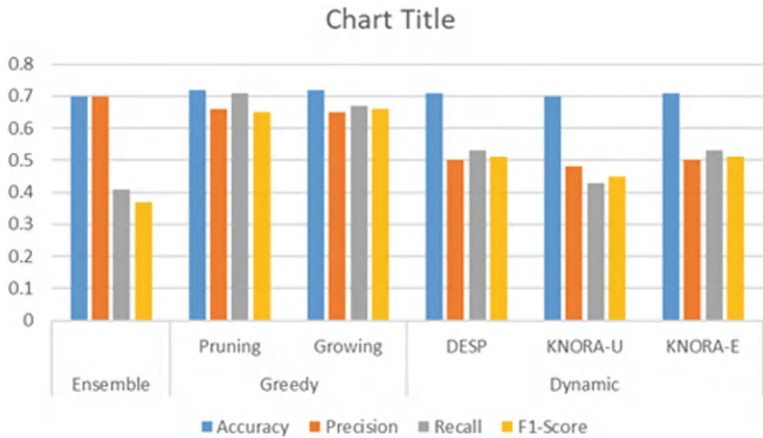


Fig. 3 Visualization of comparative analysis of ensembling techniques

Table 4 Comparative analysis of ensembling techniques

	Ensemble	Greedy	
		Pruning	Growing
Accuracy	0.70	0.72	0.72
Precision	0.7	0.66	0.65
Recall	0.41	0.71	0.67
F1-score	0.37	0.65	0.66

Thus the comparative analysis demonstrated in this section clearly indicates that ensemble model always outperforms base ML models. Although greedy model gives better performance in comparison to dynamic model; dynamic model outperforms base ensemble models. Thus this research work clearly advocates the efficacy and effectiveness of ensemble model over base ML models.

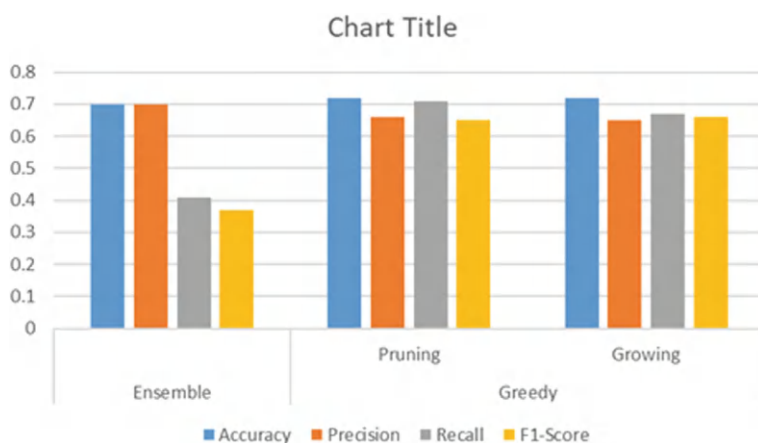


Fig. 4 Comparison of random approach versus greedy approach

5 Conclusion

This study presents a comprehensive evaluation of ensemble optimization techniques to improve classification accuracy using a credit card fraud detection dataset. The prime objective of the research is to validate the effectiveness of ensemble modeling. It is clearly demonstrated using the experimental evaluation that Greedy Optimization approach, employing both growing and pruning strategies, achieved the highest accuracy at 0.72 outperforming dynamic ensemble methods. This makes Greedy Optimization particularly suitable for applications demanding precision with manageable computational costs. The results indicate that while DES techniques are valuable for instance-based dynamic model selection, Greedy Optimization provides a more computationally efficient alternative for real-world predictive tasks. These findings highlight the significance of optimization in ensemble modeling, suggesting Greedy Optimization as an effective strategy for enhanced model performance across diverse datasets. Future work could explore the integration of hybrid methods or additional datasets to further refine ensemble strategies and expand their application across various domains.

References

1. Sharma, N., Dev, J., Mangla, M., Wadhwa, V. M., Mohanty, S. N., Kakkar, D.: A heterogeneous ensemble forecasting model for disease prediction. *New Gener. Comput.* 1–15 (2021)
2. Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. *Front. Comp. Sci.* **14**, 241–258 (2020)
3. Mahajan, A., Sharma, N., Aparicio-Obregon, S., Alyami, H., Alharbi, A., Anand, D., Sharma, M., Goyal, N.: A novel stacking-based deterministic ensemble model for infectious disease prediction. *Mathematics* **10**(10), 1714 (2022)

4. Dietterich, T.G.: Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems, pp. 1–15. Springer, Berlin (2000)
5. Sharma, R., Sharma, N., Kumar, A.: Enhancing IoT botnet detection through machine learning-based feature selection and ensemble models. *EAI Endorsed Trans. Scalable Inf. Syst.* **11**(2) (2024)
6. Mangla, M., Shinde, S. K., Mehta, V., Sharma, N., Mohanty, S.N. (Eds.): *Handbook of Research on Machine Learning: Foundations and Applications*. CRC Press (2022)
7. Yadav, S., Sharma, N.: Homogenous ensemble of time-series models for Indian stock market. In: Mondal, A., Gupta, H., Srivastava, J., Reddy, P., Somayajulu, D. (eds.) *Big Data Analytics. BDA 2018. Lecture Notes in Computer Science()*, vol. 11297. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04780-1_7
8. Lee, S., Park, J., Kang, K.: Assessing wine quality using a decision tree. In: 2015 IEEE International Symposium on Systems Engineering (ISSE) (2015). <https://doi.org/10.1109/syseng.2015.7302752>
9. Nilsen, M.M., Stordal, A.S., Lorentzen, R.J., Raanes, P.N., Eikrem, K.S.: Accelerated Ensemble Optimization using Momentum Methods (2024). <https://doi.org/10.21203/rs.3.rs-4648690/v1>
10. Cervellera, C., Macciò, D., Sanguineti, M.: *Ensemble Aggregation Approaches for Functional Optimization*. AIRO Springer Series (2024). https://doi.org/10.1007/978-3-031-47686-0_18
11. Ojha, E., Sharma, N., Mangla, M.: An ensemble framework for predictive maintenance of WaterPumps. In: 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, pp. 552–558 (2024). <https://doi.org/10.1109/IDCIoT59759.2024.10467905>
12. Tan, P.: Ensemble-based Hybrid Optimization of Bayesian Neural Networks and Traditional Machine Learning Algorithms (2023). <https://doi.org/10.21203/rs.3.rs-3312122/v1>
13. Bertsimas, D., Boussiou, L.: Ensemble modeling for time series forecasting: an adaptive robust optimization approach (2023). <https://doi.org/10.48550/arXiv.2304.04308>
14. Sharma, N., Mangla, M., Mohanty, S.N., Pattanaik, C.R.: Employing stacked ensemble approach for time series forecasting. *Int. J. Inf. Technol.* **13**, 2075–2080 (2021)
15. Mangla, M., Sharma, N., Mohanty, S.N.: A sequential ensemble model for software fault prediction. *Innovations Syst. Softw. Eng.* **18**(2), 301–308 (2022)
16. Zhang, C., Wang, Y., Zhang, D.: Greedy deep stochastic configuration networks ensemble with boosting negative correlation learning. *Inf. Sci.* (2024). <https://doi.org/10.1016/j.ins.2024.121140>
17. Portinale, L., Locatelli, M.: *Investigating the Role of Ensemble Learning in High-Value Wine Identification* (2018)

Strategic Deployment of Machine Learning in Combating Email Spam and Cyber Threats



Sarita Mohanty and Anupa Sinha

Abstract With the popularity of email as a significant communication tool in the digital environment, the emergence of cyber threats seeking to exploit email as an essential service comes through email scams commonly referred to as email spam that has proven to be a vector for more serious attacks such as phishing or malware distribution. This paper examines using several machine learning (ML) models in fighting against email-based threats and how they can bolster cybersecurity defenses. This study uses traditional models such as Logistic Regression and Decision Trees, as well as advanced algorithms, including Random Forests, Gradient Boosting, MLP, GRU, and LSTM, to generate a detailed analysis of each of these models concerning the accuracy, precision, recall, and F1 score. Furthermore, the performance of these models is optimized using Particle Swarm Optimization (PSO). Our results demonstrate the need to tailor cybersecurity frameworks to continuously adapt to the escalating cyber threat landscape by adopting advanced ML techniques.

Keywords Machine learning · Email spam · Cybersecurity · Particle swarm optimization · Neural networks · Logistic regression · Random forest

1 Introduction

With the increase of cyber threats, mainly by email spam in the age of widespread digital communication, it has become a significant concern of cybersecurity. Spam has become the favored vector for malicious entities to exploit email systems, often the cornerstone of both personal and professional life in conveying a catalog of cyber attacks, from phishing to malware deployment. Traditional spam detection and prevention methods have struggled to keep up with the complexity and volume

S. Mohanty (✉) · A. Sinha

Department of Computer Science and Engineering, Kalinga University, Naya Raipur, India
e-mail: mohantysarita104@gmail.com

A. Sinha

e-mail: anupa.sinha@kalingauniversity.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
M. Yang et al. (eds.), *Demystifying AI and ML for Cyber-Threat Intelligence*,
Information Systems Engineering and Management 43,
https://doi.org/10.1007/978-3-031-90723-4_14

203

of email-based threats, which continue to grow. The adequacy of such fixed and isolated security measures demands evolving paradigms to accommodative, more dynamic, and adaptive security measures. The growing problem of Agile planning becomes more and more tractable using Machine learning (ML), a technique able to learn from and adapt to new data, without explicit programming [1, 2].

Email spam, often defined as unsolicited and almost always malicious content, not only fills up inboxes but is also used to launch more damaging security breaches. Spam is dynamic, and its enormous volume undermines the efficiency of conventional, often rule-based, spam detection systems, which cannot evolve as quickly as spam tactics. Spam emails have consequently improved their sophistication to spoof legitimate communications [3, 4], making them more challenging to detect and dangerous.

During the rapid growth of the Internet of Things (IoT) in recent decades, the Internet of Things has become an integral part of modern life. It has become a cornerstone for developing smart cities and the basis for different social media platforms and applications. Aditya et al. [5] and Aski and Sourati [6] parallel the proliferation of IoT with an increase in spam, posing serious challenges against cybersecurity efforts globally.

Many researchers have devised ways of detecting and stomping out spam and spammers. These methods generally fall into two main categories: Behavioral pattern-based and semantic analysis-based. While each category is effective in its place, each has drawbacks and limitations. More spam emails are created, which can, in theory, be anonymized anywhere in the world; as the internet and global communications networks grow, so does the volume of malicious emails [7].

Spam remains prevalent despite the development of sophisticated anti-spam technologies. Spams containing links to malicious websites—where personal information can be stolen from innocent people—are particularly harmful. In addition, spam emails can bring down the server's memory and processing- capacity, making server response slower. Realizing this, organizations must rigorously evaluate spam detection technologies and determine spam to the extent of network intrusion identification and elimination. Common practices include allowlists and blocklists, mail header analysis, and keyword verifications to filter incoming emails [8, 9].

For example, 40 percent of all social network users engage in spamming. Spammers use popular social networking platforms to attack specific sectors, reviews, and fan pages, deliberately inserting malicious links in the seemingly innocuous content. Malicious emails tend to share similar traits, however, primarily if they target the same audiences or targets. By analyzing these characteristics, we can increase email detection and classification ability for various types of emails. The process heavily depends on AI extracting features from message headers, subjects, and bodies to classify them as spam or legitimate [10, 11]. However, there is an excellent need for state-of-the-art machine-learning algorithms that can be used to detect mail.

Spam and GPMS

For now, spam detection often relies on learning-based classifiers. The spam emails bypass classical classification techniques due to these patterns, which are unique

to spam emails. Nevertheless, the application of these learning-based models in the field of spam detection is cumbersome due to several factors, including the subjective nature of spam, concept drift, linguistic nuances, processing overhead, and delays in text analysis [12].

Machine learning in cybersecurity employs the strengths of pattern recognition and anomaly detection that are common in this field and necessary for detecting and fighting email threats. Because of vast amounts of data to sort through and differentiate benign from malicious emails, machine learning algorithms can find and analyze hidden patterns or anomalies invisible to human-minded analysts. One of the main reasons for being ML-based is that ML-based systems allow the continuous update of their models in response to new threats as a fundamental basis [13, 14].

This research examines several machine learning models to see which ones work best for spam detection. We look at the old standards like Decision Trees and Logistic Regression, the newer computational models like Support Vector Machines (SVM), and neural network designs like MLP, GRU, and LSTM. The models' non-linear data handling characteristics, processing speed, sensitivity, and specificity differ. Central to our work is a comprehensive evaluation of the models, including six fundamental performance measures essential for gauging each model's effectiveness in real-world settings: accuracy, precision, recall, and F1 score [15].

Lastly, although there are several obstacles to using machine learning models for spam detection, the following demonstrates that advancements in the field may still be made without compromising on interest, utility, or quality in the future. They are the dataset demands for training and the computational required to get processed or optimized. In addition, this introduction explores how optimization techniques like Particle Swarm Optimization (PSO) can be employed strategically to improve the performance of machine learning models. Further, using PSO enables us to fine-tune the algorithm parameters that better facilitate our model's ability to classify and predict unseen data [16] accurately.

Machine learning integration with cybersecurity operations has the potential to revolutionize how spam filters detect and respond at a speed and scale previously unattainable. Strategically, this deployment sits squarely in the context of the modern cybersecurity frameworks that tend to desire to counteract threats preemptively instead of reactively. This paper provides a call for employing advanced techniques to build more resilience. And proactive defenses against email spam that keep growing [17].

2 Literature Review

For the last 10 years, email spam, commonly known as bulk or unsolicited email, has emerged as a cyber security issue. Spambots that scrape email addresses from the Internet have made traditional spam filters less effective; however, spurred by increasingly sophisticated spam tactics and the need for more sophisticated solutions that attempt to catch spammers by their syntax. Because there are strong models

and methodologies for creating new spam detection and filtering systems, machine learning is a huge step forward in the spam detection field [18].

Deep learning methods for cyber security, including intrusion detection systems and spam detection databases, are reviewed in this overview by Tait et al. [19]. Here, they test their models on 35 widely-used cyber datasets, which are classified according to the kinds of traffic they include, such as data sent over the Internet or a network. According to Tait et al. [19], deep learning models outperform standard machine or lexicon-based intrusion and spam detection models. This suggests that deep learning might help combat sophisticated cyber threats.

In their study on supervised learning methods for spam email delivery, Vyas et al. [20] found that, while the Naïve Bayes strategy yields results relatively fast, machines such as SVM and ID3 provide high accuracy but grow somewhat more slowly. Here, emphasize the trade-off between efficiency and accuracy while selecting a machine learning method for spam identification.

Machine learning models for spam detection cannot afford to do feature selection. Yang et al. [21] used numerous supervised learning methods and demonstrated using the N-Gram algorithm for feature selection. One specific classifier strongly suited for text analysis and spam detection is N-Grams, which predict the probability of the next word in a sequence depending on the previous words.

Their work reviewed the existing approaches to email spam filtering and summarised the effectiveness of different proposed systems and the accuracy measurements under various parameters. Based on the many blogs that address email spam filtering, I checked with a few datasets like ECML and UCI datasets. I discussed the prevalence of Naïve Bayes and SVM algorithms used to filter emails. The review of their work draws attention to the repeatedly finding ways to upgrade spam bug filtering precision and react to the changing email attacks.

In their work presented in the survey of intelligent spam email detection models, artificial immune systems were presented. For example, they noted that supervised learning algorithms are exceptionally adopted because of their accuracy and consistency. Moreover, the study highlighted the efficiency of multi-algorithm frameworks over single-algorithm solutions and showed that more effective spam detection strategies could be inferred by combining different methods.

Zhuang et al. [22] described different learning-based spam filtering approaches. Other features were explained in spam emails, and economic and ethical spam issues were reviewed. According to their study, the speed and high accuracy of the Naïve Bayes classifier make it particularly effective among different learning algorithms, constituting an indispensable reference for building and improving learning.

The strand of literature on machine learning applications for spam detection shows a clear trend toward more sophisticated, efficient, and adaptive solutions. The more complex and abundant spam becomes, the more complicated and abundant the content being labeled becomes, and the more critical machine learning (particularly advanced machine learning models, such as deep learning and ensemble methods) becomes. These studies illustrate the importance of dynamic defenses, mainly driven by machine learning approaches—to a dynamic threat landscape and

provide a solid basis for ongoing research and application in machine learning-driven spam detection.

3 Methodology

To assess how well different machine learning classifiers identify spam emails, this study's methodology section details the analytical methodologies and processes employed. This all-inclusive study of machine learning strategies for fighting email-based cyber threats covers experimental design, data collection, model training, optimization, and performance assessment.

Experimental Setup and Data Collection

Our study relies on a primary dataset of labeled email messages, with spam versus non-spam, extracted from publicly available spam detection repositories. The primary datasets used include:

Enron Dataset: Large sets of emails from employees at the Enron Corporation have been widely used for machine learning tasks such as spam detection and are available as this dataset. An example of a mixed set of corporate and spam emails that serves to test spam detection algorithms is this.

Spam Assassin Corpus: This corpus is another key resource used to study publicly available marked content for spam emails. The datasets available relate to diverse email characteristics, including content-based attributes such as the body text and metadata attributes such as sender information and header details.

Classifier Selection and Training

Several machine learning classifiers were used for text classification and spam detection effectiveness. The standard data split was used to train each classifier (75% sampling for training, 25% sampling for test). That implementation ran with Python's sci-kit-learn and TensorFlow libraries, which are standard for machine learning tasks in the industry.

Performance Metrics

The effectiveness of each classifier was assessed using key performance metrics: We experiment with accuracy, recall, precision, F1-score, and training time. This kind of metrics enables the evaluation and comparison of classifier performance in terms of efficiency and effectiveness in spam detection, i.e., evaluating and comparing one classifier with the other.

Application of Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) was used to classify performance improvement further. As a bioinspired optimization technique, PSO is applied to the hyperparameters tuning of the machine learning model for high precision between spam and legitimate emails.

Data Visualization and Analysis

Results were visualized using various charts and graphs to depict the performance of the classifiers clearly.

Bar Charts: Fitted to compare accuracy, recall, precision, and F1-score among multiple classifiers.

Confusion Matrices: The true positives, false positives, true negatives, and false negatives are provided for each of the Random Forest and SVM classifiers as further details about its performance.

Horizontal Bar Chart: Showed training times of each classifier and illustrated trade-offs in the accuracy and computational efficiency.

ROC Curves: As classifiers, they showed each locator's diagnostic ability and effectiveness in discriminating spam from non-spam emails judged by the area under the curve (AUC).

We can evaluate the capabilities and limits of different machine learning classifiers for performing spam detection through this approach. This research tests the efficacy of traditional and contemporary machine learning models through structured training, optimization, and evaluation. Still, it also provides critical insights for their practical use in email security for dealing with current and future cyber threats.

4 Result and Discussion

This study designates a comprehensive evaluation of the machine learning classifiers for detecting and reducing cyber threats within emails. A series of metrics and visualization techniques are applied to empirical evidence supporting the strategic use of these classifiers.

This bar chart shown in Fig. 1 meticulously displays the accuracy, recall, precision, and F1-score of several classifiers, including Logistic Regression, KNN, Decision Trees, Extra Trees, Random Forest, Gradient Boosting, MLP, GRU, LSTM, for comparison. This figure shows that all of the classifiers do a good job.

Of detecting spam emails and contributing significantly to making email security better. No one classifier can guarantee an effective spam detection system with low false favorable rates and high accurate positive rates.

This confusion matrix, as shown in Fig. 2, displays the results of the Random Forest classifier, which is an essential tool in the spam detection arsenal. The matrix shows the classifier's performance in real-life scenarios, which includes the number of true positives, true negatives, false positives, and false negatives. If you want to know where the deployed machine learning model is strong and where it is weak, you need these thorough breakdowns.

The following horizontal bar chart in Fig. 3 reveals the training times of various classifiers used for spam detection. In environments where resources and response times are the paramount factors to achieve, it is essential to understand the computational efficiency of each classifier.

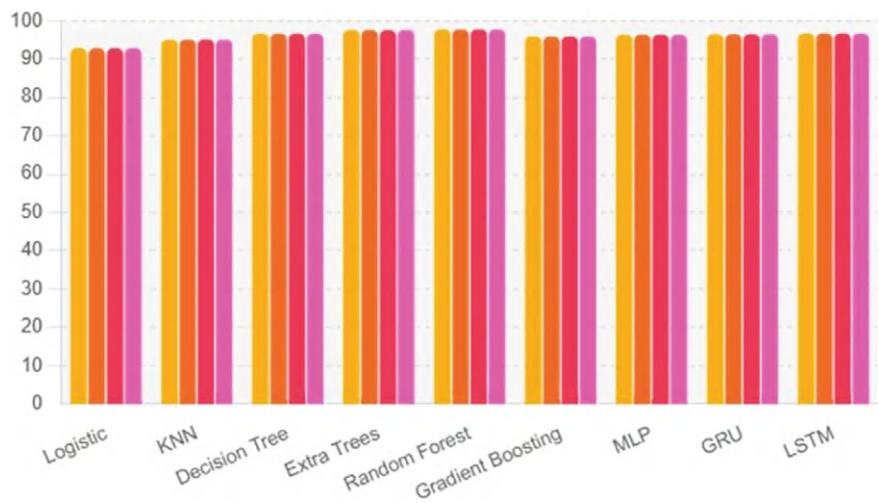


Fig. 1 Performance metrics comparison of different classifiers

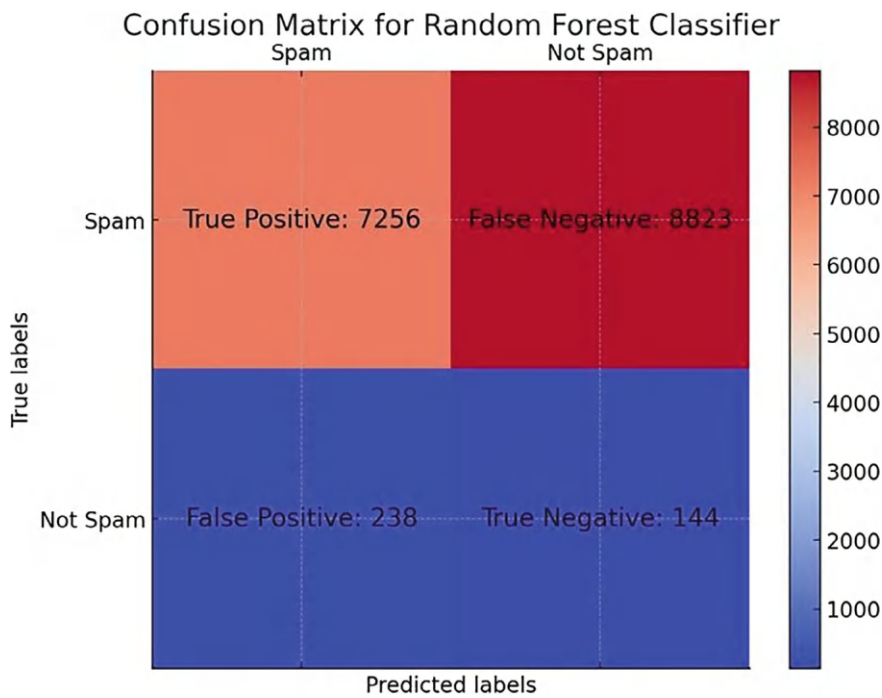


Fig. 2 Confusion matrix for random forest classifier

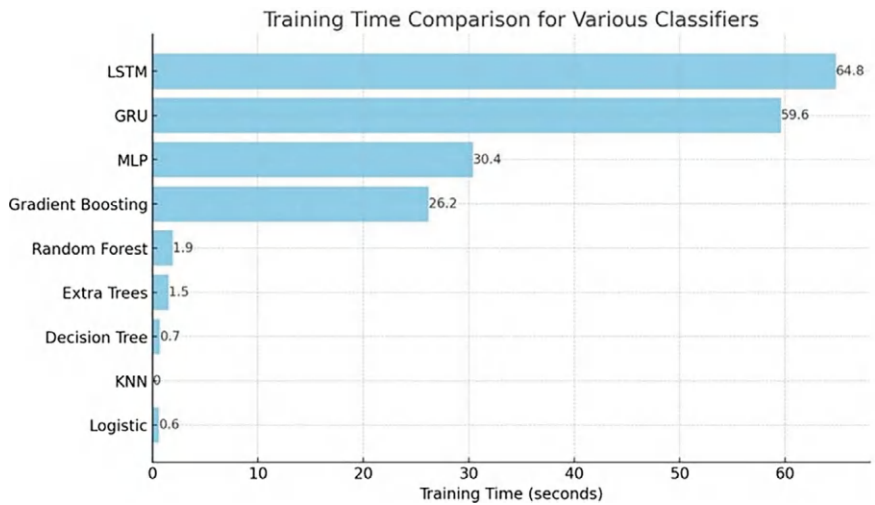


Fig. 3 Training time comparison for various classifiers

Figure 4 shows the global best fitness with particle swarm optimization (PSO), which monitors the optimization process to boost the classifier’s performance. The trend illustrates the role of PSO in tuning the models to a greater extent to achieve higher accuracy in spam detection.

An overview of the performance of the SVM classifier in classifying spam and nonspam emails is given by the confusion matrix shown in Fig. 5. It then discusses

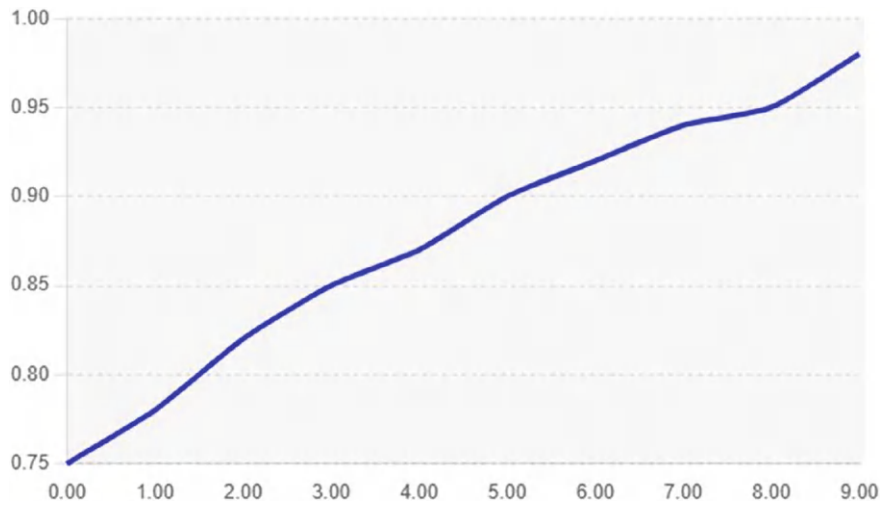


Fig. 4 Performance of PSO over epochs

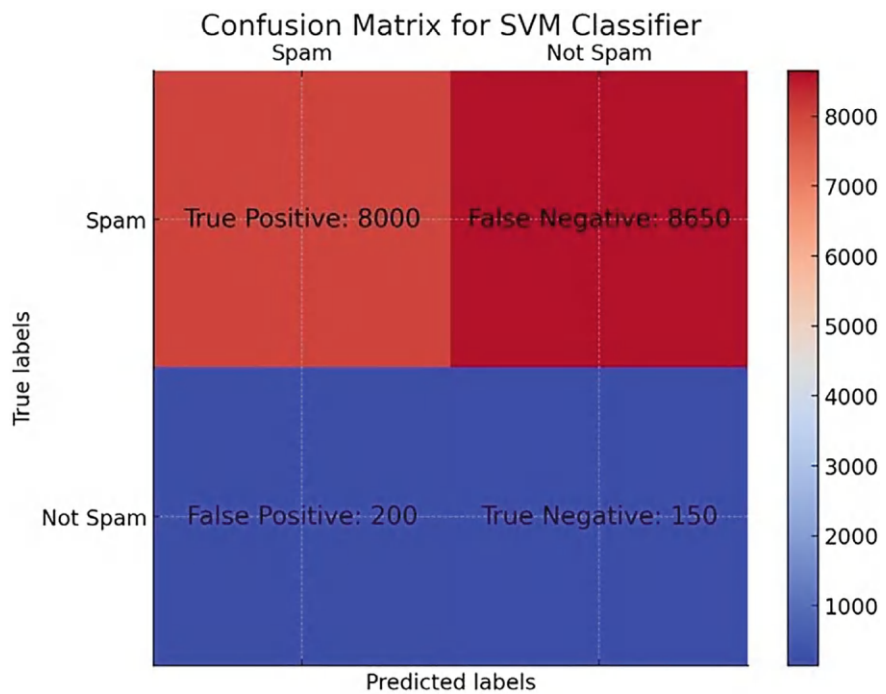


Fig. 5 Confusion matrix for SVM classifier

how many correct and incorrect classification numbers help evaluate and improve the modeling accuracy.

The Receiver Operating Characteristic (ROC) curve shown in Fig. 6 best compares the diagnostic ability of several different classifiers (Logistic Regression, Random Forest, SVM, etc.). AUC represents the effectiveness of discriminating classes for each model as a visual and quantitative measure of model performance.

These visualizations and tables support the paper’s narrative with empirical evidence of the capability and efficiency of machine learning techniques shown in Table 1 for threat identification and mitigation in anonymous email abuse. This evidence supports the paper’s thesis of agents who can effectively and efficiently mitigate the dangerous use of anonymous email abuse. Therefore, these insights are critical to improving strategic cyber security deployment.

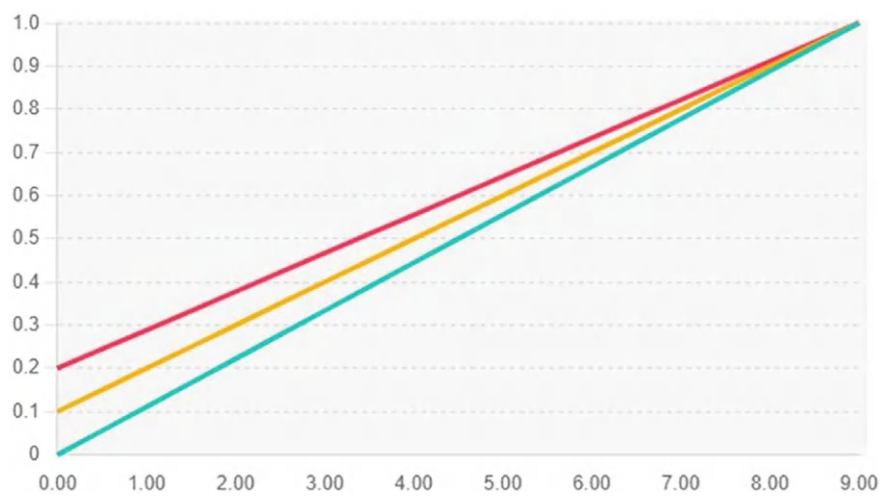


Fig. 6 ROC curve for different classifiers

Table 1 Performance metrics of classifiers for spam email detection

Classifier	Accuracy (%)	Recall (%)	Precision (%)	F1-score (%)	Training time (s)
Logistic	92.81	92.81	92.84	92.82	0.6
KNN	95.00	95.04	95.09	95.05	0.0
Decision tree	96.54	96.54	96.54	96.54	0.7
Extra trees	97.53	97.53	97.55	97.53	1.5
Random forest	97.68	97.68	97.69	97.68	1.9
Gradient Boosting	95.85	95.85	95.86	95.85	26.2
MLP (Keras)	96.30	96.30	96.30	96.30	30.4
GRU (Keras)	96.40	96.40	96.40	96.40	59.6
LSTM (Keras)	96.60	96.60	96.60	96.60	64.8

5 Conclusion

This investigation of strategic machine learning deployment to win the fight against email spam has provided several key insights and further contributions to the cybersecurity field. A comparative analysis of different machine learning classifiers was first done, finding significant differences in performance, with ensemble methods like Random Forest and even advanced neural networks outperforming them. Moreover,

these models demonstrated high accuracy and outperformed precision and recall, which were instrumental in bringing down false positives and providing reliable spam detection.

Particle Swarm Optimization was the tool that helped us refine the models, especially when tuning the hyperparameters that influence how quickly the models train and run once trained. The fact that models could use PSO to find the configuration that gave them the highest degree of accuracy demonstrated the ability of PSO to do so under the characteristics of the original dataset, which includes a wide variety of email content from multiple sources.

Additionally, this study uncovered the problems of deploying machine learning solutions in the real world: the amount of data preprocessing necessary and the computational requirements required to train even the simplest machine learning models. However, the research also showed that spam was a dynamic phenomenon, constantly evolving to stay ahead of the game against traditional detection techniques. That highlights the need to develop adaptive systems to reconfigure their parameters to counter new threats.

Finally, the facts of this research promote an educated defense against spam detection, recommending that models be improved in succession with further modeling and integration of machine learning into overall cybersecurity plans. Future work will explore the ability of unsupervised learning models and deep learning techniques to detect sophisticated spam tactics without a considerable amount of labeled data. Moreover, positioning these machine learning models in the context of network security, alongside anomaly detection systems, threat monitoring, and others, will enhance cyber defenses against the multiplicity of cyber threats.

Acknowledgements There is no funding for this work.

Disclosure of Interests The authors have no competing interests to declare relevant to this article's content.

References

1. Aditya, B.L.V.S., Mohanty, S.N.: Heterogenous social media analysis for efficient deep learning fake-profile identification. IEEE Access (2023)
2. Aditya, B.L.V.S., Mohanty, S.N.: System and method for social media fake profile identification. iN Patent 202341067941 A (2023)
3. Aditya, B.L., Mohanty, S.N.: Unveiling the underworld: Detecting fake profiles through social media network analysis and behavioral modeling. In: International Conference on Pervasive Knowledge and Collective Intelligence on Web and Social Media, pp. 342–352. Springer Nature, Switzerland, Cham (2023)
4. Aditya, B.L., Mohanty, S.N., Salini, Y.: Temporal sentiment analysis (TSMFPMSM) model for multimodal social media fake profile detection. In: International Conference on Pervasive Knowledge and Collective Intelligence on Web and Social Media, pp. 329–341. Springer Nature, Switzerland, Cham (2023)

5. Aditya, B.L., Rajaram, G., Hole, S.R., Mohanty, S.N.: F2PMSMD: design of a fusion model to identify fake profiles from multimodal social media datasets. In: International Conference on Intelligent Systems and Machine Learning, pp. 13–23. Springer (2022)
6. Aski, A.S., Sourati, N.K.: A proposed efficient algorithm to filter spam using machine learning techniques. *Pac. Sci. Rev. A Nat. Sci. Eng.* **18**(2), 145–149 (2016)
7. Bhuiyan, H., Ashiquzzaman, A., Juthi, T.I., Biswas, S., Ara, J.: A survey of existing e-mail spam filtering methods considering machine learning techniques. *Global J. Comp. Sci. Technol.* **18**(2), 20–29 (2018)
8. Blanzieri, E., Bryl, A.: E-Mail Spam Filtering with Local SVM Classifiers (2008)
9. Erden, C., Demir, H.I., K  k  am, A.H.: Enhancing machine learning model performance with hyperparameter optimization: a comparative study. [arXiv:2302.11406](https://arxiv.org/abs/2302.11406) (2023)
10. Ferrag, M.A., Maglaras, L., Moschoyiannis, S., Janicke, H.: Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study. *J. Inf. Secur. Appl.* **50**, 102419 (2020)
11. Fraley, J.B., Cannady, J.: The promise of machine learning in cybersecurity. In: SoutheastCon 2017, pp. 1–6. IEEE (2017)
12. Kumar, N., Sonowal, S., et al.: Email spam detection using machine learning algorithms. In: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 108–113. IEEE (2020)
13. Moustafa, N., Turnbull, B., Choo, K.K.R.: An ensemble intrusion detection technique based on proposed statistical flow features for protecting network traffic of Internet of Things. *IEEE Internet Things J.* **6**(3), 4815–4830 (2018)
14. Saleh, A.J., Karim, A., Shanmugam, B., Azam, S., Kannoorpatti, K., Jonkman, M., Boer, F.D.: An intelligent spam detection model based on the artificial immune system. *Information* **10**(6), 209 (2019)
15. Salman, T., Bhamare, D., Erbad, A., Jain, R., Samaka, M.: Machine learning for anomaly detection and categorization in multi-cloud environments. In: 2017 IEEE 4th international conference on cyber security and cloud computing (CSCloud), pp. 97–103. IEEE (2017)
16. Sangani, N.K., Zarger, H.: Machine learning in application security. In: Advances in Security in Computing and Communications. IntechOpen (2017)
17. Sarker, I.H., Kayes, A., Badsha, S., Alqahtani, H., Watters, P., Ng, A.: Cybersecurity data science: an overview from a machine learning perspective. *J. Big Data* **7**, 1–29 (2020)
18. Susilo, B., Sari, R.F.: Intrusion detection in IoT networks using a deep learning algorithm. *Information* **11**(5), 279 (2020)
19. Tait, K.A., Khan, J.S., Alqahtani, F., Shah, A.A., Khan, F.A., Rehman, M.U., Boulila, W., Ahmad, J.: Intrusion detection using machine learning techniques: an experimental comparison. In: 2021 International Congress of Advanced Technology and Engineering (ICOTEN), pp. 1–10. IEEE (2021)
20. Vyas, T., Prajapati, P., Gadhwal, S.: A survey and evaluation of supervised machine learning techniques for spam e-mail filtering. In: 2015 IEEE international conference on electrical, computer, and communication technologies (ICECCT), pp. 1–7. IEEE (2015)
21. Yang, B., Jiang, J., Li, N.: Towards discovering covert communication through email spam. In: Intelligent Information Processing VIII: 9th IFIP TC 12 International Conference, IIP 2016, Melbourne, VIC, Australia, November 18–21, 2016, Proceedings 9, pp. 191–201. Springer (2016)
22. Zhuang, L., Dunagan, J., Simon, D.R., Wang, H.J., Osipkov, I., Tygar, J.D.: Characterizing botnets from email spam records. *LEET* **8**(1), 1–9 (2008)

AI-Powered Multi-layered Phishing Defense Framework (AIPDF)



Pallavi Bhujbal, Jayashree Pasalkar, Madhura Eknath Sanap,
Bhagyashree Shendkar, Rajkumar Patil, and Moushmee Kuri

Abstract AI-powered multi-layered Phishing Defense Framework (AIPDF) is developed in response to the increasing complexity and sophistication of phishing attacks in the digital age. Traditional cybersecurity measures often fail to address these evolving threats, necessitating a more dynamic, intelligent, and multi-faceted approach. AIPDF leverages AI, machine learning, and blockchain technologies to provide comprehensive, real-time defense mechanisms against phishing, ensuring more accurate detection and faster response. The AI-powered multi-layered Phishing Defense Framework (AIPDF) is designed to counter evolving phishing threats by integrating multiple layers of defense. Each layer addresses different aspects of phishing attacks, from real-time anomaly detection to automated incident response. The framework combines AI-powered phishing detection engines, blockchain-based email authentication, and real-time threat intelligence to enhance email security.

P. Bhujbal · R. Patil

Department of Information Technology, MIT School of Computing, MIT Art Design and Technology University, Pune, India

e-mail: pallavi.bhujbal@mituniversity.edu.in

R. Patil

e-mail: rajkumar.patil@mituniversity.edu.in

J. Pasalkar

AISSM's, Institute of Information Technology, Pune, India

e-mail: jaysh26@gmail.com

M. E. Sanap

Department of Computer Science Software Engineering, Vishwakarma Institute of Technology, Pune, India

e-mail: madhura.sanap@vit.edu

B. Shendkar (✉) · M. Kuri

Department of Computer Science and Engineering, MIT School of Computing, MIT Art Design and Technology University, Pune, India

e-mail: bhagyashree.shendkar@mituniversity.edu.in

M. Kuri

e-mail: moushmee.kuri@mituniversity.edu.in

It also incorporates multi-factor authentication and user training to ensure continuous protection. This system enables rapid detection and response to phishing attacks while providing cybersecurity infrastructure transparency, adaptability, and resilience. Future iterations of the framework can integrate additional AI/ML models and extend to broader cybersecurity challenges.

Keywords Phishing detection · AI in cybersecurity · Blockchain email validation · Real-time threat detection · Multi-layered defense · Automated incident response

1 Introduction

Phishing attacks have become one of the most prevalent and damaging cyber threats, targeting individuals and organizations worldwide. Sensitive information, including login credentials, bank account information, and other personal details, are usually tricked into being revealed using phony emails or websites [1]. Phishing techniques have become more complex as hackers continue to change their strategies. They now use artificial intelligence, clever domain names, and innovative social engineering to create compelling messages [2]. More effective protection mechanisms are needed because the availability of phishing kits on the dark web has significantly reduced the barrier to entry for cybercriminals.

In growing cyber threats, single-layered security solutions have proven inadequate in providing comprehensive protection. Multi-layered defense systems are essential because they offer multiple layers of security, boosting the possibility that an attack will be intercepted and neutralized before it can do any damage [3, 4]. Multi-layered defenses can identify malicious activities in phishing cases at several stages, including email transmission, malware execution, and suspicious user behavior. This strategy can prevent or at least lessen damage even if one protective layer fails [5]. Multi-layered systems strengthen the overall security posture by reducing the possibility of false positives while simultaneously increasing detection accuracy.

As phishing attacks grow, conventional security measures struggle to keep pace. Static filters are often bypassed by more sophisticated, modern kinds of phishing, which conventional rule-based or heuristic approaches frequently struggle to identify [6]. Because of this, there is an urgent need for intelligent, adaptable systems that can change and grow in response to new dangers. The capacity to analyze massive amounts of data in real time, identify subtle trends, and produce insights that humans or simple algorithms could miss makes artificial intelligence (AI) a possible option [7, 8]. AI-enabled multi-layered defensive systems can provide enterprises with automated, real-time phishing detection and reaction. To provide a more effective and transparent defense against complex phishing threats, this research aims to create and assess an AI-powered phishing defense framework that blends machine learning, blockchain technology, and real-time threat intelligence.

2 Background and Literature Review

2.1 Evolution of Phishing Attacks and Techniques

Since the mid-1990s, when phishing attacks first appeared, they have evolved substantially. Phishing techniques from the past primarily consisted of spoof emails that led victims to fake websites intended to steal personal data [9]. On the other hand, phishing tactics have advanced significantly in sophistication. Attackers today use techniques like whaling, which goes after senior executives, and spear phishing, which sends tailored communications to particular people or organizations. Additional sophisticated methods include vishing (voice phishing), which uses phone calls to obtain information, and clone phishing, which replicates safe emails with harmful links. The popularity of “phishing-as-a-service” has made it simpler for adversaries to launch intricately planned campaigns [10]. Because of this, it is now harder to identify phishing using conventional protection mechanisms, calling for more sophisticated solutions like AI-powered systems.

2.2 Traditional Phishing Defense Mechanisms Strengths and Limitations

Blocklists, spam filters, two-factor authentication (2FA), and heuristic-based detection are examples of traditional phishing protection techniques. These methods have significant drawbacks even if they provide minimal security [11]. While heuristic approaches and spam filters can identify well-known phishing patterns, they are less successful in identifying new, zero-day phishing assaults that employ creative evasion techniques. Although blocklists are reactive and can only be updated after a phishing attempt has been detected, they can prevent harmful websites [12]. Even with the extra protection layer that 2FA provides, some highly skilled phishing attempts can get past it. These drawbacks emphasize the demand for more alert and sophisticated systems that can instantly recognize new phishing attacks.

2.3 Role of AI in Cyber Security

AI has completely changed a lot of cyber security, especially in identifying and thwarting online attacks. Artificial intelligence (AI) is well-suited to spotting patterns and abnormalities that can point to phishing assaults since it can quickly evaluate massive datasets [13]. A subset of artificial intelligence called machine learning algorithms can be trained on historical phishing data to identify phishing attempts based on various characteristics such as email content, sender reputation, and user behavior. AI systems can also continuously learn from new threats, allowing adaptive

protection mechanisms that improve with time [14, 15]. Artificial intelligence (AI) can potentially enhance phishing detection and response times by shortening the interval between detecting and stopping an attack.

Overview of related work: AI in phishing detection, blockchain in email authentication, threat intelligence systems.

Recent studies have explored the application of AI in phishing detection with promising results. Machine learning models like decision trees, support vector machines (SVM), and deep learning techniques such as convolutional neural networks (CNN) and recurrent neural networks (RNN) have been widely used for phishing detection [16]. These models determine whether an email is malicious by examining various email properties, such as content, links, and metadata. There has been an increasing interest in email authentication with distributed ledger technology regarding blockchain. By offering a safe, decentralized method of sender identity validation, blockchain can lessen the possibility of phishing attempts [17]. Furthermore, the usage of threat intelligence systems—which incorporate real-time inputs from several cybersecurity sources—to proactively identify and counteract phishing threats is growing. These technologies provide a more dynamic and well-coordinated approach to cybersecurity defense by gathering, analyzing, and sharing data about new threats.

3 Proposed Methodology

The AI-powered multi-layered Phishing Defense Framework (AIPDF) depicted in Fig. 1 is designed to provide comprehensive protection against phishing attacks through advanced detection techniques and user-centric security measures. At its core, the Incident Detection and Response Layer identifies anomalies in real-time. Once a phishing attempt is detected, this layer isolates the email and notifies the user while initiating automated incident response protocols to quickly and efficiently contain the threat.

Next, the Email Filtering Layer thoroughly analyses all incoming emails. The AI-powered phishing detection engine within this layer checks the authenticity of senders and scrutinizes email content, including attachments and links. By doing so, it prevents malicious emails from reaching the user. Complementing this, the Attachment and Malware Scanning Layer focuses on detecting harmful attachments and malware. It uses AI to identify potential malware and zero-day exploits that might otherwise bypass traditional security systems.

The Multi-Factor Authentication (MFA) and Dynamic Access Layer employ context-aware access control and behavioral biometrics to secure user accounts further, ensuring that only authorized users can access sensitive information. This layer adapts based on the context of the user's actions, making unauthorized access significantly more difficult.

The User Layer plays a crucial role by continuously training users to recognize phishing threats. Through behavioral biometrics and a User Awareness and Training

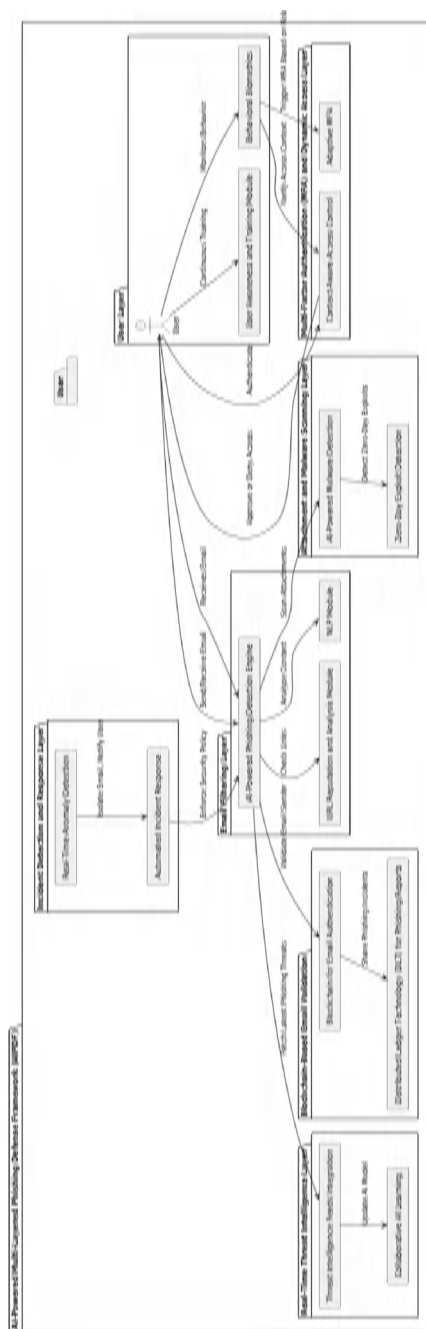


Fig. 1 AI-powered multi-layered phishing defense framework (AIPDF)

Module, this layer helps enhance the user's cyber security skills while monitoring their behavior to detect potential breaches.

A unique aspect of this framework is the Blockchain-Based Email Validation Layer, which uses blockchain technology for email authentication. Distributed Ledger Technology (DLT) facilitates sharing phishing incidents and email validation across the network, bolstering the system's overall robustness.

Finally, the Real-Time Threat Intelligence Layer integrates the latest threat intelligence feeds and updates the AI models used across the framework. It ensures continuous learning and adaptation to new phishing threats while promoting collaborative AI learning, allowing the framework to evolve and remain effective against emerging attack vectors. This multi-layered approach enhances proactive and reactive phishing defenses, ensuring comprehensive email security.

4 Summary and Discussions

4.1 Summary

The AI-powered multi-layered Phishing Defense Framework (AIPDF) integrates several defense layers, each focusing on a distinct component of email communication and cybersecurity threats, to offer complete protection against phishing assaults. Real-time anomaly detection is built into the Incident Detection and Response Layer, which isolates potentially harmful emails before they reach the user [18]. The system ensures that threats are rapidly neutralized by automatic incident response, reducing the potential for damage. Proactive defense is made possible by sophisticated anomaly detection algorithms and isolation methods.

The Email Filtering Layer scans and analyzes emails for phishing indicators using an AI-powered phishing detection engine. Emails are evaluated for content and sender validity using URL reputation analysis and Natural Language Processing (NLP). By utilizing several preventative techniques, the system successfully screens out harmful emails before they are seen by the user [19]. The Attachment and Malware Scanning Layer, which focuses on finding malware in email attachments, complements this [20]. AI-based methods find known and unidentified dangers, such as zero-day exploits. This layer thoroughly covers all possible attachment-based dangers by integrating scanning engines with the email system.

The framework in the Blockchain-Based Email Validation Layer also utilizes blockchain technology. By using Distributed Ledger Technology (DLT) to document and exchange phishing instances, blockchain establishes a decentralized system for email authentication, guaranteeing the legitimacy of email senders [21]. This method assists in identifying and thwarting phishing efforts in addition to authenticating valid senders. In addition, the Real-Time Threat Intelligence Layer incorporates threat intelligence streams to prevent new phishing techniques. The framework's AI models

are updated regularly in response to fresh threat intelligence, and collaborative AI learning makes sure the system changes to meet new security risks.

The Multi-Factor Authentication (MFA) and Dynamic Access Control Layer use behavioral biometrics for continuous authentication and context-aware access control methods to bolster access security. Adaptive MFA approaches ensure only authorized users can access sensitive data by dynamically adjusting security based on user behavior [22, 23]. Ultimately, reducing the likelihood of phishing attacks is greatly aided by the User Awareness and Training Layer. Users are informed about phishing dangers through ongoing training and awareness initiatives, and behavioral monitoring assists in identifying any strange activity. This layer allows users to identify and react to phishing efforts, improving the overall cybersecurity posture.

4.2 Discussions

Ensuring real-time protection against phishing requires the Incident Detection and Response Layer. The technology can quickly identify anomalous patterns in email correspondence—which can point to a phishing attempt—thanks to the application of anomaly detection algorithms. By identifying questionable emails and triggering an automated reaction, the technology shortens the time hackers might take advantage of users [24]. By taking a proactive stance, threats are dealt with before they can cause damage, enhancing the system's overall security. The automated response and real-time detection integration provide a solid base for the multi-layered security system.

Strong defense against email-based threats is offered by the combination of the Attachment and Malware Scanning Layer and the Email Filtering Layer. AI-powered phishing detection uses sophisticated content analysis tools, like natural language processing (NLP) and URL reputation analysis, to improve the system's capacity to distinguish between fraudulent and legitimate emails [25–27]. The system ensures that harmful payloads are prevented from reaching consumers, even if they evade the initial phishing detection, by checking email attachments for malware and identifying zero-day exploits. Attackers find it challenging to use email as a gateway for distributing malware or phishing scams because of the combined resilience of these two levels.

Using blockchain technology to verify email senders' legitimacy, the Blockchain-Based Email Validation Layer offers a fresh strategy against phishing. Because blockchain is decentralized and unchangeable, it is the perfect technology for preserving the integrity of email correspondence. The technology makes it more difficult for phishing attempts to succeed by recording hostile activity and certifying senders by storing phishing occurrences on a distributed ledger [28]. This method is transparent and dependable. Thanks to this innovation, email authentication procedures now have far greater security and trustworthiness.

The AI models in the framework are kept up to speed with the most recent phishing techniques thanks to the Real-Time Threat Intelligence Layer. The system can react to

new threats as they materialize, thanks to the constant integration of threat intelligence streams. Furthermore, the framework can change and adapt based on shared data from many sources thanks to collaborative AI learning [29]. This layer guarantees the defense system is resilient against novel attack techniques while enhancing phishing detection accuracy.

Finally, by guaranteeing that system access is highly secure, the Multi-Factor Authentication (MFA) and Dynamic Access Control Layer add another line of defense. Adaptive MFA and behavioral biometrics combine to offer a flexible and reliable authentication solution that modifies security protocols in response to user activity. By enabling users to participate actively in the phishing defense, the User Awareness and Training Layer enhances these technical defenses. Using ongoing training and oversight, users improve their ability to identify phishing attempts and react suitably, diminishing the probability of successful assaults.

Together, these layers create a holistic defense framework that addresses the complexities of modern phishing attacks, offering a strong, adaptable, and proactive cybersecurity solution.

5 Conclusion

The AI-powered multi-layered Phishing Defense Framework (AIPDF) offers a comprehensive and adaptive approach to phishing detection, combining several layers of defense that integrate AI, blockchain, and threat intelligence. Key contributions include the real-time detection of phishing attempts, validation of email authenticity via blockchain, and proactive incident response mechanisms. The framework emphasizes continuous user training and behavioral monitoring, ensuring that cybersecurity's technical and human aspects are addressed. The findings suggest that by leveraging advanced anomaly detection algorithms, collaborative AI learning, and multi-factor authentication, AIPDF significantly enhances real-time threat detection and response. Its layered approach reduces the risk of phishing attacks and ensures rapid incident mitigation, contributing to improved cybersecurity resilience. For future research, extending the framework's integration with more sophisticated AI/ML models could enhance phishing detection accuracy. Additionally, broadening its application to other cyber threats such as ransomware, social engineering, and data breaches may make it a more versatile and powerful tool for cybersecurity in diverse environments.

References

1. Ramos-Cruz, B., Andreu-Perez, J., Martínez, L.: The cybersecurity mesh: a comprehensive survey of involved artificial intelligence methods, cryptographic protocols and challenges for

- future research. *Neurocomputing* **581**, 127427 (2024). <https://doi.org/10.1016/j.neucom.2024.127427>
2. Tonhauser, M., Ristvej, J.: Cybersecurity automation in countering cyberattacks. *Transp. Res. Procedia* **74**, 1360–1365 (2023). <https://doi.org/10.1016/j.trpro.2023.11.283>
 3. Kaur, R., Gabrijelčič, D., Klobučar, T.: Artificial intelligence for cybersecurity: literature review and future research directions. *Inf. Fusion* **97** (2023). <https://doi.org/10.1016/j.inffus.2023.101804>
 4. Chandre, P.R., Shendkar, B.D., Deshmukh, S., Kakade, S., Potdukhe, S.: Machine learning-enhanced advancements in quantum cryptography: a comprehensive review and future prospects. *Int. J. Recent Innov. Trends Comput. Commun.* **11**(11s), 642–655 (2023). <https://doi.org/10.17762/ijritcc.v11i11s.8300>
 5. Marques, C., Malta, S., Magalhães, J.P.: DNS dataset for malicious domains detection. *Data Br.* **38**, 107342 (2021). <https://doi.org/10.1016/j.dib.2021.107342>
 6. Promyslov, G., Semenov, K.V., Shumov, A.S.: A clustering method of asset cybersecurity classification. *IFAC-PapersOnLine* **52**(13), 928–933 (2019). <https://doi.org/10.1016/j.ifacol.2019.11.313>
 7. Schmitt, M., Flechais, I.: Digital deception: generative artificial intelligence in social engineering and phishing. *SSRN Electron. J. ML*, 1–18 (2023). <https://doi.org/10.2139/ssrn.4602790>
 8. Makubhai, S.S., Pathak, G.R., Chandre, P.R.: Predicting lung cancer risk using explainable artificial intelligence. *Bull. Electr. Eng. Inform.* **13**(2), 1276–1285 (2024). <https://doi.org/10.11591/eei.v13i2.6280>
 9. Ogundairo, O.: AI-driven phishing detection systems, August 2024 (2024)
 10. Alsubaie, S., Almazroi, A.A., Ayub, N.: Enhancing phishing detection: a novel hybrid deep learning framework for cybercrime forensics. *IEEE Access* **12**, 8373–8389 (2024). <https://doi.org/10.1109/ACCESS.2024.3351946>
 11. Alkhalil, Z., Hewage, C., Nawaf, L., Khan, I.: Phishing attacks: a recent comprehensive study and a new anatomy. *Front. Comput. Sci.* **3**, 1–23 (2021). <https://doi.org/10.3389/fcomp.2021.563060>
 12. Loh, P.K.K., Lee, A.Z.Y., Balachandran, V.: Towards a hybrid security framework for phishing awareness education and defense. *Future Internet* **16**(3), 86 (2024). <https://doi.org/10.3390/fi16030086>
 13. Mittal, A., Sivaraman, R.: Phishing detection using natural language processing and machine learning phishing detection using natural language processing and machine learning. *SMU Data Sci. Rev.* **6**(2), 14 (2022)
 14. Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., Shukla, S.: Applications of deep learning for phishing detection: a systematic literature review **64**(6) (2022). Springer London
 15. Gaddekar, B.P., Hiwarkar, T.: Tryambak Hiwarkar, “A Conceptual Modeling Framework to Measure the Effectiveness using ML in Business Analytics.” *Int. J. Adv. Res. Sci. Commun. Technol.* **2**(1), 399–406 (2022). <https://doi.org/10.48175/ijarsct-7703>
 16. Basit, A., Zafar, M., Liu, X., Javed, A.R., Jalil, Z., Kifayat, K.: A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommun. Syst.* **76**(1), 139–154 (2021). <https://doi.org/10.1007/s11235-020-00733-2>
 17. Rendall, K., Nisioti, A., Mylonas, A.: Towards a multi-layered phishing detection. *Sensors (Switzerland)* **20**(16), 1–18 (2020). <https://doi.org/10.3390/s20164540>
 18. Shad, R., Brooklyn, P., Egon, A.: The evolving thread landscape pf AI-powered cyberattacks: a multi-faceted approach to defense and mitigate (2023)
 19. Guebe, B., Azeta, A., Misra, S., Osamor, V.C., Fernandez-Sanz, L., Pospelova, V.: The emerging threat of AI-driven cyber attacks: a review **36**(1) (2022). Taylor & Francis
 20. Gaddekar, B., Hiwarkar, T.: A critical evaluation of business improvement through machine learning: challenges, opportunities, and best practices. *Int. J. Recent Innov. Trends Comput. Commun.* **11**(10s), 264–276 (2023). <https://doi.org/10.17762/ijritcc.v11i10s.7627>
 21. Gautam, H., Kumar, V., Sharma, V.: Phishing prevention techniques: past, present and future, pp. 83–98 (2021). https://doi.org/10.1007/978-981-33-6307-6_10

22. Chawla, M., Singh Chouhan, S.: A survey of phishing attack techniques. *Int. J. Comput. Appl.* **93**(3), 32–35 (2014). <https://doi.org/10.5120/16197-5460>
23. Damre, S.S., Shendkar, B.D., Kulkarni, N., Chandre, P.R., Deshmukh, S.: Smart healthcare wearable device for early disease detection using machine learning. *Int. J. Intell. Syst. Appl. Eng.* **12**(4s), 158–166 (2024)
24. Sadiq, et al.: A review of phishing attacks and countermeasures for the Internet of things-based smart business applications in industry 4.0. *Hum. Behav. Emerg. Technol.* **3**(5), 854–864 (2021). <https://doi.org/10.1002/hbe2.301>
25. Shankar, Shetty, R., Nath, B.: A review on phishing attacks. *Int. J. Appl. Eng. Res.* **14**(9), 2171–2175 (2019). <http://www.ripublication.com>
26. Chatterjee, S., Pattnaik, L.M., Satpathy, S.: An intrusion detection system and attack intension used in network forensic exploration. In: *International Conference on Intelligent Systems and Machine Learning*, December 2022, pp. 334–345. Springer Nature Switzerland, Cham (2022)
27. Chatterjee, S., Satpathy, S., Paikaray, B.K.: Forecasting DDoS attack with machine learning for network forensic investigation. *Int. J. Reason.-Based Intell. Syst.* **16**(5), 352–359 (2024); Priestman, W., Anstis, T., Sebire, I.G., Sridharan, S., Sebire, N.J.: Phishing in healthcare organizations: threats, mitigation and approaches. *BMJ Heal. Care Inform.* **26**(1), 1–6 (2019). <https://doi.org/10.1136/bmjhci-2019-100031>
28. Dhotre, D., Chandre, P.R., Khandare, A., Patil, M., Gawande, G.S.: The rise of crypto malware: leveraging machine learning techniques to understand the evolution, impact, and detection of cryptocurrency-related threats. *Int. J. Recent Innov. Trends Comput. Commun.* **11**(7), 215–222 (2023). <https://doi.org/10.17762/ijritcc.v11i7.7848>
29. Abbas, S.G., et al.: Identifying and mitigating phishing attack threats in IoT use cases using a threat modeling approach. *Sensors* **21**(14), 1–25 (2021). <https://doi.org/10.3390/s21144816>

Blockchain Innovations for Cyber Security and Digital Trust

Ethereum Blockchain-Based Decentralized Voting Platform



Natasha Wanjari, Pratiksha Chafle, and Rahul Moriwal

Abstract At the same time, electronic voting has emerged as an alternative voting method that could reduce voter turnout and inequality. Traditional international elections pose a challenge to the importance of stability and transparency. Elections are still centralized and controlled by an organization. Some issues that can arise in the traditional election process include managing and operating the organization's storage space and systems. This article examines historical voting systems used by some countries and organizations. Blockchain is perhaps the most unique technology today that promises to expand the power of electronic voting. The technology offers the opportunity to support blockchain decisions such as cryptographic perspective and transparency of principles in electronic voting. The course is designed to follow the basic rules of the electronic voting process and complete the final analysis. The system uses end-to-end electronic voting scores for higher education after in-depth evaluation.

Keywords Ethereum · Voting · Decentralized · Blockchain

1 Introduction

Whether it is traditional elections or electronic voting (e-voting), voting is the foundation of democracy today. People Electronic voting is seen as a solution that will attract young voters. Many features and security must be specified to achieve strong electronic voting, including transparency, accuracy, verification, performance and

N. Wanjari · P. Chafle (✉) · R. Moriwal

Department of Computer Science and Engineering, G H Raisoni College of Engineering Nagpur, Nagpur, India

e-mail: pratiksha.chafle@raisoni.net

N. Wanjari

e-mail: natasha.wanjari.trs@ghrce.raisoni.net

R. Moriwal

e-mail: rahul.moriwal@raisoni.net

data integrity fairness, confidentiality/stealth, legality, and distribution. Blockchain technology is supported by the collaboration of many people at the intersection [1]. In traditional voting, we use different voting methods such as EVM and ballot paper, where we use machines to control the voting process, the voter receives a ballot paper (paper), and the polling station releases the candidates. However, traditional elections have many disadvantages, such as paper processing, being time-consuming, having no direct leadership accountability, and machine damage. Bulk updates do not allow users to update multiple projects simultaneously [2]. One way to solve the security problem is to use blockchain technology. Blockchain technology has many applications [3].

Blockchain can secure transactions using cryptographic algorithms that protect against manipulation and fraud. Many researchers have proposed voting via blockchain to increase the accuracy and transparency of voting [4–6]. Question: The report also discusses the current status of some blockchain projects. Now, it is designed for a small selection on-site in applications, offices, meeting rooms, etc. Statistically, I am reasonable, honest, ambitious, successful, and strong. It also provides vote verification, secret ballot, vote, and public evidence. Blockchain technology is used to accomplish this task.

1.1 System of Paper Voting

Voting is the most optional. It will be used until electronic voting is done. The ballot paper consists of a ballot paper and a voting card. The ballot papers are available to all voters but cannot be seen. The disadvantages of this system are time and speed.

1.2 Online Voting System

The new platform for safe voting and voting is the online voting system. Online voting is a website where votes are submitted online through a web browser. Voters worldwide have the right to vote online. Calls have been made from central offices for online voting. The application offers an alternative to the long, understandable, secure, and easy process for voters. Voters can easily vote for local candidates without going to the polls and save time. This application divides the following contents into three groups according to users.

Admin Panel: This category is used only by the members of the Election Commission to manage the entire election process, including the registration of candidates and voters.

Voter Panel: This panel is only available to anyone with the right to vote (i.e., people who are 18 years old or older). These are the primary users of the application development.

The new platform for safe voting and voting is the online voting system. Online voting is a website where votes are submitted online through a web browser. Voters worldwide have the right to vote online. Central offices have made calls for online voting. The application offers an alternative to the long, understandable, secure, and easy process for voters. Voters can easily vote for local candidates without going to the polls and save time. This application divides the following contents into three groups according to users.

2 Literature Review

The primary purpose of our project is to provide a secure voting platform and prove that electronic solutions can be implemented using blockchain. When everyone with a computer or mobile phone can vote, citizens and members can make all operational decisions, or at least people will have open minds and policies because landowners and leaders will have more fun. This will eventually bring people to justice [7]. The process [8] is similar to an election, and the electronic voting process is the same as the previous voting. So this article will review blockchain technology and how it can be used for electronic voting. Each vote will be counted as one item. These votes will be counted, and the results will be announced immediately. Secret or secret ballots play a significant role in many countries. This process can affect the vote, undervote, etc., and has many disadvantages. We have started to investigate further to overcome this problem [9]. He [10] stated that electronic voting should provide security by being transparent (privacy is important) and not allowing re-election. He suggests using an innovative contract-based electronic voting application that allows users with a valid EOA to vote on the contract (once per address). However, this decision does not include an accurate address verification process because voters receive voting rights from the central government. He said electronic elections should provide security by being transparent (confidential) and not allowing re-election. He suggests using an innovative contract-based electronic voting application that allows users with a valid EOA to vote on the contract (once per address). However, this solution does not have an actual address verification process because voters based on EOs receive voting rights from the central government. In this paper [11], we implemented and tested an intelligent contract-based electronic voting on the Ethereum network using blockchain technology and the Solidity wallet.

A study by [12] shows that weak electronic voting has the potential to affect voter privacy and integrity, leading to false votes being entered into the system or votes being miscounted. In a private [13] blockchain, only the entity that owns the blockchain can grant anyone the authority to use the blockchain and vote. In this model, the government owns the election process and, therefore, is responsible for allowing its citizens to vote using blockchain technology. The government is also the only entity that puts voters on the blockchain. On private [14] blockchains, creating a block and changing the nonce (mining process) before obtaining a signature should be cost-effective on public blockchains, as fewer nodes exist. The rise of digital voting

increases security, efficiency, and fairness. Designed to prevent fraud by limiting the speed of new ballots being processed on electronic machines, the device has been used in many countries where elections are held.

In addition to predicting future developments, this literature review [15] discusses the challenges and solutions to scalable blockchain-based electronic voting systems. They analyze the proposals from previous studies, their implementation, analysis, and various cryptographic solutions to evaluate the cost and time. They analyze the performance, advantages, and limitations of different systems and the most common methods for blockchain scalability. Singh et al. [16] blockchain technology is one way to solve the problems that often occur in elections. Its benefits also make the body more secure. Electronic voting has been controversial in the United States since 2001, when the government announced the process in the summer of 2003. Considering these issues, integrating blockchain technology into electronic voting has become a good way to go [17–19]. Today, the development of technology has improved the lives of many people. The paper has many more uses than today's ballot papers. The potential of legacy systems continues to grow and threatens security and transparency [20]. Electronic voting should make receiving and counting ballots in elections easy, convenient, and secure. Advances in mobile, wireless, and network technologies have led to the emergence of new applications that will make the voting process easier and more efficient [21]. Blockchain-based electronic voting systems provide instant review and verification of the voting process. The transparency of Blockchain allows independent auditors and stakeholders to monitor voting activities and verify the accuracy of the vote count. This increases the electoral process's transparency, accountability, and trust [22]. By using smart contracts and encryption technology, the system ensures the integrity and confidentiality of the vote while completing public verification. Most citizens do not comply with this restriction and avoid their obligations. In this case, electronic voting is generally considered a good option. Blockchain technology is a new technology that can provide immutable, transparent, anonymous, distributed, and real solutions [23]. According to the Times of India, on January 24, 2009, 1.1 million fake votes were found in Delhi. Later, according to a report by India News, in June 2013, The Election Commission detected 30,000 illegal voters at the Hilla Dikshit polling station. Another source, claimed by LJP (LJP) president Ram Vilas Paswan, said there were 30 voters in Bihar [24]. The system is designed to ensure secure voting, save money, reduce waiting time, eliminate discrepancies due to various errors, increase efficiency, and ensure work without physical exertion. Therefore, reliable elections will help in the development of democratic institutions. Thanks to our initiative, voters can now vote from the comfort of their homes, saving time and reducing voter errors [25]. In democratic countries, election security is a matter of national security. For ten years, computer security has been working on electronic voting to reduce the cost of elections nationwide. Since the beginning of democratic elections, voting has been done by paper and pencil. Replacing traditional pen and paper with new options is important to prevent fraud and make the voting process detectable and verifiable [26]. Create an electronic voting system that meets the needs of legislators. Our current elections are conducted using EVMs, which have been proven to be hackable and tamper-proof in

many places. This creates doubts about the election among candidates and the public. Hyperledger developers assumed that commercial blockchains would operate on a single trust system [27]. At this stage, technology is essential to help meet people's needs. Considering that today's majority do not trust politicians and elections are important in today's democracies, the increasing use of technology poses new challenges to democracy. Elections are important in deciding who will govern a country or institution, or we can say that elections are events that determine the future of a country [28]. Election security is an important issue during elections in a country. Computer security has been working on many electronic voting systems for years to increase security and reduce energy costs. Elections in India are conducted by voting in front of EVMs. EVMs have replaced India's local, state, and parliamentary elections [29]. With the popularity of blockchain technology, electronic voting systems increasingly use blockchain technology as a central storage to make the voting process more transparent, efficient, and secure to prevent tampering with data [30]. This study is divided into two parts: a control group and a voting group. The main purpose of voting (for citizens of a country) is to find the leader of their choice [31]. The application of blockchain technology has attracted great attention as a secure and open online election. To ensure the efficiency and reliability of the voting process, this paper focuses on the design and implementation of online voting systems based on blockchain technology [32]. When a hundred guides are matched, voters can choose their favorite candidates from the group. People can share a hyperlink to vote (as long as they know the link), and people who know the link can vote, and only one vote can be cast per browser. Weakness in voting, re-voting, and rejection of votes [33]. This has caused doubts about the election among candidates and the public. This article aims to analyze the use of blockchain technology to create a decentralized electronic voting system [34, 35].

3 Proposed Methodology

During this process, the administrator logs in to the system from the imported account. You can use the security key and password to import your account or the file. The two parts of the job are money and content. Administrators can add candidates through the Add Candidate function defined in the smart contract on the left side. When you click it, a box will open with details such as the candidate's name and the political party they represent. By calling this function, an instance of the model is created to store the object as a variable. Data is stored using the data provided in the user interface that feeds the data to the converter. It makes a smart contract, uses a template to execute it, and stores it as a brilliant contract instance. NOTE: Create a new template for each new candidate. This latest challenge is the need to exchange contracts on the blockchain. This is not the driver of the mining price; it is only for miners to verify the blocks. After the manager and the candidate sign the transaction, the information is entered into the system via a smart contract. This smart contract creates a smart contract based on the competitor's product, where all the information

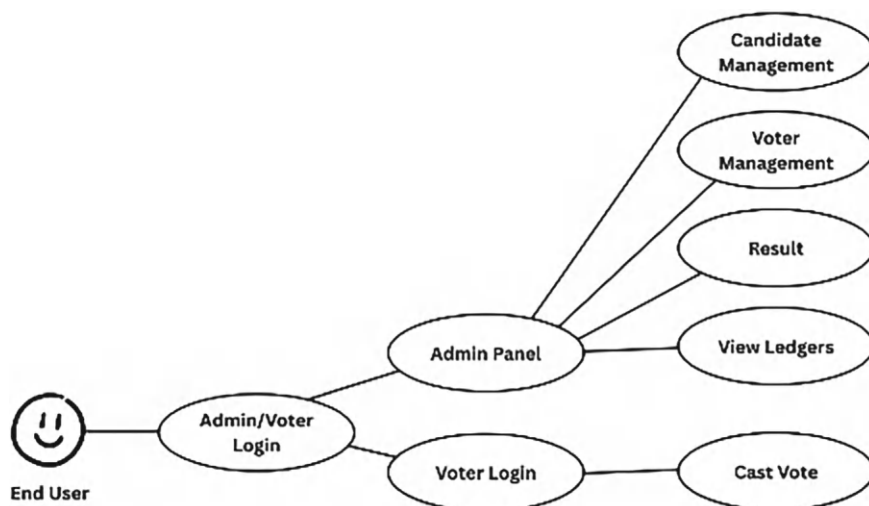


Fig. 1 Flow chart of voting system

about the competitor's product is stored. Voters first log in to the system. They can do this by creating a new account or registering a new one in the system. When voting, they sign the transaction by verifying it using a signature code or token. Once a transaction is signed, it is broadcast to the network and mined using methods such as Boot. Since most votes are public, the vote data is not encrypted, but we encrypt it for security purposes, as shown in Fig. 1.

Voting registration is done using Html/CSS/Bootstrap on the front end and SQL on the back end, and it stores the user's personal information, e.g., this can be considered an Aadhar database. Biometric devices will be used as evidence. The voter will provide their ID/address information as an access certificate if the user is valid. Administrators enter the system by sending money. Funds can be sent using key encryption. Both stores are linked to accounts and their associated content. Administrators can use the "Add Candidate" function to add new candidates. When clicked, a card with details such as the candidate's name and the political party they represent will appear. Calling this function creates an instance of the template sentence that stores the input as a variable. Data is stored using app.js. This will contrast with the data provided in the UI. App.js creates smart contracts and uses objects to implement and store them as smart contracts. Note: 17 new events are made for each new contestant. This new feature is a contract on the blockchain called transactions. It works in a monitoring mode only and does not change the blockchain value's state except for the mining value of the leader miner who adds the block to the blockchain for external verification. The whole process is shown in Figs. 2, 3, 4, 5, 6.

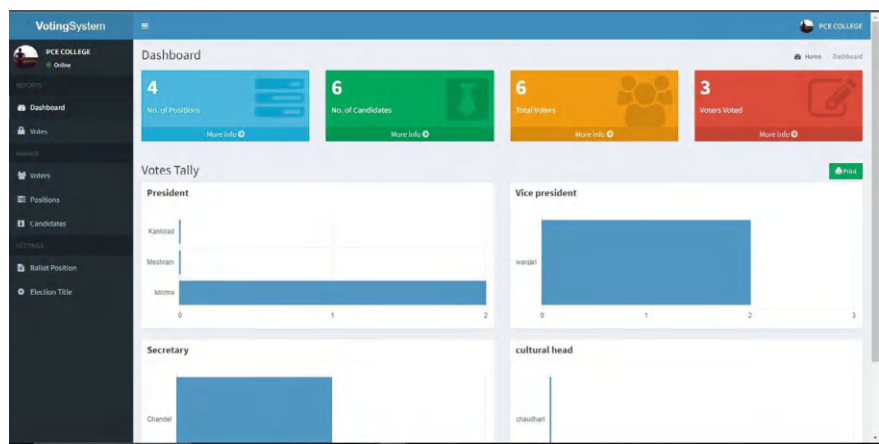


Fig. 2 This is the login page of the admins

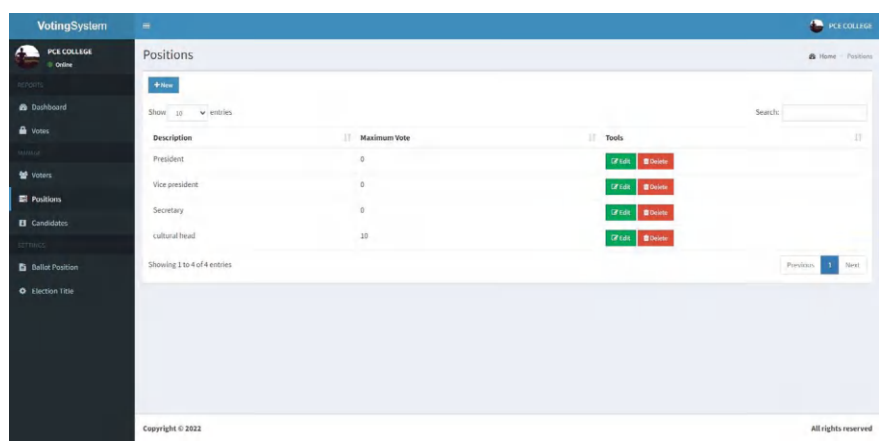


Fig. 3 Admin dashboard

4 Result and Discussion

This Is the Voter Dashboard Or Voter Page; in This Voter, Can Only Vote Once The Voter Not Voted looks like in Fig. 7. Once they vote, they will not vote again, and the page will look like this: they can only see who they voted for but can't edit or cancel their vote.

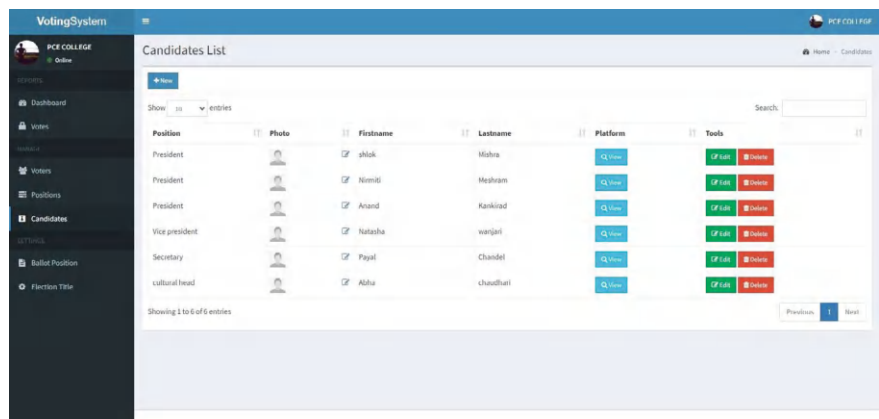


Fig. 4 Add voter details

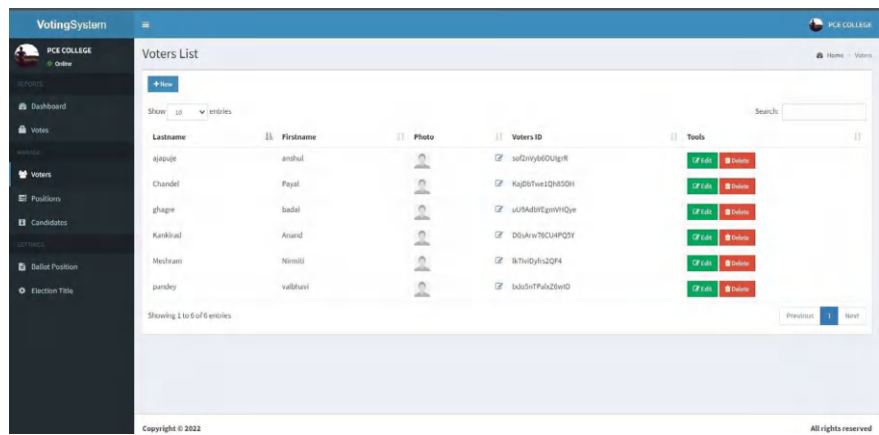


Fig. 5 Position details

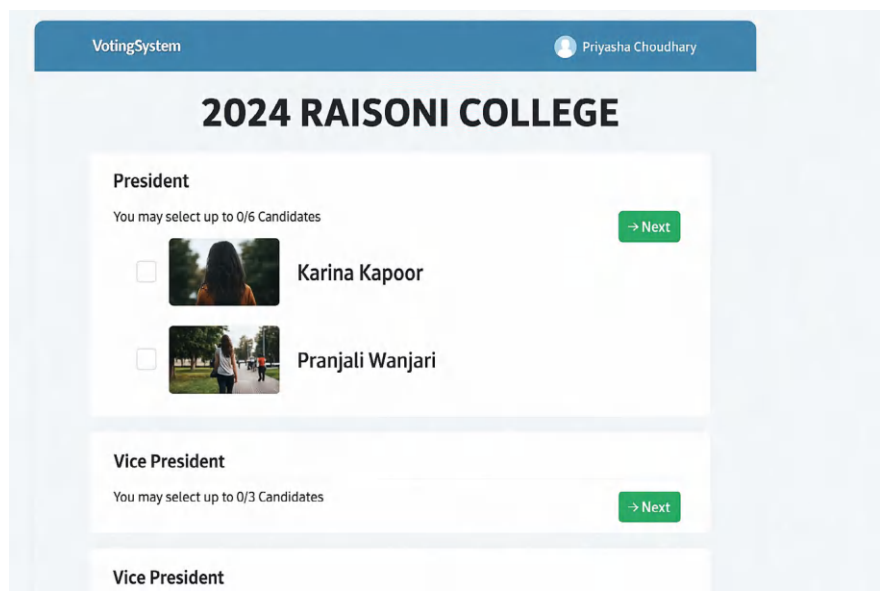


Fig. 6 Add candidates and all candidates list

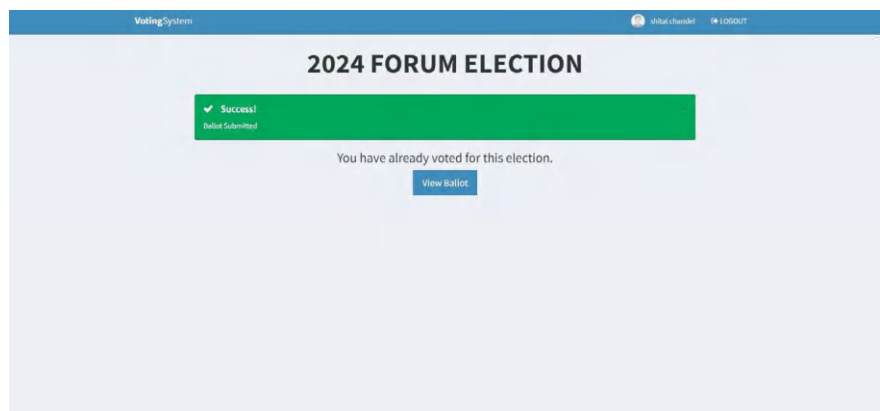


Fig. 7 Successfully voted

5 Conclusion

Companies in organizations such as presidential, parliamentary, and various office elections currently use the system. Additional fees for electronic voting will be determined according to the decision’s requirements. Blockchain-based electronic voting is designed to meet the requirements of the electronic voting process. All

votes in the blockchain are cryptographically linked, block by block. If other blocks have the same duration, the block with the highest signature value is selected.

References

1. Kuma, S., Singhi, N., Patankar, A.: A survey on the smart electronic voting system through blockchain technology. *J. Emerg. Technol. Innov. Res. (JETIR)* (2020)
2. Sah, A.K., Gupta, S., Patel, N., Harshitha, P., Basavaraju, D.R.: Effective e-voting mechanism using blockchain and IOT. In: 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, pp. 2277–2280, March 2024. IEEE (2024). <https://doi.org/10.1109/ICACCS60874.2024.10717141>
3. Ayed, A.B.: A conceptual secure blockchain-based electronic voting system. *Int. J. Netw. Secur. & Appl.* **9**(3), 01–09 (2017). <https://doi.org/10.1109/WorldS4.2018.8611593>
4. Pereira, B.M.B., Torres, J.M., Sobral, P.M., Moreira, R.S., Soares, C.P.D.A., Pereira, I.: Blockchain-based electronic voting: a secure and transparent solution. *Cryptography* **7**(2), 27 (2023). <https://doi.org/10.3390/cryptography7020027>
5. Garg, K., Saraswat, P., Bisht, S., Aggarwal, S.K., Kothari, S.K., Gupta, S.: A comparative analysis of e-voting system using blockchain. In: 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), pp. 1–4, April 2019. IEEE (2019). <https://doi.org/10.1109/IoT-SIU.2019.8777471>
6. Alvi, S.T., Uddin, M.N., Islam, L., Ahamed, S.: DVTChain: a blockchain-based decentralized mechanism to ensure the security of digital voting system voting system. *J. King Saud Univ.-Comput. Inf. Sci.* **34**(9), 6855–6871 (2022). <https://doi.org/10.1016/j.jksuci.2022.06.014>
7. Indapwar, A., Chandak, M., Jain, A.: E-voting system using Blockchain technology. *Int. J. Adv. Trends Comput. Sci. Eng.* **9**(3) (2020). <https://doi.org/10.30534/ijatcse/2020/45932020>
8. Ramesh, S.S., Venkataraja, D., Bharadwaj, R.N., Kumar, M.S., Santhosh, S.: E-voting is based on blockchain technology. *Int. J. Eng. Adv. Technol.* **8**(5), 107–109 (2019)
9. Yadav, A.S., Urade, Y.V., Thombare, A.U., Patil, A.A.: E-voting using blockchain technology. *Int. J. Eng. Res. Technol.* **9**(7) (2020)
10. Taş, R., Tanrıöver, Ö.Ö.: A systematic review of challenges and opportunities of blockchain for E-voting. *Symmetry* **12**(8), 1328 (2020). <https://doi.org/10.3390/sym12081328>
11. Khoury, D., Kfoury, E.F., Kassem, A., Harb, H.: Decentralized voting platform based on the Ethereum blockchain. In: 2018 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), pp. 1–6, November 2018. IEEE (2018). <https://doi.org/10.1109/IMCET.2018.8603050>
12. Dalvi, Y., Jaiswal, S., Sharma, P.: E-voting using blockchain. *Int. J. Eng. Res. & Technol. (IJERT)* (2021)
13. Malkawi, M., Yassein, M.B., Bataineh, A.: Blockchain-based voting system for Jordan parliament elections. *Int. J. Electr. Comput. Eng.* **11**(5), 4325 (2021). <https://doi.org/10.11591/ijece.v11i5.pp4325-4335>
14. Al-Zoubi, A., Aldmour, M., Aldmour, R.: Preserving transparency and integrity of elections utilizing blockchain technology. *J. Telecommun. Digit. Econ.* **10**(4), 24–40 (2022)
15. Hajian Berenjestanaki, M., Barzegar, H.R., El Ioini, N., Pahl, C.: Blockchain-based e-voting systems: a technology review. *Electronics* **13**(1), 17 (2023). <https://doi.org/10.3390/electronics13010017>
16. Singh, S., Wable, S., Kharose, P.: A review of e-voting system based on blockchain technology. *Int. J. New Pract. Manag. Eng.* **10**(04), 09–13 (2021). <https://doi.org/10.17762/ijnpm.v10i04.125>
17. Fatih, R., Arezki, S., Gadi, T.: A review of blockchain-based e-voting systems: comparative analysis and findings. *Int. J. Interact. Mob. Technol.* **17**(23), 49–67 (2023). <https://doi.org/10.3991/ijim.v17i23.45257>

18. Jafar, U., Ab Aziz, M.J., Shukur, Z., Hussain, H.A.: A systematic literature review and meta-analysis on scalable blockchain-based electronic voting systems. *Sensors* **22**(19), 7585 (2022). <https://doi.org/10.3390/s22197585>
19. Ranjbari, P., Sheikahmadi, S.A.: A systematic literature review of blockchain-based e-voting. *Soft Comput. J.* **9**(2), 14–33 (2021). <https://doi.org/10.22052/scj.2021.242836.0>
20. Vivek, S.K., Yashank, R.S., Prashanth, Y., Yashas, N., Namratha, M.: E-voting systems using blockchain: an exploratory literature survey. In: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 890–895, July 2020. IEEE (2020). <https://doi.org/10.1109/ICIRCA48905.2020.9183185>
21. Vladucu, M.V., Dong, Z., Medina, J., Rojas-Cessa, R.: E-voting meets blockchain: a survey. *IEEE Access* **11**, 23293–23308 (2023). <https://doi.org/10.1109/ACCESS.2023.3253682>
22. Oprea, S.V., Băra, A., Andreescu, A.I., Cristescu, M.P.: Conceptual architecture of a blockchain solution for e-voting in elections at the university level. *IEEE Access* **11**, 18461–18474 (2023). <https://doi.org/10.1109/ACCESS.2023.3247964>
23. Granata, D., Rak, M., Palmiero, P., Pastena, A.: A methodology for vulnerability assessment and threat modeling of an e-voting platform based on the Ethereum blockchain. *IEEE Access* (2024). <https://doi.org/10.1109/ACCESS.2024.3495981>
24. Çabuk, U.C., Adiguzel, E., Karaarslan, E.: A survey on feasibility and suitability of blockchain techniques for the e-voting systems (2020). [arXiv:2002.07175](https://arxiv.org/abs/2002.07175), <https://doi.org/10.48550/arXiv.2002.07175>
25. Sahib, R.H., Al-Shamery, E.S.: A review on distributed blockchain technology for e-voting systems. *J. Phys. Conf. Ser.* **1804**(1), 012050 (2021). <https://doi.org/10.1109/ICE3IS59323.2023.10335317>. IOP Publishing
26. Ohize, H.O., Onumanyi, A.J., Umar, B.U., Ajao, L.A., Isah, R.O., Dogo, E.M., Nuhu, B.K., Olaniyi, O.M., Ambafi, J.G., Sheidu, V.B., Ibrahim, M.M.: Blockchain for securing electronic voting systems: a survey of architectures, trends, solutions, and challenges. *Clust. Comput.* **28**(2), 132 (2025)
27. Daraghmi, E., Hamoudi, A., Abu Helou, M.: Decentralizing democracy: secure and transparent e-voting systems with blockchain technology in the context of palestine. *Future Internet* **16**(11), 388 (2024). <https://doi.org/10.3390/fi16110388>
28. Lindmark, M., Salihovic, A.A.: Investigating the security of end-to-end and blockchain-based electronic voting systems: a comparative literature review (2022)
29. Aidynov, T., Goranin, N., Satybalina, D., Nurusheva, A.: A systematic literature review of current trends in electronic voting system protection using modern cryptography. *Appl. Sci.* **14**(7), 2742 (2024). <https://doi.org/10.3390/app14072742>
30. Khan, K.M., Arshad, J., Khan, M.M.: Investigating performance constraints for blockchain-based secure e-voting system. *Future Gener. Comput. Syst.* **105**, 13–26 (2020). <https://doi.org/10.1016/j.future.2019.11.005>
31. De Vega Alforte, A.: Assessing the implementation challenges of e-voting on the electoral integrity of Asian democracies: a systematic review of literature. *SSRN* 4909917 (2024). <https://ssrn.com/abstract=4909917>
32. Daramola, O., Thebus, D.: Architecture-centric evaluation of blockchain-based smart contract e-voting for national elections. *Informatics* **7**(2), 16 (2020). <https://doi.org/10.3390/informati cs7020016>. MDPI
33. Muyambo, E., Baror, S.: Systematic review to propose a blockchain-based digital forensic ready internet voting system. In: International Conference on Cyber Warfare and Security, vol. 19, no. 1, pp. 219–230, March 2024 (2024). <https://doi.org/10.34190/iccws.19.1.2188>
34. Gochhi, S.K., Sahoo, S., Samanta, P.K., Panda, S.K., Sahoo, J.R.: Blockchain-based comparative analysis of e-voting systems: a review. In: 2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC), pp. 1–6, January 2024. IEEE (2024). <https://doi.org/10.1109/ASSIC60049.2024.10507924>
35. Satpathy, S., Mahapatra, S., Singh, A.: Fusion of blockchain technology with 5G: a symmetric beginning. In: Tanwar, S. (eds.) *Blockchain for 5G-Enabled IoT*. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67490-8_3

BLESS: Blockchain-Enhanced Intelligent Security System Using BPSO and AVOA for Smart Home Network



Amrutanshu Panigrahi, Nilachakra Dash, Abhilash Pati, Bibhuprasad Sahu, Bidya Bhusan Panda, and Ghanashyam Sahoo

Abstract Blockchain and machine learning assure data, resilience, and intelligent threat detection, including blockchain and machine learning. The tamper-resistant blockchain protects sensitive information with a decentralized storage mechanism; machine learning adds the 'smartness' in real-time concerning emerging risks in this home. However, a more systematic framework that integrates secure storage mechanisms and streamlined, accurate classification of IoT data must be developed to enhance threat detection capabilities. To bridge this gap, the present work presents BLESS, a hybrid model designed explicitly for IoT security within a smart home combining Binary Particle Swarm Optimization (BPSO) and African Vulture Optimization Algorithm (AVOA) for feature selection, and as the fitness evaluator, it uses a Support Vector Machine. For secure data storage, BLESS has used a Blockchain Inter Planetary File System (IPFS) server so that management is decentralized and has tamper-proofing of IoT data. The model was tested using two prominent IoT datasets, NSL-KDD and UNSW-NB15. It shows an accuracy of 98.97% and 97.75%, which surpasses standalone BPSO and AVOA in accuracy, precision, recall, specificity, and F1-score. All these results show that BLESS is very effective for robust adaptive threat detection. Future work would be in the direction of scaling BLESS for other applications in IoT and improving feature selection with enhanced capabilities toward adaptability and accuracy.

A. Panigrahi (✉) · A. Pati

Department of CSE, Siksha 'O' Anusandhan (Deemed to Be University), Bhubaneswar, Odisha, India

e-mail: amrutansup89@gmail.com

N. Dash

Department of CSE, Nalla Malla Reddy Engineering College, Hyderabad, India

B. Sahu

Department of IT, Vardhaman College of Engineering (Autonomous), Hyderabad, India

B. B. Panda

School of Electrical Engineering, KIIT University, Bhubaneswar, India

G. Sahoo

Department of CSE, GITA Autonomous College, Bhubaneswar, Odisha, India

Keywords Blockchain · Machine learning · Binary Particle Swarm Optimization (BPSO) · African Vulture Optimization Algorithm (AVOA) · Inter Planetary File System (IPFS)

1 Introduction

The Internet of Things (IoT) is an interconnected global system of physical devices, services, and people that aims to make people's lives easier and better by connecting everyday objects and services with unique identifiers for communication and collaboration. Manufacturing, transportation, and use statistics are just a few areas that may benefit from the information sharing made possible by the Internet of Things. One of the most common applications of the Internet of Things is in smart homes, which provide users with an enhanced quality of life by implementing automated appliance controls and assistive services. By employing context awareness and predetermined restrictions that emerge from situations inside the home environment, devices that are part of the Internet of Things (IoT) work together to improve consumer results [1].

Through a home-based application, smart homes provide services that enable individuals to live their lives in a safer, more comfortable, and more convenient manner. However, the essential thing to remember is that each of these applications generates a substantial amount of personal data that may be transferred to various service providers. As a consequence of this, hostile attackers may target the functionalities of the network that are responsible for the interchange of data [2]. A bright house is connected to the internet and may have a variety of innovative gadgets controlled by its residents. In the house, every appliance serves an essential purpose for the person and their loved ones. An intelligent home network built on the Internet of Things links various smart devices, including smartphones, innovative laptops, and wearables [3]. Homeowners might make their lives easier and safer by enhancing the accessibility and security of their homes. Customers and system developers have been motivated to do extensive studies to take advantage of the smart home's valuable features, including monitoring behaviors and even safety testing [4].

It is possible to find solutions to these issues by using blockchain-like technologies and unified computing networks similar to the cloud. Blockchain technology comes with a time-stamped collection of harmful-proof documentation controlled by a network of independent networks [5]. The blockchain architecture comprises a sequence of blocks connected via straightforward cryptography. Rigidity, decentralization, and transparency are core concepts that all blockchain technologies possess. All three roles have performed exceptionally well, allowing them to gain experience in every conceivable kind of digital money technology. These technologies include the functionalities of mobile vehicles, cellular devices, and embedded systems. Despite the security and anonymity of blockchain technology, specific issues persist in its implementation. An illustration of this is the increasing intricacy of Sybil

assaults, which include the creation of several fraudulent identities to dominate the community [6].

1.1 Objective

This research mainly focuses on providing a model for security in the smart home. The proposed hybrid method adopts the Binary Particle Swarm Optimization (BPSO) and African Vultures Optimization Algorithm (AVOA) as a feature selection algorithm. In addition, the proposed model BLESS implements the Support Vector Machine (SVM) as the fitness calculator. The objectives of this research work can be summarized as follows:

- To include the BPSO and AVOA methods for selecting appropriate features.
- To include SVM as the fitness calculator for calculating the fitness function of the BPSO and AVOA.
- To develop BLESS, a hybrid model with an SVM classifier for effectively classifying the IoT-based data.
- To evaluate the performance of the proposed model using different machine learning-based evaluative parameters.

2 Literature Survey

Andoni et al. [7] conducted a comprehensive examination of many alternative blockchain applications inside a peer-to-peer resource-sharing network and published their results. The research offers comprehensive insights into the execution and functionalities of diverse smart home networks, including security within the smart grid, artificial intelligence (AI), data analytics, and payment systems. Conversely, their study inadequately covered subjects pertinent to smart homes, like smart home security and financial planning for smart cities. Khan et al. [8] presented a user-centric blockchain architecture to improve the security of edge data transfer inside the Internet of Things. J. Wu et al. [9] introduced a software-defined blockchain interface to discern changing configurations.

Khan et al. [10] contributed to maintaining secrecy and integrity. This sensor provides secure data collection, encryption, and querying for applications specifically designed for smart homes. The information sent back and forth between the person, the gateway, the network operator, and the system is preserved, which helps to ensure that the information is verified and that privacy is maintained. In this day and age of digital technology, the exponential expansion of Internet of Things (IoT) devices presents enterprises with various design issues connected to privacy and security. Previous studies suggest blockchain technology is a crucial response to data security issues associated with the Internet of Things (IoT). Using blockchain technology,

numerous data suppliers can securely and dependably communicate information with one another [11].

In [12], Lee et al. present a smart home solution built on Ethereum. This solution aims to minimize the difficulties of secrecy, integrity, and authentication associated with the Internet of Things devices. In addition, the architecture overcomes issues around centralized gateways; nevertheless, it does not address the additional computational complexity that blockchain brings.

Xu et al. [13] built and executed a decentralized smart home system based on Ethereum. Blockchain technology gives rise to Ethereum, a software platform that enables developers to construct and deploy decentralized applications. Ethereum was created to empower developers in this endeavor. In light of this, the authors have used Ethereum to build smart contracts to store sensor data. This is accomplished by continuous monitoring.

On the other hand, the authors of this work do not indicate that their system is expensive to operate and that certain aspects of the design need to be improved further. She et al. [14] discussed the architecture of their consortium's blockchain-based smart home system to address data privacy concerns precisely. The model's performance is appraised favorably via simulation techniques; nevertheless, there is no explanation for the energy it consumes or the time it takes to process.

3 Methodology and Dataset Description

The proposed BLESS models utilize the BPSO and AVOA algorithms with the SVM classifier as the fitness evaluator and classification algorithm. The proposed BLESS is evaluated over two IoT-based datasets, NSL-KDD and UNSW-NB15. The key motivation for choosing the BPSO as a feature selector is that it efficiently searches large solution spaces and can handle complex, nonlinear relations between features. BPSO is also computationally efficient and often finds the optimum or near-optimum subsets of features with fewer evaluations than conventional methods. Similarly, AVOA presents healthy benefits to feature selection due to its balance of exploration and exploitation. This will help prevent local optima and enhance its global search capabilities. Inspired by the social and foraging behaviors of the vultures, it will be efficient in finding the best subsets of features within complicated and high-dimensional datasets.

3.1 Dataset Description

The proposed BLESS model is evaluated over two publicly available datasets, NSL-KDD [15] and UNSW-NB15 [16]. The NSL-KDD dataset contains 42 features with 148,517 numbers samples with two classes, regular and anomaly. Similarly, the UNSW-NB15 contains 45 features with 257,363 samples. The UNSW-NB15 consists

Table 1 Dataset description

Dataset	Number of features	Number of samples	Class
NSL-KDD	42	148,517	Normal-77,054
			Anomaly-71,463
UNSW-NB15	45	257,363	Normal-93,000
			Anomaly-164,673

of attacks, including DoS, Backdoor, fuzzer, shellcode, etc. To label the attack and standard samples, all types of attacks are considered as a single class ‘anomaly,’ and the remaining is named ‘normal.’ The exact distribution of the datasets is depicted in Table 1.

3.2 Binary Particle Optimization Algorithm (BPSO)

A binary version of the traditional Particle Swarm Optimization algorithm, BPSO takes for the binary search space instead of continuous ones. In BPSO, each particle within the swarm is a potential solution for optimization. Each dimension of the location vector is bounded between 0 and 1 for any particle since it only represents the binary values 0 or 1. The algorithm iteratively updates every particle’s location concerning its velocity, which considers both the particle’s own best historical or personal best and the global best position for the entire swarm. The velocity values are computed in the continuous space and then mapped to probabilities using the sigmoid function, which guides the process of updating the binary position of the particle through a decision of whether to move from 0 to 1 or vice versa for each dimension. Such transformation allows BPSO to efficiently explore the binary solution space based on the principles of swarm intelligence to find the optimal or near-optimal solution by balancing exploration and exploitation to solve different applications in feature selection, network security, and combinatorial optimization challenges [17]. The velocity of each particle in BPSO can be represented in Eq. 1.

$$vel_{i,d}(t+1) = \omega \cdot vel_{i,d}(t) + \rho_1 \cdot \tau_1 \cdot (P_{best(i,d)}) + \rho_2 \cdot \tau_2 \cdot (G_{best} - P_{i,d}(t)) \quad (1)$$

where $vel_{i,d}(t+1)$, and $vel_{i,d}(t)$ are the velocity of the i th particle P_i at time $t+1$ and t at dimension d , ω is the inertia weight which controls the $vel_{i,d}(t)$, $P_{best(i,d)}$, and G_{best} are the local best of particle P_i in dimension d , and G_{best} is the global best obtained in the population. ρ_1 and ρ_2 are the cognitive and social coefficients. High ρ_1 will be stuck in the search process in local optima. So, keeping ρ_1 low and ρ_2 as high is highly recommended. For the current scenario, the ρ_1 is kept at 0.1, and ρ_2 is kept at 2.0 to maintain the balance between the local and global search processes. τ_1 and τ_2 are the random numbers drawn from a uniform distribution [0, 1]. To convert the obtained continuous velocity to a binary one, the sigmoid function (S) is applied.

The sigmoid function (S) can be represented by using Eq. 2.

$$S(vel_{i,d}(t+1)) = \frac{1}{1 + e^{-vel_{i,d}(t+1)}} \quad (2)$$

Based on the obtained velocity using the Sigmoid function, the position of the P can be adjusted by using Eq. 3 with r as the uniform random distribution in the range of [0, 1].

$$P_{i,d}(t+1) = \begin{cases} 1 & \text{if } r < S(vel_{i,d}(t+1)) \\ 0 & \text{Otherwise} \end{cases} \quad (3)$$

3.3 African Vulture Optimization Algorithm (AVOA)

The vultures' hunting and scrounging inspire the AVOA. As vultures in the search space, agents offer one alternative solution that adaptively explores the space utilizing exploitation and exploration behavior. Exploration resembles how vultures scan the terrain for food or solutions, covering the search space. Vultures roam around and circle the potential food sources to identify the best ones to exploit. It enhances the search for good-quality solutions, fine-tuning the discovery. AVOA uses the mathematical model of phases that adjusts accordingly to every location of the vulture so that all the best solutions and probabilistic principles converge together. AVOA solves complicated optimization issues with high-quality solutions by adapting between broad and targeted search patterns and achieving efficient global optima convergence. The AVOA initializes with an initial population of randomly positioned vultures representing candidate solutions. In the iteration step, the vultures balance exploration (more exhaustive search) and exploitation (refining search in promising regions) depending on the dynamic control factors f. Phase-dependent updates of the position of the vultures will be applied according to the best-known solution, mean positions, or random positions. The algorithm iterates all the positioning of vultures and continues until it fulfills a stopping criterion given by the maximum number of iterations [18]. The position of vulture V in the exploration phase can be calculated using Eq. 4. The excessive movement of V is controlled by applying the Sigmoid function (S) to the position of the V. The utilized S can be represented as Eq. 5.

$$V_i(t+1) = V_i(t) + S(F * (V_{best}(t) - V_i(t)) * R + F * (V_{random} - V_i(t))) \quad (4)$$

$$S(x) = \frac{1}{1 + e^{-x}} \quad (5)$$

F is the controlling parameter that balances between the exploration and exploitation of V, which can be represented using Eq. 6.

$$F = F_{max} - \left(\frac{t}{T} \right) * (F_{max} - F_{min}) \quad (6)$$

F_{max} and F_{min} are the maximum and minimum values of F . For the current work, F_{max} is set to 2.0, and F_{min} is set to 0.2 to enhance the global search ability over local search. The Position of the V in the exploitation phase can be updated using Eq. 7.

$$V_i(t+1) = V_i(t) + S(c1 * D * (V_{best}(t) - V_i(t)) * R + c2 * D(V_{mean} - V_i(t))) \quad (7)$$

where $c1$ and $c2$ are the control parameters for the acceleration of V , D can be represented as the distance factor indicating the closeness to the best position, which can be described using Eq. 8 with α as the scaling factor of the sigmoid function's steepness which is set to 0.5 in the current work.

$$D = S(\alpha * |V_{best}(t) - V_i(t)|) \quad (8)$$

3.4 Fitness Function

For calculating the fitness function (Fit) for the selected feature f_i , the SVM [19] classifier is represented using Eq. 9 with BA as the Balanced Accuracy of SVM using f_i .

$$Fit(f_{selected}) = 1 - BA(SVM(f_{selected}), C) \quad (9)$$

BA is the balanced accuracy, which can be represented using Eq. 9 with $t1$, $t2$, $f1$, and $f2$ as the true positive, true negative, false positive, and false negative, respectively.

$$BA = \frac{1}{2} \left(\frac{t1}{t1 + f2} + \frac{t2}{t2 + f1} \right) \quad (10)$$

3.5 Workflow of the Proposed Model

The proposed BLESS model adopts the BPSO and AVOA algorithms to select the features from the dataset. Based on the fitness function, the features are evaluated by the Fitness function using SVM. The workflow of the proposed model is represented in Fig. 1. The work of the proposed model can be summarized below.

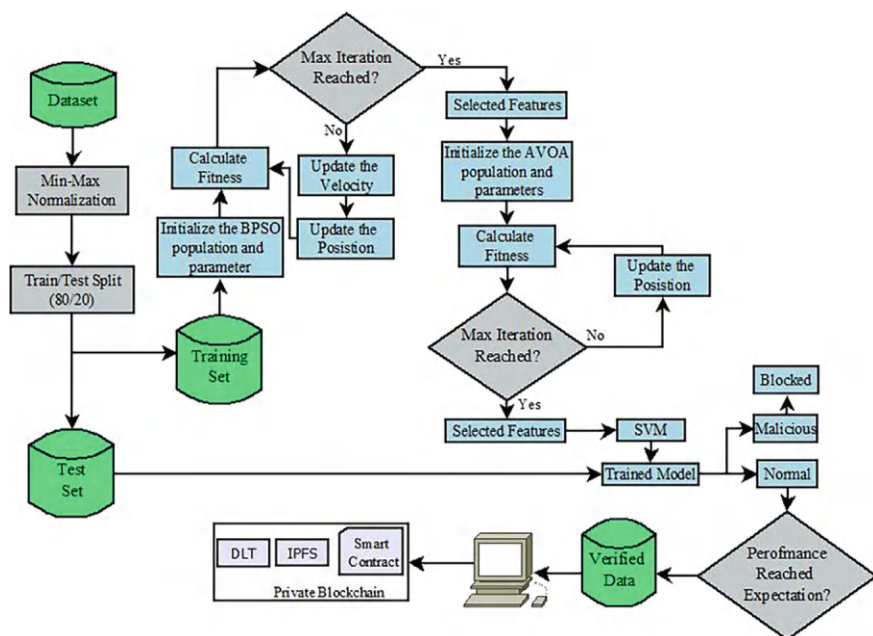


Fig. 1 Workflow of the proposed BLESS model

Step 1: Initialize the Dataset.

Step 2: Min Max normalization.

Step 3: Train/Test Split with a distribution ratio of 0.2.

Step 4: To the Training set, initiate the parameters and population of PSO.

- i. Initiate the population of the particle ($n = 100$)
- ii. Find the Fit() using Eq. 9
- iii. If $t < T$, where t is the current iteration, and T is the maximum iteration

- Update the Velocity
- Update the Position
- Find the Fit() using Eq. 9
- Go to step iii.

iv. Otherwise, identify the optimal feature set.

Step 5: To the optimal set selected by BPSO, initiate parameters and population of AVOA.

- i. Initiate the population of the particle ($n = 100$)
- ii. Find the Fit() using Eq. 9
- iii. If $t < T$, where t is the current iteration, and T is the maximum iteration

- Update the Velocity
- Find the Fit() using Eq. 9
- Go to step iii.

iv. Otherwise, identify the optimal feature set.

Step 6: Apply SVM for classification to classify regular or malicious activity.

Step 7: If the activity is found normal (accuracy > 95%).

- i. Send the instruction to the access point
- ii. Store the activity in DLT of the private blockchain
- iii. Send the instruction to home appliances to start in the smart home

Step 8: Otherwise, mark the activity as malicious and block it.

4 Result and Discussion

The proposed model BLESS is evaluated in a system with an Intel core i7 processor with 4.1 GHz clock speed, 16 GB of RAM, 1 GB of NVIDIA GeForce Graphics, and Ubuntu 22.04 LTS operating system. Six different parameters, including accuracy (A_Y), precision (P_N), recall (R_L), specificity (S_Y), F-1 Score (F_1S), and AUC score, are calculated to evaluate the proposed model. Equations 11–15 quantify the evaluative parameters. Table 2 shows the average number of features selected (AVG_F) by the PSO, AVOA, and the proposed model BLESS. Table 3 shows the performance of the proposed BLESS model.

$$A_Y = \frac{t_1 + t_2}{t_1 + t_2 + f_1 + f_2} \quad (11)$$

$$P_N = \frac{t_1}{t_1 + f_1} \quad (12)$$

$$R_L = \frac{t_1}{t_1 + f_2} \quad (13)$$

$$S_Y = \frac{t_2}{t_1 + f_2} \quad (14)$$

$$F_1S = \frac{2 * \frac{t_1}{t_1 + f_1} * \frac{t_1}{t_1 + f_2}}{\frac{t_1}{t_1 + f_1} + \frac{t_1}{t_1 + f_2}} \quad (15)$$

Table 3 Representation of the thorough performance evaluation of the BLESS algorithm against the two benchmark datasets, namely NSL-KDD and UNSW-NB15, in terms of critical parameters: Accuracy (A_Y), Precision (P_Y), Recall (R_L), Specificity (S_Y), and F1-Score (F_1S). Anyhow, regarding the NSL-KDD data set, BLESS

Table 2 Average number of selected features

Dataset	AVG_F		
	BPSO	AVON	BLESS
NSL-KDD	26.7	20.5	12.4
UNSW-NB15	26.7	22.8	13.2

Table 3 Performance evaluation of BLESS

	A_Y	P_Y	R_L	S_Y	$F1_S$
NSL-KDD	98.97	97.12	96.61	97.66	96.86
UNSW-NB15	97.75	96.56	95.89	97.17	96.23

Table 4 Accuracy comparison between BLESS, BPSO, AVOA

	BLESS	BPSO	AVON
NSL-KDD	98.97	93.45	95.12
UNSW-NB15	97.75	92.67	93.78

revealed better performance-related outcomes in terms of accuracy of 98.97%, precision of 97.12%, recall of 96.61%, specificity of 97.66%, and an F1-score of 96.86%. On the UNSW-NB15 dataset, too, BLESS maintains great metrics, with accuracy at 97.75%, precision at 96.56%, recall at 95.89%, specificity at 97.17%, and an F1-score of 96.23%. BLESS is consistent and robust across the evaluation measures on both datasets. Table 4 shows the performance analysis of two datasets with BPSO and AVOA feature selection techniques with the proposed BLESS methodology. For the NSL-KDD dataset, BLESS acquired an accuracy of 98.97%, surpassing BPSO (93.45%) and AVOA with 95.12%. For the UNSW-NB15 dataset, BLESS also reported an accuracy of 97.75%, substantially more than the 92.67% accuracy for BPSO and 93.78% for AVOA. Figures 2 and 3 show the ROC analysis of the proposed BLESS for the NSL-KDD and UNSW-NB15 datasets, respectively

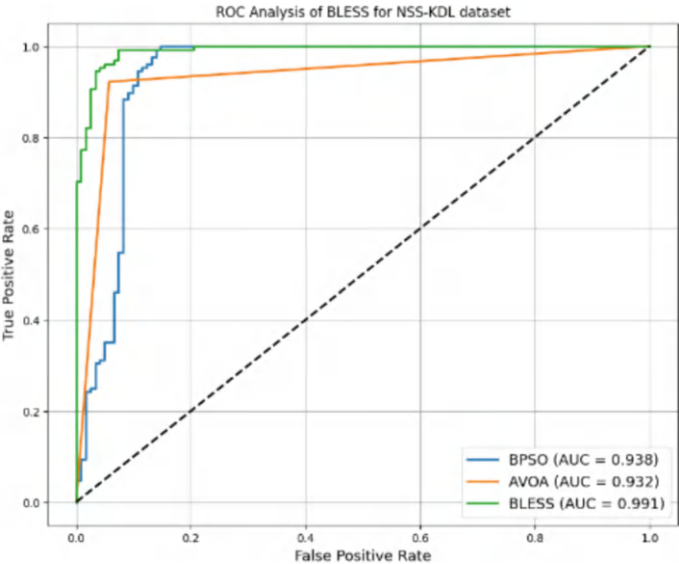


Fig. 2 ROC analysis of BLESS, BPSO, and AVOA for the NSL-KDD dataset

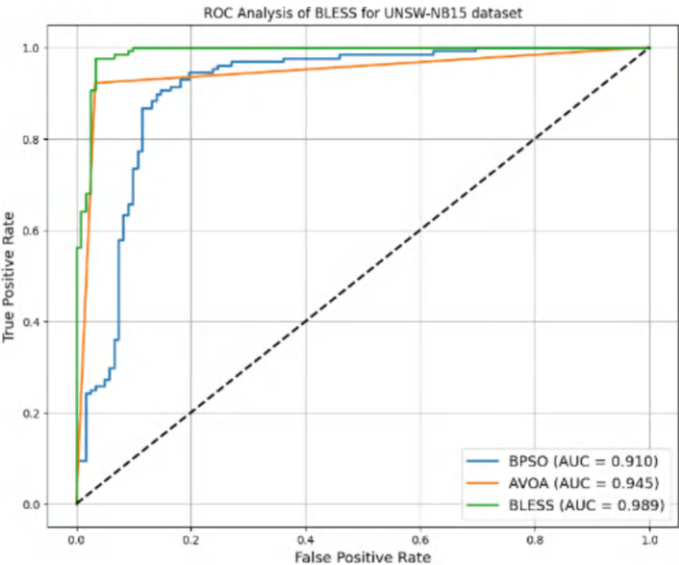


Fig. 3 ROC analysis of BLESS, BPSO, and AVOA for the UNSW-NB15 dataset

5 Conclusion

This paper proposes the hybrid approach, specifically the BLESS model, that improves security in IoT by integrating feature selection with robust classification. It uses both BPSO and AVOA algorithms for the feature selection and applies SVM algorithms as the fitness evaluator; the proposed BLESS model strongly appeared to be very efficient on critical metrics like accuracy, precision, recall, specificity, and F1-score on both datasets of NSL-KDD and UNSW-NB15. BLESS outperformed BPSO and AVOA in isolation, achieving an accuracy of 98.97% on NSL-KDD and 97.75% on UNSW-NB15 datasets. It demonstrates to successfully function correctly for IoT data in detecting security threats. The good ROC performance of this model also puts it into the category of reliable models for IoT security applications. Future work will be on the extension of BLESS to manage even more complex and large-scale IoT environments and integrate other sophisticated feature selection techniques and machine learning algorithms to enhance its adaptability in detection across various IoT domains. Finally, furthering BLESS into real-time deployment and scalability within distinct IoT-based smart infrastructures will firmly solidify its applicability in practical settings.

References

1. Magara, T., Zhou, Y.: Internet of Things (IoT) of smart homes: privacy and security. *J. Electr. Comput. Eng.* **2024**(1), 7716956 (2024)
2. Pati, A., Panigrahi, A., Parhi, M., Pattanayak, B.K., Sahu, B., Kant, S.: Simulating fog of medical things: research challenges and opportunities. *IEEE Access* (2024)
3. Netinant, P., Utsanok, T., Rukhiran, M., Klongdee, S.: Development and assessment of internet of things-driven smart home security and automation with voice commands. *IoT* **5**(1), 79–99 (2024)
4. Pati, A., Parhi, M., Pattanayak, B.K.: IoT-fog-edge-cloud computing simulation tools: a systematic review. *Int. J. Smart Sens. Adhoc Netw.* **3**(2) (2022)
5. Panigrahi, A., Nayak, A.K., Paul, R., Sahu, B., Kant, S.: CTB-PKI: clustering and trust enable blockchain-based PKI systems to facilitate efficient communication in P2P networks. *IEEE Access* **10**, 124277–124290 (2022)
6. Panigrahi, A., Nayak, A.K., Paul, R.: Smart contract assisted blockchain-based public key infrastructure system. *Trans. Emerg. Telecommun. Technol.* **34**(1), e4655 (2023)
7. Andoni, M., Robu, V., Flynn, D., Abram, S., Geach, D., Jenkins, D., McCallum, P., Peacock, A.: Blockchain technology in the energy sector: a systematic review of challenges and opportunities. *Renew. Sustain. Energy Rev.* **100**, 143–174 (2019). [CrossRef]
8. Khan, M.A., Abbas, S., Rehman, A., Saeed, Y., Zeb, A., Uddin, M.I., Nasser, N., Ali, A.: A machine learning approach for blockchain-based smart home network security. *IEEE Netw.* **35**, 223–229 (2020)
9. Wu, J., Dong, M., Ota, K., Li, J., Yang, W.: Application-aware consensus management for software-defined intelligent blockchain in IoT. *IEEE Netw.* **34**, 69–75 (2020)
10. Khan, M.A., Ghazal, T.M., Lee, S.W., Rehman, A.: Data Fusion-based machine learning architecture for intrusion detection. *Comput. Mater. Contin.* **70**, 3399–3413 (2022)
11. Le Nguyen, B., Lydia, E.L., Elhoseny, M., Pustokhina, I., Pustokhin, D.A., Selim, M.M., Nguyen, G.N., Shankar, K.: Privacy-preserving blockchain technique to achieve secure and reliable sharing of IoT data. *Comput. Mater. Contin.* **65**, 87–107 (2020)

12. Lee, Y., Rathore, S., Park, J.H., Park, J.H.: A blockchain-based smart home gateway architecture for preventing data forgery. *Hum.-Centric Comput. Inform. Sci.* **10**(1), 1–14 (2020)
13. Xu, L., Bao, T., Zhu, L.: Blockchain empowered differentially private and auditable data publishing in industrial IoT. *IEEE Trans. Ind. Inform.* (2020)
14. She, W., Gu, Z.-H., Lyu, X.-K., Liu, Q., Tian, Z., Liu, W.: Homomorphic consortium blockchain for smart home system sensitive data privacy-preserving. *IEEE Access* **7**, 62058–62070 (2019)
15. NSL-KDD. (n.d.). www.kaggle.com, <https://www.kaggle.com/datasets/hassan06/nslkdd>
16. UNSW_NB15. (n.d.). www.kaggle.com, <https://www.kaggle.com/datasets/mrwellsdavid/unswnb15>
17. Sahu, B., Panigrahi, A., Rout, S.K., Pati, A.: Hybrid, multiple filters embedded political optimizer for feature selection. In: 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP), pp. 1–6, July 2022. IEEE (2022)
18. Abdollahzadeh, B., Gharehchopogh, F.S., Mirjalili, S.: African vultures optimization algorithm: a new nature-inspired metaheuristic algorithm for global optimization problems. *Comput. Ind. Eng.* **158**, 107408 (2021)
19. Panigrahi, A., Pati, A., Sahu, B., Das, M.N., Nayak, D.S.K., Sahoo, G., Kant, S.: En-MinWhale: An ensemble approach based on MRMR and Whale optimization for Cancer diagnosis. *IEEE Access* **11**, 113526–113542 (2023)

Navigating Security and Privacy in Blockchain: Challenges and Future Directions



Navjot Kaur  and Ramandeep Kaur 

Abstract This paper gives an overview of blockchain technology security and privacy issues, potential solutions, and possible research avenues. The study focuses on the first principles in analyzing the security risks of blockchain systems concerning consensus layer risks, smart contract risks, and network layer risks. It describes different attack vectors, including vulnerabilities within cryptographic protocols, Distributed Denial of Service attacks, transaction manipulation and exploitation of state channels, and prevention strategies. The paper focuses on privacy frameworks and protection measures, where special attention is paid to zero-knowledge proof, identity and access management, and compliance with the Data Protection Regulation. Future development of security solutions includes the integration of artificial intelligence, quantum-safe encryption, and cross-chain security architecture, which are under consideration. The paper also emphasizes the growing issues of size, security, and privacy in the evolution of the blockchain network in the quantum regime. The paper also posits that the security and confidentiality of blockchain systems are still evolving, with future issues expecting to add new cryptographic approaches, intelligent frameworks, and quantum-safe solutions. This continuous change requires constant research and development to ensure Blockchain stays a key attribute of a secure, scalable, and privacy-preserving distributed system.

Keywords Block chain · Security · Privacy mechanism · Protection mechanism · Zero-knowledge proof · Risks

N. Kaur
Chandigarh University, Mohali, Punjab, India
e-mail: navkaurtoor@gmail.com

R. Kaur (✉)
Dayananda Sagar University, Bangalore, India
e-mail: ramangrewalg@gmail.com

1 Introduction

Blockchain technology has become a revolutionary paradigm in distributed systems and digital transactions, and the paper presents a broad overview of critical security and privacy challenges in Blockchain, along with proposed solutions and future directions. Naturally, the decentralized paradigm offered by Blockchain has many merits yet introduces unique vulnerabilities and demands sophisticated measures of security and privacy-preserving mechanisms. The essential characteristics of this technology mutation, but transparency and distributed consensus, make this landscape complex, with the necessity of profound balancing of considerations about security and privacy [1].

Although there are various advantages of using a decentralized paradigm of Blockchain, it also poses unique vulnerabilities that need advanced security and privacy-preserving measures. With the increase of blockchain adoption in various sectors, there is a need to examine multiple security challenges related to it to ensure its sustainable and trustworthy integration into applications.

We got motivation for researching blockchain security and privacy from the growth of blockchain implementation in blockchain management, healthcare, and finance sectors, and the consequences of security and privacy breaches in these extend beyond financial losses. And even risks to personal safety in identity management applications.

It is critical to talk about blockchain security and privacy; its fundamental understanding is fundamental. Blockchain can record transactions across a computer network, a distributed, undisputable ledger. Various blockchains exist, each having its own characteristics and security considerations. The basic types of blockchains are explained below:

1. **Public blockchains:** This type of Blockchain allows anyone to participate in the network, and it is called open and permissionless. Bitcoin and Ethereum are the most common public blockchains. This open nature presents security challenges related to anonymity, Scalability, and the potential for malicious actors to engage in illicit activities.
2. **Private blockchains:** These are permissioned, restricting access to authorized participants. Hyperledger blockchains are considered private blockchains where participants need to be verified. They could enhance security and offer greater control over the network while raising concerns about centralization.
3. **Consortium blockchains:** These are also known as federated blockchains. This is a hybrid approach where a selected group of organizations manages the network. It provides a balance between security, decentralization, Scalability, and efficiency.

Every type of Blockchain employs a different consensus mechanism to validate transactions and maintain the ledger's integrity. Consensus Mechanism is a set of rules that are considered to be the heart of Blockchain Security. A few of the consensus mechanisms are Proof-of-Work (PoW), Proof-of-Stake (PoS), Delegated Proof-of-Stake (DPoS), and Practical Byzantine Fault Tolerance (PBFT). These offer different

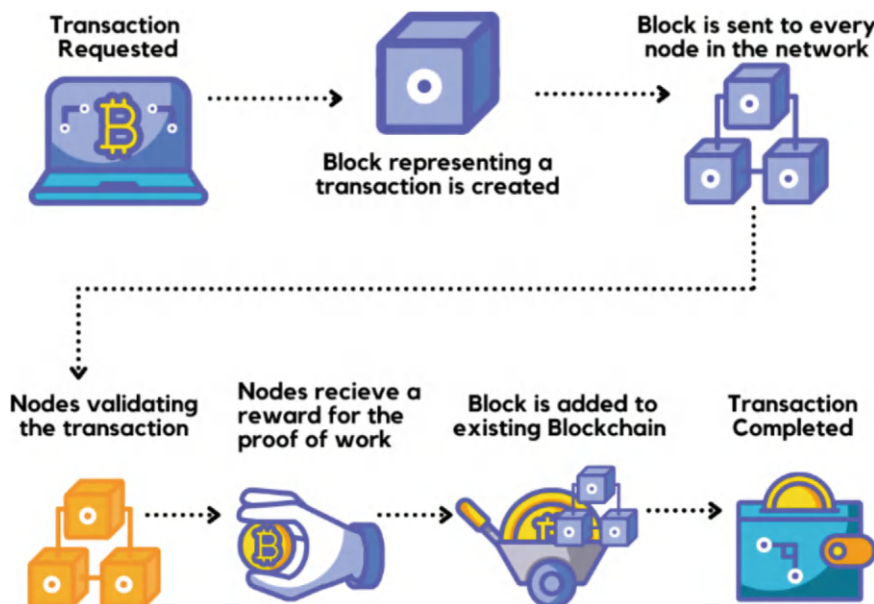


Fig. 1 How blockchain works

trade-offs regarding energy efficiency, Scalability, and security. Figure 1 explains how Blockchain works [2].

This paper provides a broad overview of the critical security and privacy challenges that must be considered in blockchain systems. We also proposed solutions and future work for these challenges. The study focuses on security challenges, including the potential for consensus mechanism attacks, such as 51% attacks and selfish mining, which can compromise the integrity of the Blockchain. This paper is structured to provide a complete investigation of these topics. Section 2 digs into the fundamental security architecture of blockchain systems, examining vulnerabilities in consensus mechanisms, smart contracts, and the network layer. Section 3 presents an attack surface analysis, focusing on cryptographic protocol vulnerabilities, Distributed Denial of Service (DDoS) attacks, transaction malleability, and state channel vulnerabilities. Defense mechanisms, including Segregated Witness and multi-signature schemes, are discussed. Section 4 explores privacy frameworks and protection mechanisms, focusing on zero-knowledge systems, identity and access management, and data protection and compliance. The mathematical underpinnings of zero-knowledge proofs are presented. Section 5 discusses advanced security solutions and future considerations, including AI-enhanced anomaly detection, quantum-safe encryption, and cross-chain security frameworks. The trade-offs between Scalability and security are also examined. Finally, Sect. 6 concludes the paper by summarizing key findings and highlighting future research directions.

2 Fundamental Security Architecture and Vulnerabilities

The security of blockchain systems and the integrity of their underlying architecture are linked intrinsically, which comprehends consensus mechanisms, smart contracts, and network infrastructure. Each of these presents unique weaknesses that, if not addressed, can weaken the security and reliability of the entire system. Understanding these weaknesses is essential for developing effective defense strategies for blockchain networks.

2.1 Consensus Mechanism Vulnerabilities

Consensus Mechanisms are the backbone of blockchain Technology. Several security challenges are related to consensus mechanisms, specifically within Proof-of-Work (PoW) and Proof-of-Stake (PoS). However, these mechanisms are not immune to attack. According to this, any network in PoW networks is vulnerable to 51% attacks since attackers have more than the transaction validation [3]. Under such conditions, the security limit can thus be written as:

$$P(\text{attack}) = \left(\frac{N}{n}\right)^k$$

where, 50% of the computing power can always be manipulated.

- P(attack) represents the probability of a successful attack
- n denotes the attacker's computing power
- N represents the total network computing power
- k indicates the number of confirmation blocks.

This equation highlights the importance of the attacker's relative computing power (n/N) and the number of confirmation blocks (k) in determining the success of a 51% attack. If the ratio of attacker computing power to total network computing power is higher, it increases the probability of a successful attack. If more confirmation blocks are added, it becomes more difficult for the attacker to overtake the main chain.

2.2 Smart Contract Security Issues

Smart contracts are one of the essential parts of the Blockchain that write very nonsubjective program code or decision points to define precisely how that transaction will be managed and what steps will be taken when that situation occurs, also known as chain code. Security issues in smart contracts arise at different levels: contract code level, virtual machine level, and business level. Vulnerabilities in smart

contracts (self-executing agreements stored on the Blockchain) present significant security risks, including issues like reentrancy attacks, integer overflow and underflow, and dependency on timestamps. These vulnerabilities can lead to unauthorized access, financial losses, and compromise. In other words, vulnerabilities make a smart contract susceptible to unauthorized access, financial losses, and system compromise. Because of that, smart contract security is one of the outstanding challenges for Blockchain. Smart contracts require high security because they deal with valuable information, e.g., cryptocurrencies, tokens, and other digital assets.

Deployed smart contracts cannot be easily modified, and the transactions built with smart contracts are irreversible. So, addressing these vulnerabilities is crucial for ensuring the security and reliability of blockchain applications. A few constraints are there on smart contracts to secure the blockchain environment from attackers and to cover vulnerabilities. Competent contract developers should design smart contracts with great attention against known or unknown attacks since not all contracts are secure enough.

2.3 Network Layer Threats

In the blockchain system, the network layer is exposed to various threats that can compromise its integrity. Including eclipse attacks, DDoS attacks, and routing attacks, threats in the network layer compromise blockchain integrity. These attacks can disrupt the operation of the Blockchain, censor transactions, and even lead to blockchain forks. DDoS attacks disrupt the consensus process and transaction validation. DDoS overwhelms all network services and makes resources unavailable to end users. Blockchain technology's decentralized and open nature makes it more susceptible to such attacks, as they flood the network with requests. An eclipse attack allows an attacker to eclipse a target node by isolating it from honest nodes in the network. Network connectivity can be lost, and it can cause blockchain forks because of routing attacks through BGP hijacking. On the other hand, Eclipse attacks are also hazardous because they are challenging to detect and prevent. Routing attacks discard the consensus and delay the transactions, leading to inconsistent blockchain data [4].

More secure and resilient blockchain systems can be developed by understanding the vulnerabilities in the consensus mechanisms, smart contracts, and network layer. Proactive identification and mitigation of these attacks to maintain the security of blockchains. A multi-layered advanced defense approach is required that contains secure coding practices, the usage of static analysis tools, and the implementation of vigorous defense mechanisms, etc. Implementing encryption of data and use of secure communication protocols for node communication can safeguard against network layer attacks.

3 Attack Surface Analysis and Defense Mechanisms

Blockchain networks present a complex attack surface while designed with robust security features. Such a network demands careful analysis and the implementation of effective defense mechanisms. This section of the paper explores several key attack vectors, including susceptibilities in cryptographic protocols, Distributed Denial of Service (DDoS) attacks, transaction malleability exploitation, and state channel vulnerabilities. Each of these areas poses unique challenges, requiring specialized mitigation strategies to maintain the integrity and availability of the blockchain network.

3.1 *Cryptographic Protocol Vulnerabilities*

To enable secure and trusted transactions between untrusted parties, Blockchain creates a layer of trust between them. It eliminates the need for a centralized authority or third party to act as intermediaries. Blockchain networks possess a notable vulnerability in their cryptographic protocol implementations, requiring highly sophisticated defense mechanisms. Mainly, the issues concern the implementation of the Elliptic Curve Digital Signature Algorithm that forms the backbone of any transaction verification system. There are potential vulnerabilities through collisions in hash functions within the Merkle tree. There are also issues concerning predictability in pseudo-random number generators, especially where the key generation methods in a system are deterministic.

Advanced strategies have been devised to mitigate the vulnerabilities involved. State-of-the-art encryptions like Advanced Encryption Standard (AES) for symmetric key operations and enhanced versions of Rivest-Shamir-Adleman (RSA) protocols for asymmetric encryption are used in present-day blockchain systems. Periodic policies for key rotation were set to ensure Perfect Forward Secrecy, meaning that captured data transmissions cannot be used for decryption even after the policy has passed.

3.2 *Distributed Denial of Service Vector Analysis*

Because it is decentralized, blockchain networks face unique challenges on the defense end against Distributed Denial of Service (DDoS) attacks [5], which attack the ability of the network to process transactions by using protocol-level vulnerabilities. The attacks are more about creating an overwhelming load in the network validation mechanisms by using complex transactions with high computation resources. Figure 2 explains the DDoS attack scenario in a cloud computing environment [5, 6].

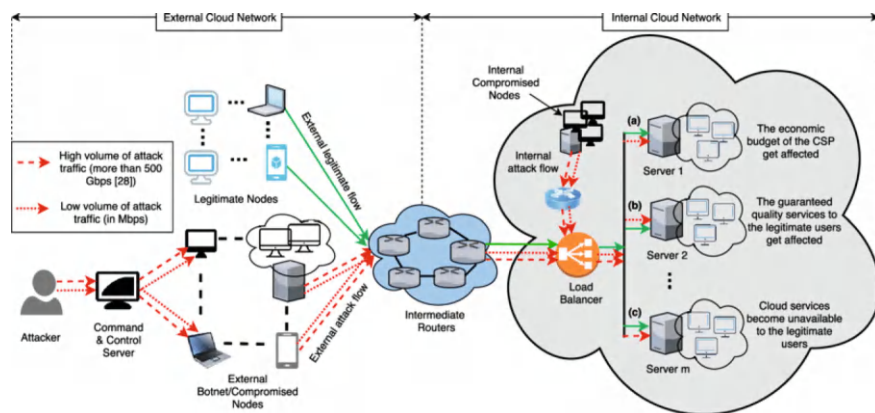


Fig. 2 DDoS attack scenario in a cloud computing environment

Blockchain networks adopt a multi-layered approach to security to mitigate the threat of DDoS attacks. For instance, TCP SYN cookie validation acts as a protective layer, blocking legitimate contact attempts. Furthermore, DDoS attacks are allowed through only once as the algorithms based on exponential backoff models are used for rate limiting [7]. Advanced systems also utilize Adaptive Proof-of-Work challenges when submitting transactions to ensure fair and efficient utilization of available resources.

3.3 Transaction Malleability Exploitation

The other advanced forms of attacks in the blockchain network include transaction malleability. This type primarily affects individuals whose smart contracts are more complicated. It works when it manipulates signatures of the transactions before confirmation to the network, and at the same time, it can lead to inconsistency in identifying the transaction. The cascading effect on the network may affect dependent transactions and execution paths in smart contracts.

Modern blockchain architectures address the issue of transaction malleability through the advanced implementation of the Segregated Witness architecture. This approach decouples transaction signatures from other transaction uses and protects against forgery [8]. The solution is effective because witness identifiers are assigned to be both unique and unchanging throughout the transaction.

3.4 State Channel Vulnerability Assessment

State channels, crucial for scaling blockchain networks, bring particular security considerations that need thorough evaluation and resolution. These vulnerabilities mainly come into play in multi-signature schemes, where race conditions are exploitable during signature verification. Besides this, manipulating a timeout mechanism is a danger in time-sensitive transaction protocols, especially in cross-chain operations.

There exist multiple strategic layers within a defense framework for state channel security. First, such an implementation of mandatory multi-signature validation protocols makes sure the transaction is indeed authentic because of distributed consensus. In addition, the atomic swap protocols maintain atomicity for transactions across all state channel operations. On the other hand, state-verifying mechanisms can be time-constrained to achieve transitions within some temporal bounds, so any verification mechanism would fail to stop timing-based attacks.

These consensus mechanisms that rely on the Byzantine Fault-Tolerant mechanism make security heightened at the state channel with assured system reliability, considering adverse parties [9]. Advanced protocols keep the integrity of the network by verifying complex algorithms that detect and isolate potentially harmful state transitions before they affect the overall network.

In such a comprehensive breakdown of the attack vectors, in addition to the corresponding countermeasures, blockchain systems can thus implement strong and sound security measures that focus on transaction integrity and network resilience, given that these security policies evolve in real time with emerging attack methodologies.

4 Privacy Frameworks and Protection Mechanisms

Privacy in the blockchain system requires complicated framework designs that balance transparency and confidentiality. Zero-knowledge-proof systems represent promising solutions to ensure simultaneous private transactions and system verifiability.

Advanced privacy-preserving solutions include ring signatures, zero-knowledge systems, and privacy-focused consensus mechanisms. These solutions must tackle the challenges of identity management, data exposure risks, and regulatory compliance while preserving the fundamental advantages of blockchain technology. Figure 3 explains the privacy-preserving blockchains [10].

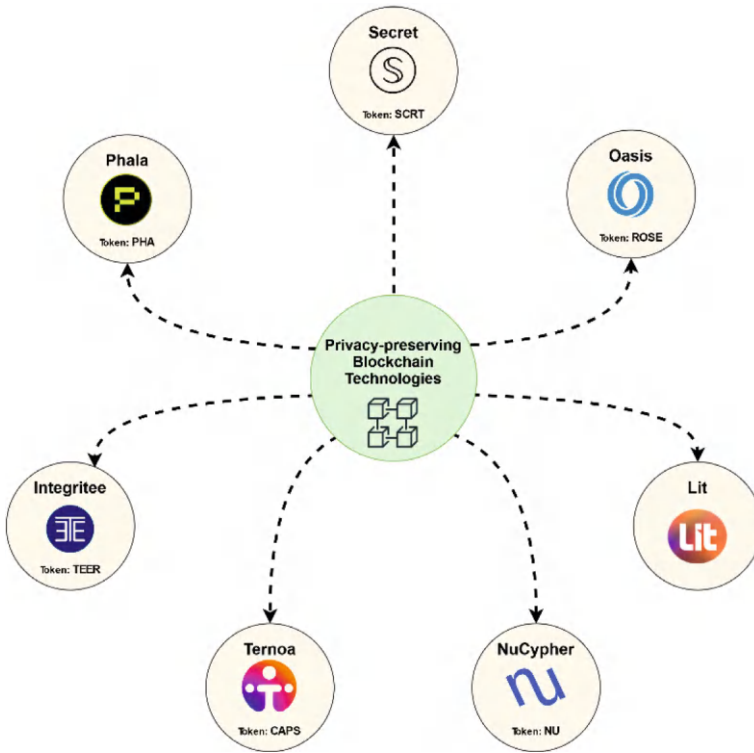


Fig. 3 Privacy preserving blockchains

4.1 Zero-Knowledge Systems

Zero-knowledge proof (ZKP) systems offer promising solutions for maintaining transaction privacy while ensuring system verifiability. ZKPs enable one party (the prover) to convince another party (the verifier) that a statement is true without revealing any information beyond the statement's veracity. This can be mathematically represented as:

$$ZKP = \{(P, V), x | P \rightarrow V : P \text{ knows } \omega, V \text{ verifies}(x, \omega) \in R\}$$

where:

- P represents the prover
- V represents the verifier
- x is the public input
- w is the private witness
- R is the relation being proved.

The properties of ZKP include completeness, soundness, and zero-knowledge. Completeness ensures that an honest prover can convince the verifier if the statement is true. Soundness ensures that no prover can convince the verifier if the statement is false. Zero-knowledge ensures that the verifier learns nothing beyond the validity of the statement. Various advanced forms of ZKPs offer enhanced efficiency and security [11].

4.2 Identity and Access Management

Effective identity and access management is crucial for controlling access to sensitive data and resources on the Blockchain. There is a need to balance privacy requirements with regulatory compliance while implementing a robust identity and access management system. Some of the regulatory compliances include:

- Role-based access control
- Multi-factor authentication
- Privacy-preserving identity verification
- Decentralized identity management.

Role-based access Control (RBAC) restricts access to sensitive data and operations based on pre-defined roles and permissions. MFA (multi-factor authentication) uses two or more methods to enhance security. It is done by providing multiple verification forms before granting access to required users. Users can verify their identities without revealing sensitive information through privacy-preserving identity verification. Decentralized Identity Management (DID) is a secure and user-centric method for providing users with self-sovereign identities. With this, they can control and manage independently.

4.3 Data Protection and Compliance

Data protection and compliance mechanisms are essential for adhering to privacy regulations and protecting sensitive information on the Blockchain. Ring signatures and privacy-focused consensus mechanisms are advanced solutions aimed at preserving privacy. Ideally, these solutions would alleviate challenges in data exposure risk and regulatory compliance but still retain the essence of blockchain technology.

Privacy frameworks and protection mechanisms are essential for enabling secure and trustworthy blockchain applications. Zero-knowledge systems, identity and access management, and data protection and compliance are critical components of a comprehensive privacy strategy for blockchain networks. By implementing these mechanisms, blockchain systems can balance transparency and confidentiality, enabling new and innovative applications while protecting user privacy [12].

4.4 Hash Function

A hash function is a mathematical formula used in blockchain technology that provides a way to link all blocks of the Blockchain together. At the block level, the hash of the previous Block–1 header is stored in Block i. In a block, there are multiple transactions. Blockchain also hashes every transaction using the hash functions that have the following properties:

1. **Fixed-size output:** It is possible to take anything as an input, and the Hash function creates an output with a fixed size. So, blockchains use different hash functions to compact messages for digital signatures.
2. **Collision resistance:** It is computationally infeasible to produce the same output from distinct inputs.
3. **Significant change:** If a single bit changes the input, it will result in an entirely different output.
4. **Preimage resistance:** It is not feasible to randomly input the data into the hash function until the same output is produced.
5. **Second preimage resistance:** If an input and its hash output are given, it is computationally infeasible to get the second input that produces the same hash output.

5 Advanced Security Solutions and Future Considerations

The world of security is constantly evolving, and due to this, there are many updates in the blockchain security sector. These updates demand further innovations in security measures—for example, fraud detection, healthcare, etc. With updates in the blockchain security sector, new developments such as artificial intelligence enhanced with anomaly detection, quantum secure encryption algorithms, and cross-chain security frameworks exist. These implementations allow adaptive responses to new challenges through advanced access control systems and dynamic security architectures.

5.1 Artificial Intelligence (AI) Enhanced Anomaly Detection

Blockchain technology and AI integration play a significant role in enhancing security. AI systems can analyze vast amounts of network data to detect cloud anomalies, such as odd patterns and other malicious activity. Such systems are capable of finding the problems before they occur. Such detections can be improved through AI. AI can watch everything happening and analyze transaction patterns, user behavior, and innovative contract interactions. This allows Blockchain to identify and respond to potential attacks and fake activities. Such detections can mitigate Distributed Denial

of Service (DDoS) attacks (sudden flood of transactions) and Sybil attacks (creating tons of fake accounts). This can be stopped by continuous monitoring and learning from network activities. Further, AI systems get better at spotting new attack patterns and issuing real-time alerts, enhancing the overall security carriage of blockchain networks. Machine Learning techniques, such as reinforcement learning, mitigate adversarial attacks in PoS blockchains. Even Deep-learning techniques are also used to detect and prevent cryptographic key compromises in blockchain networks [13–16].

5.2 Quantum-Safe Encryption Algorithms

Quantum-safe encryption algorithms help protect against super-powerful computers. It could break the encryption that keeps Blockchain safe. With the arrival of quantum computing, a significant threat to the security of blockchain systems is posed. Various quantum algorithms can decrypt symmetric and asymmetric encryption techniques that ensure system integrity and security. The quantum-safe algorithms are also known as post-quantum cryptography. These are designed to be resilient to attacks from both classical and quantum computers. Lattice-based cryptography, code-based cryptography, and hash-based cryptography are examples of quantum-safe encryption algorithms. Integrating these algorithms into blockchain systems will ensure they remain secure despite quantum computing advancements [17].

5.3 Cross-Chain Security Frameworks

Different blockchain networks have to talk to each other. It is essential to keep blockchains safe, and a Cross-chain security framework helps in that situation. Cross-chain transactions and interactions introduce new security challenges. Developing robust cross-chain security frameworks is essential for ensuring the integrity and confidentiality of cross-chain transactions. These frameworks should address various issues, like Cross-chain authentication, data integrity verification, and secure communication protocols. Another approach is a cross-chain security framework involving blockchain relays and hash-lock-based mechanisms that enhance the trustworthiness of interaction between multiple blockchains [6].

5.4 Scalability Versus Security Trade-Offs

This section discusses the comparison between scalability and security trade-offs. Blockchain scalability makes today's world faster and raises some critical challenges for blockchain networks. Because of Scalability, the number of users is increasing,

and the number of transactions also increases. Due to this, the network's performance can degrade, which leads to delays and increased transaction fees. The solution to this scalability challenge requires careful consideration of the trade-offs between Scalability and security. One of the various solutions is Layer-2 scaling solutions (helping the highway). This includes state channels, plasma, and rollups, which offer promising approaches for improving Scalability without sacrificing security. Another solution is dividing the highway, known as Sharding. Sharding involves dividing the blockchain network into smaller, more manageable shards, each responsible for processing a subset of transactions. The strategy of sharding schemes must sensibly consider security consequences to prevent attacks on individual shards. Cloud platforms store the information so that the proof of the data sharing can be recorded in the Blockchain for auditing and tracing to know who did what [18].

5.5 Privacy Policies and Regulatory Changes

There are various privacy policies and regulations to keep the data safe and to follow laws. The regulatory landscape surrounding blockchain technology is constantly evolving. Various new rules, such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), impose stricter data protection and privacy requirements. While Blockchain maintains its core principles of decentralization and transparency, it should comply with these regulations. Planning effective privacy policies is essential to balance security concerns, system performance, and regulatory changes for the long-term sustainability of blockchain technology [19, 20].

The future of blockchain security lies in developing and integrating advanced security solutions that can adapt to evolving threats and technological advancements. With the acceptance of AI-enhanced anomaly detection, quantum-safe encryption algorithms, cross-chain security frameworks, and privacy policies, blockchain networks can continue to provide secure, scalable, and privacy-preserving solutions for an extensive range of applications. Continuous research and development are essential to ensure blockchain technology remains at the forefront of innovation and maintains its position as a trusted and secure platform.

With the expansion of blockchain networks, the issue of Scalability versus that of security has to be addressed with great care. The transition to the quantum era necessitates the construction of quantum-resistant algorithms and hybrid security architectures. Lastly, devising effective privacy policies must seek to optimize other security concerns, system performance, and market regulatory changes. Code obfuscation, homomorphism encryption, trusted executing platforms, and smart contracts for privacy preservation would be promising directions.

6 Conclusion

The search for privacy and security in blockchain networks discloses a complex and continuously changing environment. The study of smart contracts, consensus processes, and network layer vulnerabilities highlights the difficulties in protective decentralized systems. Even if defenses like Isolated Witness, DDoS mitigation techniques, and sophisticated cryptographic protocols provide helpful security, they must be modified to keep up with new attack methods and the growing complexity of bad actors.

The discussion of privacy frameworks highlights how important it is to find solutions that compromise secrecy and transparency. Strong identity and access control procedures, zero-knowledge systems, and compliance with data protection laws are essential elements of an all-inclusive privacy strategy. Ongoing efforts are necessary to match technical progress with changing legal and ethical norms, nevertheless, because the regulatory environment around blockchain technology is still dynamic.

The concerns regarding security and privacy in blockchain systems require constant development and never-ending progress. The suggested answers and arguments may formulate a starting point for constructing more dependable and encompassing tools for blockchain systems. Future research will deal with how the best performance will be achieved amidst the various security requirements increasing with time and the development of new rules and regulations. Attacks of the future will demand the integration of new secure cryptographic algorithms, intelligent systems, and algorithms resistant to quantum techniques.

The future of blockchain security and privacy will emerge only with the realization of the measures mentioned above and change whenever appropriate in dealing with demands for new challenges. Such a developing scenario calls for constant effort in research, development, and interaction in the blockchain space to ensure the technology remains what it should be: secure, scalable, and privacy-centric distributed systems. Continuous efforts in research and development are crucial to ensuring Blockchain remains a safe, scalable, and privacy-preserving distributed ledger system.

References

1. Mohanta, B.K., Jena, D., Ramasubbareddy, S., Daneshmand, M., Gandomi, A.H.: Addressing security and privacy issues of IoT using blockchain technology. *IEEE Internet of Things J.* 11 (2020)
2. Krishna (2024). <https://www.getastra.com/blog/base/blockchain-security-issues/>
3. Zubaydi, H.D., Varga, P., Molnár, S.: Leveraging blockchain technology for ensuring security and privacy aspects in internet of things: a systematic literature review. *Sensors* **23**(2), 788788 (2023)
4. Alfandi, O., Khanji, S., Ahmad, L., Khattak, A.: A survey on boosting IoT security and privacy through blockchain. *Clust. Comput.* **24** (2020)

5. Xihua, Z., Goyal, D.S.B.: Security and privacy challenges using IoT-blockchain technology in a smart city: critical analysis. *Int. J. Electr. Electron. Res.* **10**(2), 190–195 (2022)
6. Li, N., Qi, M., Xu, Z., Zhu, X., Zhou, W., Wen, S., Xiang, Y.: Blockchain cross-chain bridge security: challenges, solutions, and future outlook. *Distrib. Ledger Technol.* **4**(1), Article 8, 34pp. (2025). <https://doi.org/10.1145/3696429>
7. Shah, Thakkar, V., Khang, A., pp. 1–13 (2023). <https://doi.org/10.1201/9781003356189-1>
8. Patil, P., Sangeetha, M., Bhaskar, V. (2020). <https://doi.org/10.1007/s11277-020-07947-2>
9. Whig, P., Velu, A., Nadikattu, R.R.. <https://www.igiglobal.com/chapter/blockchain-platform-to-resolve-security-issues-in-IoT-and-smart-networks/306880>
10. Valadares, D.C.G., Perkusich, A., Martins, A.F., Kamel, M.B.M., Seline, C.: Privacy-preserving blockchain technologies. *Sensors* **23**, 7172 (2023). <https://doi.org/10.3390/s23167172>
11. Zhou, L., Diro, A., Saini, A., Kaisar, S., Hiep, P.C.: Leveraging zero-knowledge proofs for blockchain-based identity sharing: a survey of advancements, challenges, and opportunities. *J. Inf. Secur. Appl.* **80**, 103678 (2024). <https://doi.org/10.1016/j.jisa.2023.103678>, ISSN 2214-2126
12. Sargiotis, D.: Data security and privacy: protecting sensitive information. In: *Data Governance*. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-67268-2_6
13. Edozie, E., Shuaibu, A.N., Sadiq, B.O., et al.: Artificial intelligence advances in anomaly detection for telecom networks. *Artif. Intell. Rev.* **58**, 100 (2025). <https://doi.org/10.1007/s10462-025-11108-x>
14. Pattnaik, L.M., Swain, P.K., Satpathy, S., Panda, A.N.: Cloud DDoS attack detection model with data fusion & machine learning classifiers. *EAI Endorsed Trans. Scalable Inf. Syst.* **10**(6) (2023)
15. Chatterjee, S., Satpathy, S., Paikaray, B.K.: Forecasting DDoS attack with machine learning for network forensic investigation. *Int. J. Reason.-Based Intell. Syst.* **16**(5), 352–359 (2024)
16. Satpathy, S., Swain, P.K., Mohanty, S.N., Basa, S.S.: Enhancing security: federated learning against man-in-the-middle threats with gradient boosting machines and LSTM. In: *2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–8, July 2024. IEEE (2024)
17. Kumar, M.: Post-quantum cryptography algorithm’s standardization and performance analysis. *Array* **15**, 100242 (2022). <https://doi.org/10.1016/j.array.2022.100242>, <https://www.sciencedirect.com/science/article/pii/S2590005622000777>, ISSN 2590-0056
18. Saqib, N.A., AL-Talla, S.T.: Scaling up security and efficiency in financial transactions and blockchain systems. *J. Sens. Actuator Netw.* **12**(2), 31 (2023). <https://doi.org/10.3390/jsan12020031>
19. Satpathy, S., Mahapatra, S., Singh, A.: Fusion of blockchain technology with 5G: a symmetric beginning. In: Tanwar, S. (eds.) *Blockchain for 5G-Enabled IoT*. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67490-8_3
20. Tikkinen-Piri, C., Rohunen, A., Markkula, J.: EU general data protection regulation: changes and implications for personal data collecting companies. *Comput. Law & Secur. Rev.* **34**(1), 134–153 (2018). <https://doi.org/10.1016/j.clsr.2017.05.015>, <https://www.sciencedirect.com/science/article/pii/S0267364917301966>, ISSN 0267-3649

Revolutionizing Academic Record Management: A Blockchain-Driven Solution for Secure and Transparent Student Documents



Swapnil S. Chaudhari , Vaishnavi Vikhe, Dipanjali Bhujbal, Shreyash Pangarkar, and Gokul Arya

Abstract The growing reliance on digital archives for academic records shows critical contests connected to data security, genuineness, and confirmation effectiveness. Old-fashioned consolidated systems are exposed to data breaches, manuscript damage, and incompetence in the substantiation process. This paper recommends a blockchain-based student document organization scheme to discourse these issues by leveraging reorganized skills. In this system, pupils can upload, manage, and share academic proceedings such as transcriptions and documentation with complete control over their data. Superintendents can validate and bring up-to-date records, ensuring data truthfulness and uprightness. At the same time, peripheral verifiers, such as establishments or educational institutions, can securely endorse documents deprived of the need for intercessors.

Keywords Blockchain · Smart contracts · Academic records · Student document management · MetaMask · Ethereum · Transparency · Decentralization

S. S. Chaudhari (✉) · V. Vikhe · D. Bhujbal · S. Pangarkar · G. Arya
Department of Computer Engineering, Marathwada Mitra Mandal's Institute of Technology Pune,
Pune, India
e-mail: swapnil.chaudhari@mmit.edu.in

V. Vikhe
e-mail: vaishnavi.vikhe@mmit.edu.in

D. Bhujbal
e-mail: dipanjali.bhujbal@mmit.edu.in

S. Pangarkar
e-mail: Shreyash.pangarkar@mmit.edu.in

G. Arya
e-mail: gokul.arya@mmit.edu.in

1 Introduction

The quick shift in the direction of digital records in the scholastic sector has underlined the need for a protected and effectual system for handling academic records. Outdated document controlling schemes, often conditional on centralized databases, are defenseless to issues such as data cracks, forgery, and unofficial access. These contests clear the truthfulness and faithfulness of crucial academic evidence. This paper offers a blockchain-based student document administration system that authorizes pupils to maintain complete control over their papers, including transcripts and certificates. By utilizing regionalized technology, the system confirms data exactness and immutability, addressing the insistent defies of article tampering.

Additionally, it modernizes the authentication process for officers and verifiers, such as probable employers or educational institutions, by empowering secure admittance to records without intercessors. Incorporating smart contracts on the Ethereum blockchain develops record administration's inclusive security and proficiency. With a user-friendly boundary that combines modern front-end machinery, the system is premeditated to be reachable for all stakeholders. This explanation expands transparency and fosters user confidence, making it a robust unconventional for supervising academic papers in a more and more digital background.

2 Literature Survey

Blockchain technology [1] meaningfully develops digital identity supervision by providing regionalized control and confirming data reliability. This paper evaluates 63 relevant education to scrutinize the state-of-the-art explanations in blockchain-based identity controlling systems. This inclusive review shows that Blockchain's integral geographies, such as immutability and disseminated ledger knowledge, offer auspicious alternatives to outmoded compacted systems. However, the research also makes known gaps, predominantly regarding scalability and security, which dictate further investigation to fully force Blockchain's potential in universal digital identity networks.

This research work in [2] examines the key recompenses and obstacles connected with reorganized identity-controlling systems (RICS). These classifications encourage enhanced concealment and condense dependence on centralized specialists by offering users greater control over their particular documents. However, the scholarship also accentuates the privacy anxieties that arise when supervising digital personalities on community blockchains, even though regionalized systems can diminish the risk of unsanctioned access and openings, safeguarding wide-ranging privacy and obedience with surfacing data guidelines left over an experiment. Future developments should concentrate on decontaminating encryption performances and user authentication appliances to overwhelm these issues and unlock the probability of regionalized identity explanations.

The study in [3] offers everyday case revisions where Blockchain has been positively functional in verifying hypothetical authorizations and managing enlightening documents. These real-world employments establish Blockchain's capability to generate tamper-proof, crystal clear, and practical arrangements for handling penetrating academic data. The conclusions clarify that Blockchain can rationalize the confirmation process for speculative credentials, tumbling the need for intercessors and safeguarding the truthfulness of documents. However, the conclusions also highlight certain restrictions, such as the requirement for homogeneous structures and guidelines across dissimilar informative institutions to expedite the wider embracing of Blockchain in the educational division.

3 Methodology

The Blockchain-Based Individuality and Manuscript Management Structure for Educational Associations is a ground-breaking solution to leverage blockchain machinery's strong point, ensuring protected, scalable, and user-friendly surroundings for management theoretical proceedings. The proposal development is widespread and accurately premeditated to cater to the assorted needs of scholars, administrators, and verifiers on the condition that an all-inclusive attitude to individuality and document control lacks the need for an outdated backend system [4–7].

It commences with the enlargement of the front-end boundary using React, a powerful JavaScript archive known for generating communicating user involvements [8–10]. The front end will attend to multiple user characters, including students, administrators, and verifiers (such as establishments), each with custom-made functionalities. Students will devise the capacity to upload, manage, and share their speculative booklets effortlessly. Superintendents will be tasked with proving, updating, and upholding these accounts, ensuring accuracy and truthfulness. Verifiers, including potential proprietors, can firmly access verified archives, fostering expectation and transparency in engagement development.

A foundation of this scheme is the incorporation of MetaMask, a Web3 holder that empowers users to substantiate and cooperate with the Blockchain immediately after their browsers. This excludes the out-of-date login necessity of usernames and secret codes, rearranging user familiarity, and improving sanctuary. Instead of keyboarding personal qualifications, users can log in to MetaMask, which uses their Ethereum holder for verification. This attitude simplifies login development and makes a complex sanctuary level available, as sensitive evidence is not deposited on consolidated servers [11–14].

On one occasion, the front end was industrialized, and the focus was on modifications to produce innovative agreements using Solidity, the principal user interface design language for Ethereum. These clever conventions will administrate the core functionalities of the classification, such as distinctiveness substantiation, manuscript issuance, and access control. For occurrence, when an intellectual uploads

an academic document, the structure produces a cryptographic hash of that paper, which is formerly stored on the Blockchain. This confirms that the manuscript's truthfulness is well-maintained and supportable, while the authentic document is deposited off-chain in distributed storage explanations like IPFS (InterPlanetary File System). By employing IPFS for manuscript storing, the system equilibrums effectiveness and cost-effectiveness, as loading large files straight on the Blockchain can be individually luxurious and unreasonable.

Outstandingly, the complete construction of this organization is calculated to mean without any outmoded backend substructure. All indispensable procedures, including manuscript substantiation and identity management, are finalized on-chain. There are no cohesive databanks or servers, disregarding single points of failure and minimizing vulnerabilities typically associated with unadventurous systems. This structural design authorizes students, administrators, and verifiers to intermingle unswervingly with the Blockchain, resulting in a structure that is not only more protected but also naive and more mountable.

Security residues are a dominant concern throughout the putting-into-practice process. The use of blockchain technology fundamentally defends data from damage and deception. Every operation, including manuscript uploads and substantiations, is documented on the Blockchain, creating an absolute and translucent archive of all communications. Furthermore, vigorous encryption techniques such as AES (Advanced Encryption Standard) are laboring to protect off-chain data, whereas RSA (Rivest-Shamir-Adleman) encryption is employed to supervise private key procedures. These layers of safekeeping ensure that both on-chain and off-chain data remain safe and trustworthy.

After the insolent agreements and front-end mechanisms are enlarged, demanding analysis is accompanied to safeguard system uprightness. Tools such as Truffle and Hardhat are used to test the insolent agreements in a controlled expansion atmosphere. This systematic analysis process is crucial for recognizing and determining any susceptibilities or incompetence before deploying the agreements to a community Ethereum test net like Rinkeby. Once the sureness of the smart contracts is recognized, they will be set up to the Ethereum mainnet or a Layer-2 solution like Polygon to develop scalability and decrease contract costs. The use of Layer-2 clarifications is predominantly valuable for educational establishments, where large capacities of transactions may happen.

User acceptance testing (UAT) is a precarious phase where real users—students, administrators, and verifiers—interrelate with the organization to estimate its functionality, usability, and global presentation. Response from UAT is irreplaceable, as it services the expansion team by making compulsory modifications to increase user involvement and discourse on any issues that may arise. Following UAT, the classification will arrive at a phased distribution process, opening with a pilot package. This pilot will be directed in a controlled atmosphere, such as a detailed subdivision or between a small group of operators, permitting further calculation of system presentation and documentation of any outstanding subjects.

As the pilot accomplishes this positively, the full-scale disposition will take dwelling, creating a reachable system across the instructive establishment. The clever

conventions will be live on the Ethereum mainnet or a Layer-2 explanation, safeguarding optimal presentation and scalability. The React front-end determination must be firmly hosted on a dependable policy to guarantee incessant admission for all users. Wide-ranging user credentials will be provided to help pupils, administrators, and verifiers route the scheme professionally. In addition, exercise sessions, factories, and webinars will be prearranged to publicize users with the policy's features and functionalities, safeguarding that all stakeholders are well-equipped to exploit the system efficiently.

The scheme will be carefully watched post-disposition to ensure horizontal procedure and proper discourse on any possible issues. A helpdesk establishment organization will be recognized as long as users have possessions and support for troubleshooting any practical complications they may meet. Regular information and developments will be finished in the system based on user feedback and scientific progressions, safeguarding that it remains pertinent and operative in conference with the requirements of instructive organizations.

The Blockchain-Based Individuality and Article Administration System for Informative Organizations signifies a transformative method for handling theoretical records. By fully implementing regionalized knowledge, the system safeguards high safekeeping, data truthfulness, and limpidity while authorizing students to have widespread control over their hypothetical documents. The non-appearance of outmoded backend organization streamlines the construction, improves scalability, and condenses the risk of information gaps. This wide-ranging explanation not only happens to the instantaneous needs of enlightening establishments but also sets an instance for future revolutions in identity and manuscript management. By encouraging trust and pellucidity among all participants, this system overlays the way for a more protected and competent attitude to the supervision of hypothetical records, initially causative to the development of educational knowledge.

4 Proposed System

This segment summarizes the projected blockchain-based scheme for protected and regionalized management of hypothetical documents. The classification is ingenious to safeguard data truthfulness, limpidity, and sanctuary by utilizing blockchain knowledge to store, accomplish, and verify speculative records. The construction introduces roles for Superintendents, Students, and Verifiers, each underwriting the manuscript lifecycle in distinctive ways.

4.1 *System Design and Architecture*

The recommended coordination construction utilizes blockchain technology to create a decentralized and unchallengeable archive designed to handle hypothetical documents. This pioneering methodology disregards the essentials for integrated control, dispensing trust, and safekeeping across all play-apart nodes in the blockchain network. This confirms that students, administrators, and verifiers in place of academic establishments or establishments can faultlessly interrelate with the system in a way that agreements the truth, pellucidity, and authenticity of all hypothetical records. The immutability of blockchain technology certifies that once archives are added to the structure, they cannot be changed, thus stopping any unsanctioned tinkering with the records.

The key knowledge essential to this organization is MetaMask, a Web3 folder that serves as the doorway through which users securely interrelate with the Blockchain. MetaMask succeeds in public and private keys, confirming that only sanctioned users, such as students, administrators, and verifiers, can admit and adjust the data. It safeguards that each action in the system's interior is intensely employed and sanctioned, conserving the classification from unsanctioned access or employment of hypothetical records. Another decisive aspect of this structure is smooth conventions, which are self-executing conventions with the relationships of the promise directly surrounded in code. Innovative agreements industrialize many progressions, such as document substantiation and certificate issuance, permitting communications without disinterested parties. For occurrence, once a student uploads a manuscript and the system verifies it, a smart contract automatically triggers an update to the substantiation status. This automation streamlines the process, reduces manual intervention, and minimizes the risk of human error.

In terms of data sanctuary, the organization engages in distributed storage, allowing theoretical documents like documentation and transcriptions to be disseminated across numerous nodules rather than stockpiled on a single server. This dispersed methodology makes available redundancy, guaranteeing the information is always obtainable, even if approximately nodules go offline. Keeping an orientation or hash of the manuscript on the Blockchain guarantees the manuscript leftovers are irreversible and can be unaffected at any time without cooperating data security [5].

The system construction, as shown in Fig. 1, defines three core user roles: administrator, Verifier, and Student, each with different tasks that underwrite the inclusive security and functionality of the structure. The Superintendent accomplishes the system by adding verifiers, issuing administration certificates, and supervising student booklets. They also keep posted on the verification position of histories and have the expert delete old-fashioned documents. The Verifier, such as a speculative establishment or employer, authenticates documents by examining their authenticity against the hash stored on the Blockchain. Once substantiated, they can also bring up-to-date information about the document's verification position, adding trust. The Student preserves their academic archives and can upload, understand, and manage

booklet access. They interrelate with the Blockchain concluded MetaMask, ensuring protected and sanctioned actions and control over who can verify their archives.

As per Fig. 1, the core occupations of the arrangement include numerous essential routes. Uploading booklets allows students to upload academic proceedings to decentralized storage via MetaMask strongly. Each document is dispensed an exceptional hash, ensuring its immutability. Both scholars and verifiers can interpret documents through a protected crossing point. Students have control over accomplishing academic records and protecting their information, which is systematized and up-to-date. All communications are powerfully touched through MetaMask, confirming authorized access. Manuscript substantiation is a key feature, tolerating verifiers to

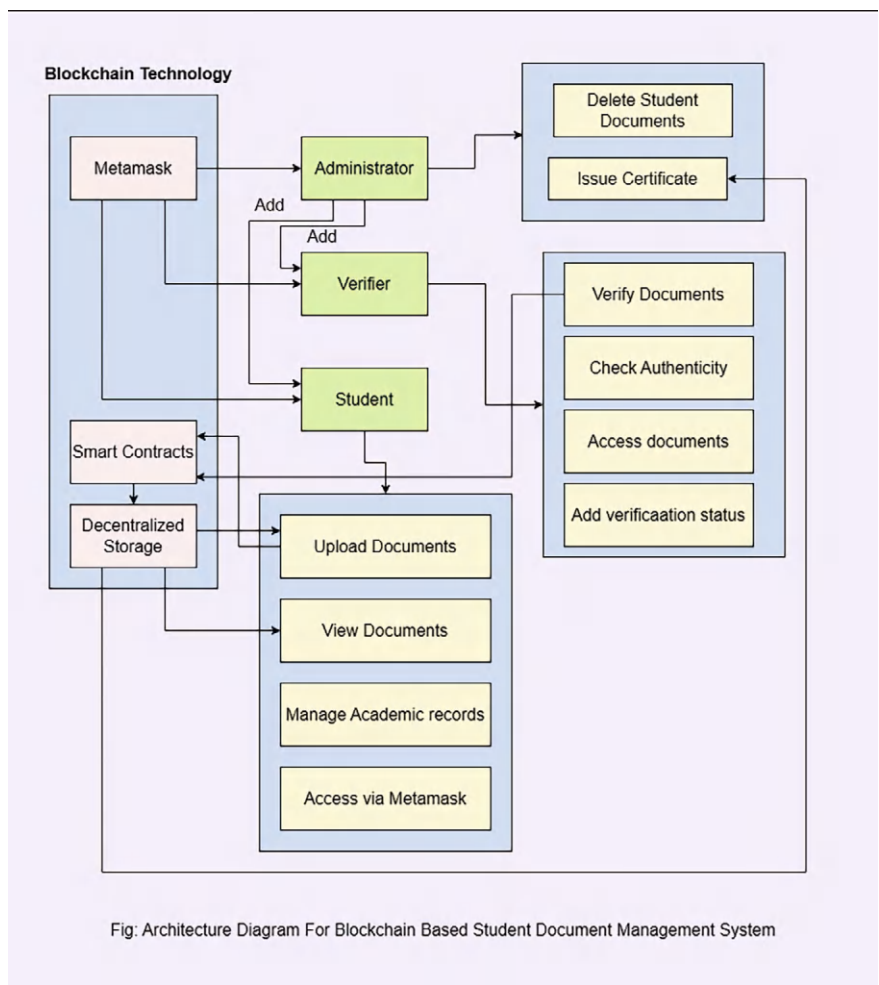


Fig. 1 System architecture of student document management system

authenticate accounts through the blockchain hash. Certificate issuance is succeeded by administrators, certifying that recently dispensed certificates are mechanically certifiable on the Blockchain. Administrators also can obliterate student documents to keep records up-to-date. Document truthfulness instructions allow students and verifiers to document the blockchain-stored hash with the authentic document to confirm it hasn't been damaged. As a final point, verifiers can add authentication prestige to a document after sanctioning its authenticity, further warranting its trustworthiness.

This blockchain-based organization propositions quite a few advantages over outmoded centralized educational record administration. Decentralization and impatience remove the need for a central specialist, dispensing trust diagonally in the network and increasing safekeeping. The system's immutability confirms that once data is stockpiled, it cannot be rehabilitated or canceled, stopping tampering. Safekeeping is guaranteed through cryptographic apparatuses and reorganized storage safeguards that data remains available even if some nodules go offline. The system also makes available pellucidity, as all communications and updates are documented on the Blockchain, fashioning a supportable audit stream. Ownership of data invests students, giving them complete control over their academic proceedings deprived of relying on intercessors. Using insolent conventions improves productivity by programming processes such as substantiation and documentation issuance, tumbling the need for manual interpolation and lessening errors. Lastly, the system's struggle against fraud ensures that academic proceedings are threatened by forgery, as the Blockchain makes available a trusted and irreversible source of actuality.

This construction makes available a robust, secure, and efficient resolution for treating academic records using blockchain equipment. By combining MetaMask, innovative agreements, regionalized storage, and Blockchain's immutability, the system safeguards hypothetical documents that are secure, translucent, and easily supportable. It invests students by open-handed proprietorship of their data, while submission verifiers are an essential method for authenticating academic records. This architecture is predominantly proper for academic organizations seeking to digitize their processes while confirming the reliability and truthfulness of their authorizations.

4.2 Document Workflow

The Blockchain-Based Document Controlling Arrangement empowers a secure and apparent process for treatment and corroborating academic brochures. The progression is instigated when a student cooperates with the system using MetaMask, a Web3 holder. The scholar chooses a manuscript (such as a certificate or diploma) for upload, and the scheme produces an exceptional hash that acts as an alphanumeric thumbprint for that manuscript. While the document is warehoused in decentralized storage, its hash and proprietorship metadata are stored immutably on the Blockchain. This ensures that any effort to modify the manuscript will be demonstrable, as the

hash will no longer match. Furthermore, an orientation link to the manuscript is created for forthcoming access.

Once the manuscript is uploaded, the Student can bring about it through a user-friendly crossing point, establishing, observing, or scrubbing documents as needed. Momentously, the Student controls admission agreements, deciding who, such as possible verifiers (e.g., employers or academic organizations), can admittance the manuscript. These agreements are administrated by clever conventions, ensuring that only sanctioned users can outlook or verify the manuscript [9].

When a verifier needs contact with a manuscript, they first gain acquiescence from the Student. The Verifier can formerly admit the document and its conforming hash to the Blockchain to authorize its faithfulness. This process ensures that the manuscript has not been damaged by any incongruities amid the blockchain-stored hash and that the manuscript will proximately indicate falsification. After fruitful corroboration, the verifier proceedings the document's authentication status on the Blockchain finished a smart contract, and the result was immutable and apparent. In cases someplace a document fails substantiation, it is flagged as "not genuine," and an alert is prompted [8, 10].

The system continues to ensure data truthfulness and safekeeping at every stage. All communications—uploading, supervision, or verifying booklets—are logged immutably on the Blockchain with timestamps, cryptographic hashes, and user particulars. Regionalized storage certifies that booklets remain reachable even if some link swellings go offline. Furthermore, encrypted passages and MetaMask combination agree that all contact is safe, sound, and sanctioned.

The arrangement also makes available organizational tasks for supervision verifiers and supervising verification processes. Superintendents can intercede in exceptional cases, such as recirculating certificates or treatment records for defunct students. They can also delete old-fashioned or inaccurate records, certifying that the blockchain location is reorganized to continue current and truthful records. Figure 2 shows the DataFlow Diagram of the Student Document Management System.

4.3 Use Case Diagram

See Fig. 3. The Use Case Diagram summarizes the exchanges between key actors and the blockchain-based student document organization system. The principal performers include the Student, Verifier, Superintendent, Smart Convention, and IPFS (InterPlanetary File System), each performing a decisive role in confirming sheltered and competent document supervision. The Student is accountable for uploading hypothetical brochures, managing access authorizations, and recovering stored files. Acknowledged by their Ethereum holder address, scholars can succumb files, store metadata, track substantiation status, and govern manuscript access [5].

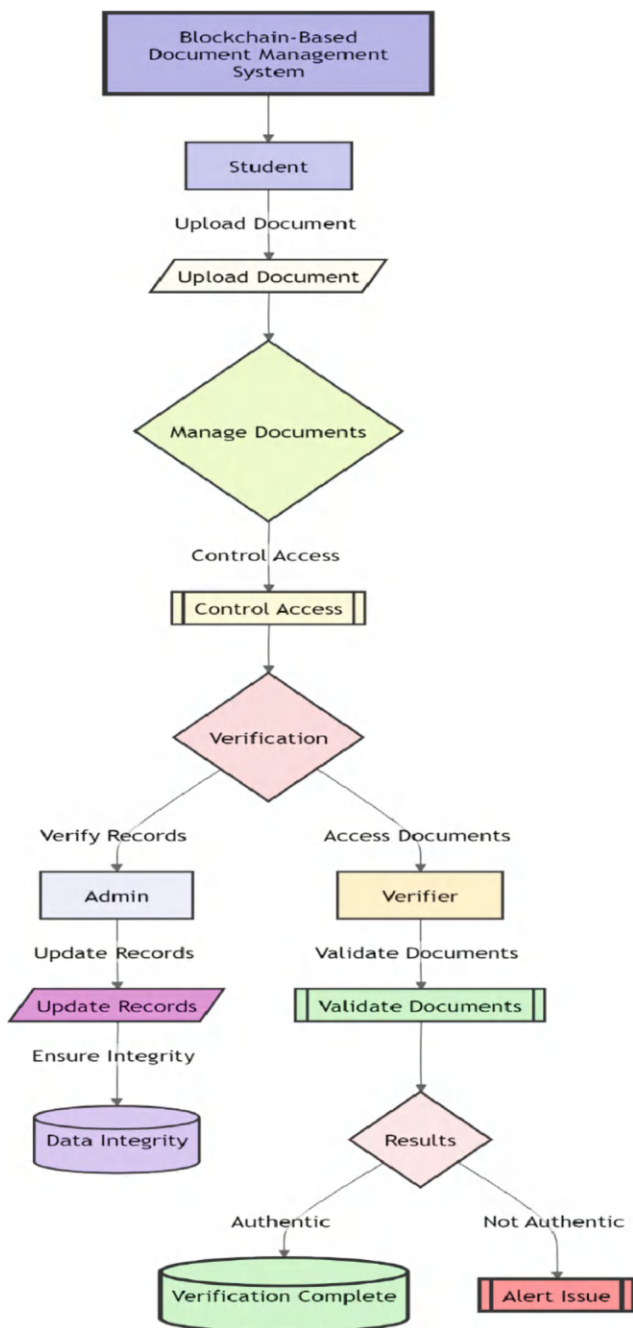


Fig. 2 DataFlow diagram of student document management system

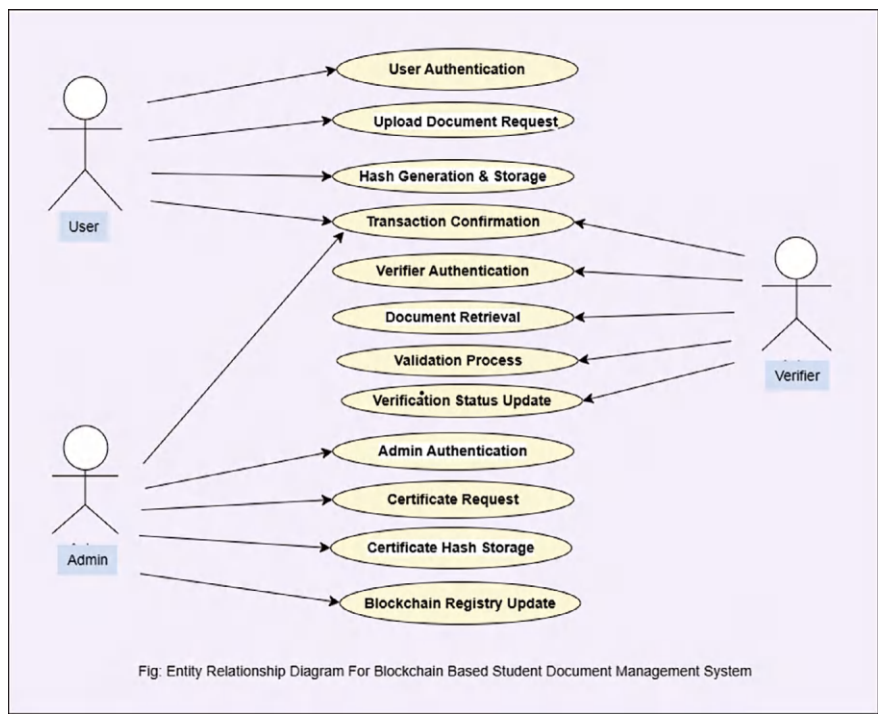


Fig. 3 Use case diagram of student document management system

The Verifier, which consists of employers and informative establishments, intermingles with the system to authenticate hypothetical official papers. Through the Document Substantiation use case, verifiers can apply admittance, confirm truthfulness, and confirm the proprietorship of student proceedings.

The Superintendent plays a fundamental role in administration system procedures by managing verifiers, make-up-your-mind disputes, and ensuring flat platform functionality. Their key accountabilities comprise Verifier Approval, System Monitoring, and Access Administration, which benefit continued platform truthfulness [6]. Smart Indentures systematize essential processes such as certificate verification, access regulator, and operation logging, guaranteeing that proprietorship, agreements, and substantiation results are powerfully recorded on the Blockchain. In addition, IPFS facilitates decentralized manuscript storage by connecting files to unique cryptographic hashes. The Document Storage and Reclamation use case consents students and verifiers to store strong and admittance academic proceedings off-chain while maintaining blockchain-backed reliability [7].

By demonstrating these exchanges, the Use Case Diagram highlights how changed actors cooperate to enhance safekeeping, pellucidity, and efficiency in academic manuscript administration. It serves as an introductory guide for significant user roles,

system functionality, and the carrying out of smart contracts to create a reorganized and tamper-proof documentation substantiation structure.

4.4 Blockchain Integration

By fully assimilating blockchain expertise into the classification construction, the future system eradicates the need for outmoded centralized databanks and backends. Blockchain nodules manage the scattered archive, where every contract related to manuscript supervision is documented. The MetaMask holder acts as the protected authentication instrument, providing continuous interaction between the user and the Blockchain.

Insolent conventions administer business logic, such as manuscript verification proprieties and issuing accreditations. This attitude reduces system convolution, develops security, and warrants that accredited parties can singly reform data, as the clever conventions demarcate.

See Fig. 4. The MMIT Document Administration System delivers a user-friendly interface for protected, decentralized theoretical record supervision. It propositions role-based admittance for Students, Verifiers, and Superintendents, enabling manuscript upload, confirmation, and system misinterpretation. As per MetaMask Account Transactions, Key features include MetaMask Web3 holder incorporation for blockchain communications and a simple triangulation system for ease of use. This interface is attached to the research aim of improving transparency, safeguarding, and efficiency in blockchain academic password authentication.

4.5 Key Benefits of the Proposed System

The blockchain-based hypothetical document administration system, as shown in Fig. 5, familiarizes numerous key assistance:

- **Immutability and Data Integrity:** Once warehoused, academic registers cannot be rehabilitated or tampered with, defending against manuscript fraud.
- **Enhanced Security:** Distributed data storage expressively reduces defenselessness associated with consolidated systems, such as data ruptures and unsanctioned admissions.
- **Efficient Verification:** Verifiers have unswerving access to scholar records on the Blockchain, consenting for instantaneous and secure corroboration of document truthfulness without peacekeeping troops.
- **Auditability and Transparency:** All exchanges with academic proceedings are publicly distinguishable on the Blockchain, which makes it easy to assess document antiquity and ensure responsibility.

Fig. 4 MetaMask account transactions

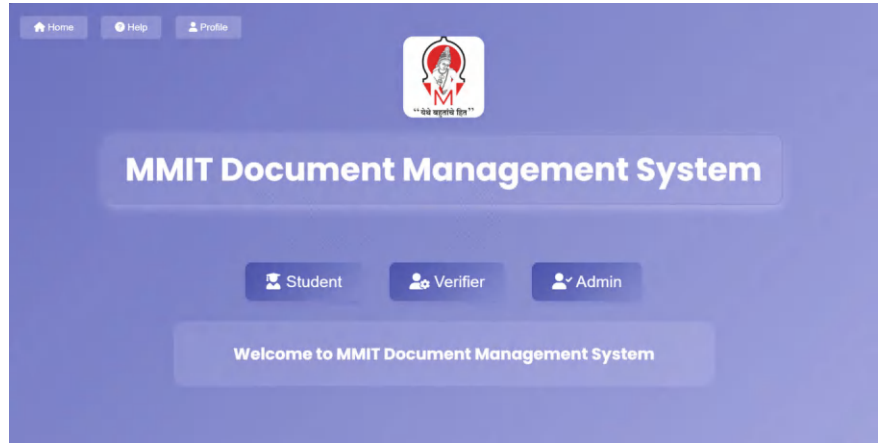
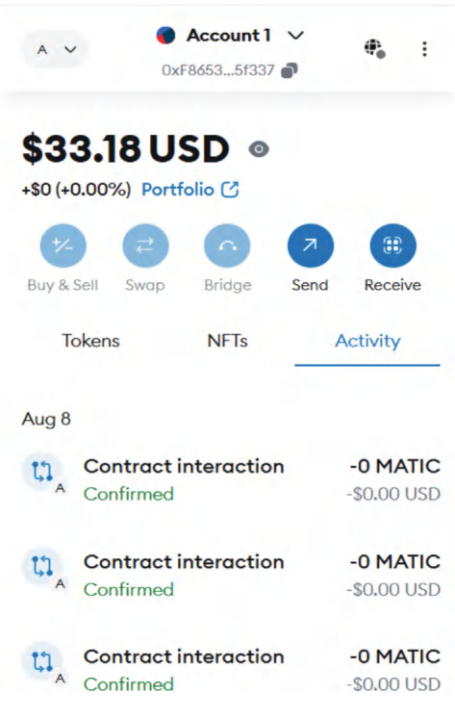


Fig. 5 Home page of the system

- **Simplified Architecture:** By disregarding the need for an outmoded backend arrangement, the system reduces operative involvement and, directly above all, meets on direct collaboration with the Blockchain for all perilous functions. [11]

5 Result

After executing the blockchain-based student article managing system, preliminary analysis has provided promising outcomes, authorizing its efficacy in firmly storing and supervising academic proceedings. The classification was experienced across innumerable circumstances to estimate its functionality and enactment. Users could effortlessly upload their academic booklets, including transcriptions and documentation, without exertion. The instinctual design of the React-based front-end edge is acceptable for horizontal direction finding, enabling operators to without difficulty manage their proceedings without being deprived of facing significant usability experiments.

The confirmation process established a high level of proficiency and pellucidity. Verifiers such as proprietors and hypothetical establishments could unswervingly admittance and verify archives inside the system, disregarding the need for arbitrators. This modernized development innocently reduced the time prerequisite for verification and validation, and subscription was a clear advantage over old-fashioned methods. The immutability feature of blockchain equipment plays a key role in guaranteeing security, as once an article is uploaded to the Blockchain, it becomes unassailable—meaning that it cannot be reformed or obliterated. This effectively decreases the risk of manuscript tinkering or forgery, encouraging all parties' assurance of the educational archives' truthfulness and integrity [11].

User feedback throughout the testing segment was insignificantly positive. Test applicants treasured the user-friendly boundary and the authorization that comes with consuming complete control over their educational records. The capability to manage booklets autonomously, deprived of depend on outmoded organizational techniques, was seen as a momentous assistance. This recommends that the system has durable potential for pervasive adoption in educational establishments. Furthermore, the combination of Blockchain resolves remaining issues interrelated to manuscript administration and builds a background of limpidity and trust for all participants to elaborate in the process.

The auspicious results of this introductory testing designate that the proposed system is everyday and valuable for users, arranging the underpinning for real-world application in educational backgrounds. To further demonstrate its benefits, a reasonable analysis was accompanied between the blockchain-based system and old-fashioned document supervision procedures [12]. Key presentation metrics such as paper upload time, authentication time, safekeeping, and user satisfaction were evaluated. The outcomes showed that the blockchain-based system suggestively outperforms out-of-date methods regarding productivity, security, and user experience.

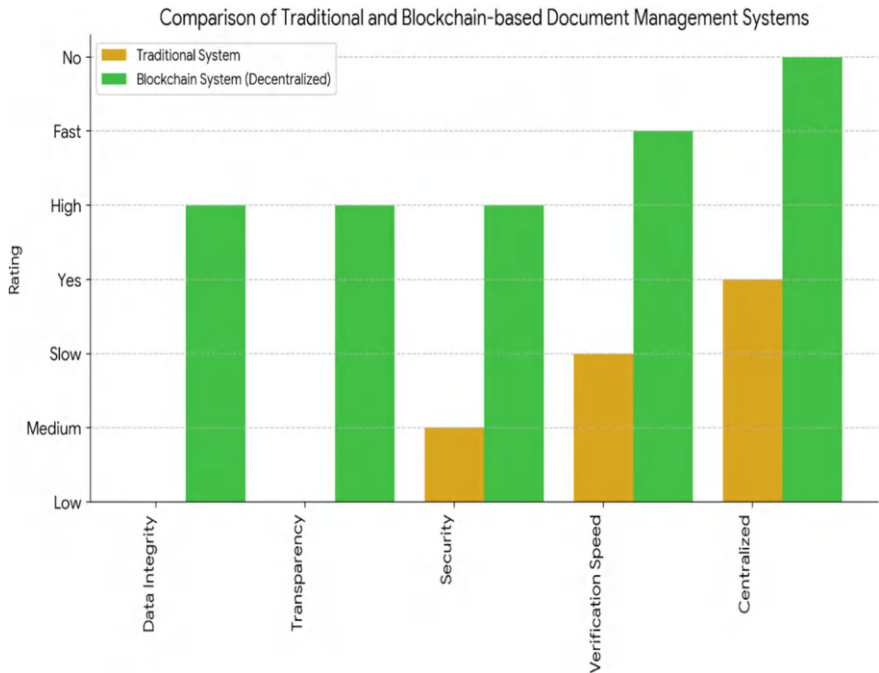


Fig. 6 Statistics with traditional system

For illustration, while the outmoded system involves manual intercession for document substantiation, which is laborious and prone to errors, the blockchain system completes this process in an element of time with enhanced accuracy. Additionally, the risk of documents interfering in traditional systems is moderated by the indisputable nature of blockchain records, certifying unequalled safekeeping [13].

Figure 6 shows the graphical representations of these proportional outcomes, which will be comprised below to afford an apparent visual consideration of the system’s presentation compensations over outmoded methods. Future recapitulations of the structure will concentrate on decontaminating its structures based on user criticism, enlightening scalability, and exploring supplementary capabilities to enhance its functionality and user understanding further.

6 Limitations

Several restrictions must be well-thought-out for practical application. One of the principal challenges is scalability. Public blockchain linkages often face slow procedure processing times and top-to-toe fees as users increase. Also, storing extensive educational archives directly on-chain is unreasonable due to significant storage

costs. Another critical concern is cost constraints. Deploying and continuing a blockchain-based classification can be exclusive, especially when using public blockchains that necessitate gas fees. Informative foundations may vacillate to accept the technology due to the high opening setup and functioning costs accompanying blockchain incorporation.

Moreover, adoption encounters pose a momentous barrier. Foundations, employers, and scholars must integrate coordination into their workflows, imposing training and conviction in blockchain technology. Numerous establishments still rely on old-fashioned document substantiation methods essential to slow embracing rates. Another ultimate limitation is the inability to change or recuperate data once documented on the Blockchain. If inappropriate or fraudulent documents are uploaded before substantiation, they remain permanently stored on the Blockchain without additional instruments, such as deletion records office or linked modification records, being implemented. These encounters highlight the prerequisite for efficient off-chain loading solutions, cost-effective organization strategies, pervasive cognizance, and robust authentication processes to develop the system's feasibility and usability.

7 Conclusion and Future Scope

This paper presents a blockchain-based explanation for student document administration that efficiently discourses life-threatening contests in sanctuary, transparency, and effectiveness when managing academic archives. By leveraging regionalized tools, the system confirms that academic booklets, such as transcriptions and documentation, are firmly deposited and easily reachable to students. This sanctions scholars by giving them complete control over their archives, justifying apprehensions about data breaks and unlicensed modifications. Furthermore, the incorporation of user-friendly crossing points makes the interaction for operators easier, making it easier for them to upload, manage, and validate their booklets.

The system also rationalizes substantiation processes for commissioners and verifiers, such as budding employers and edifying institutions. By rejecting the need for intermediaries, the system reduces the time and effort required for manuscript validation, improving overall proficiency. The promising introductory results indicate vigorous budding for broader approval of this technology within informative institutions. Impending research will be applied to refine the system based on user feedback and explore additional geographies that could enhance functionality. Additionally, research into other applications of blockchain equipment in the informative sector may make known pioneering ways to improve the supervision and security of various academic processes. Overall, this research paves the way for a safer, sound, and more efficient impending in student document treatment. The future scope of blockchain-based student document supervision systems offers several occasions for modernization and enhancement. One key area is integrating multiple blockchains, which concluded cross-chain interoperability, supporting maintenance

for countless blockchain platforms, and collective classification elasticity. Furthermore, implementing Layer-2 scaling resolutions, such as Polygon, can expressively condense operation expenses and recover structure proficiency, making construction blockchain espousal more achievable for edifying institutions. Another auspicious improvement is AI-powered substantiation, where artificial astuteness can be leveraged for computerized fraud recognition and manuscript justification, improving security and steadfastness.

Further advances can be made through regionalized identity clarifications by incorporating self-sovereign individuality (SSI) frameworks and agreeing on broader submissions for identity supervision beyond academic records. Employing mobile-based and biometric substantiation machinery can develop safekeeping and approachability, ensuring seamless user contacts. Moreover, global adoption and adjustment energies, counting relationships with educational institutions and administration bodies, can simplify constructing a comprehensively accepted blockchain-based credentialing system. Lastly, progressing innovative contract functionalities can empower dynamic document admission control and computerized substantiation, further establishing the system's malleability and productivity. These future developments will contribute to a more robust, scalable, and widely accepted blockchain-based secure academic certificate management solution.

Acknowledgements The authors would like to express their heartfelt gratefulness to Dr. Subhash Rathod, who supported the development of this project. Their visions and proficiency have been priceless throughout the research process of the blockchain-based student document management system. Distinct recognition is given to the faculty members who provided critical direction and positive feedback, safeguarding that the system effectively meets user needs.

The authors also appreciate the Marathwada Mitra Mandal's Institute of Technology, Lohegaon, Pune, who made this work possible and sponsored the entire project. The contributions of all involved have greatly enhanced this research's overall quality and effectiveness, for which the authors are deeply thankful.

References

1. Ahmed, M.R., Muzahidul Islam, A.K.M., Shatabda, S., Islam, S.: Blockchain-based identity management system and self-sovereign identity ecosystem: a comprehensive survey (2022)
2. Brown, E.: Decentralized identity management systems: applications and challenges (2024)
3. Johnson, M.: Blockchain-based document verification systems: case studies in education, July 2022 (2022)
4. Zheng, Z., Xie, S., Dai, H.N., Chen, X., Wang, H.: Blockchain technology for social impact: a systematic literature review. In: Proceedings of the 2018 International Conference on Blockchain Technology (2018)
5. Huang, S., Wu, Y., Zhao, Y.: Blockchain-based digital certificates: a case study on academic credentials. *Int. J. Inf. Manag.* **57**, 102303 (2021)
6. Rao, R.S., Kumari, V.: Blockchain technology in education: a systematic review. *J. King Saud Univ. – Comput. Inf. Sci.* (2020)
7. Swan, M.: *Blockchain: Blueprint for a New Economy*. O'Reilly Media (2015)
8. Li, J., Zhang, H.: A blockchain-based approach for secure and efficient data sharing in higher education. *J. Educ. Technol. Soc.* **23**(1), 162–175 (2020)

9. Ali, M., Badr, Y.: Blockchain-based approach for secure electronic certificates: a case study of academic credentials. *Futur. Gener. Comput. Syst.* **116**, 328–340 (2021)
10. Kouadio, A.M., Togo, D.: Blockchain technology: a solution for managing educational credentials. *Educ. Inf. Technol.* **25**(4), 2845–2865 (2020)
11. Nakamoto, S.: Bitcoin: a peer-to-peer electronic cash system (2008)
12. Christidis, K., Devetsikiotis, M.: Blockchains and smart contracts for the internet of things. *IEEE Access* **4**, 2292–2303 (2016)
13. Satpathy, S., Mahapatra, S., Singh, A.: Fusion of blockchain technology with 5G: a symmetric beginning. In: Tanwar, S. (eds.) *Blockchain for 5G-Enabled IoT*. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67490-8_3
14. Schmidt, K., He, J.: Decentralized identity management using blockchain technology. *Future Internet* **12**(6), 103 (2020)

A Block-Chain-Based Security Framework for Protecting Patient Electronic Health Records



Natasha Wanjari, Pratiksha Chafle, and Rahul Moriwal

Abstract Blockchain technology has become an essential tool for trust and security in many areas, especially in healthcare. Blockchain uses a cryptographic hash function to change blocks because it can be viewed forever continuously. Efforts to change one section will cause changes in the next section, thus increasing security. The proposed tool also ensures coordination and consistency of daily changes in child-care centers. Due to security concerns and cumbersome procedures, medical records are essential but difficult to collect and share. This article combines blockchain innovation with data recovery to promote efficient and effective data sharing. This integration will reduce data traffic and access, leveraging blockchain security capabilities, thereby increasing trust and transparency in data management distress.

Keywords Blockchain · EMR · Smart contract · Patient · Encryption · Hyperledger

1 Introduction

Blockchain innovation has gained much attention and popularity due to its various applications in different research areas and its importance in society. Although Blockchain has not yet been implemented and is in the research stage, it is considered a solution to many modern problems, such as self-governance, transparency, ownership of information, decision-making, and access to decision-making information [1]. It works cooperatively among the flat owners; they receive the same instance of

N. Wanjari · P. Chafle (✉) · R. Moriwal

Department of Computer Science and Engineering, G H Raisoni College of Engineering Nagpur, Nagpur, India

e-mail: pratiksha.chafle@raisoni.net

N. Wanjari

e-mail: natasha.wanjari.trs@ghrce.raisoni.net

R. Moriwal

e-mail: rahul.moriwal@raisoni.net

real estate and are subject to the same rules. Regarding data transfers, Blockchain continues to develop with the explosion of modern data. This innovation guarantees user security and data consistency through cryptographic transactions and protocol understanding. Big data has seven “V” characteristics: quantity, type, medium, variability, accuracy, knowledge, and respect [2]. These characteristics highlight management issues and provide insights from a wide range of literature. Capacity is related to the size of the data, while normal velocity is the speed of receiving data. Distribution indicates the origin and quality of the product, while variability indicates the difference the product will have. Reality is the source of good information, while visualization helps to understand information through visual representation. Finally, the value represents the beneficial results obtained from processing data [3].

Blockchain technology offers great solutions for various big data applications, including identity and information management, asset and supply chain management, Internet of Things (IoT) communications, and healthcare [4].

The service concept uses blockchain technology to help secure and manage data directly. Professionals have access to silent information, allowing them to easily store the medical information they need for diagnosis and treatment [5]. The block diagram and working of blockchain technology are shown in Figs. 1 and 2, respectively.

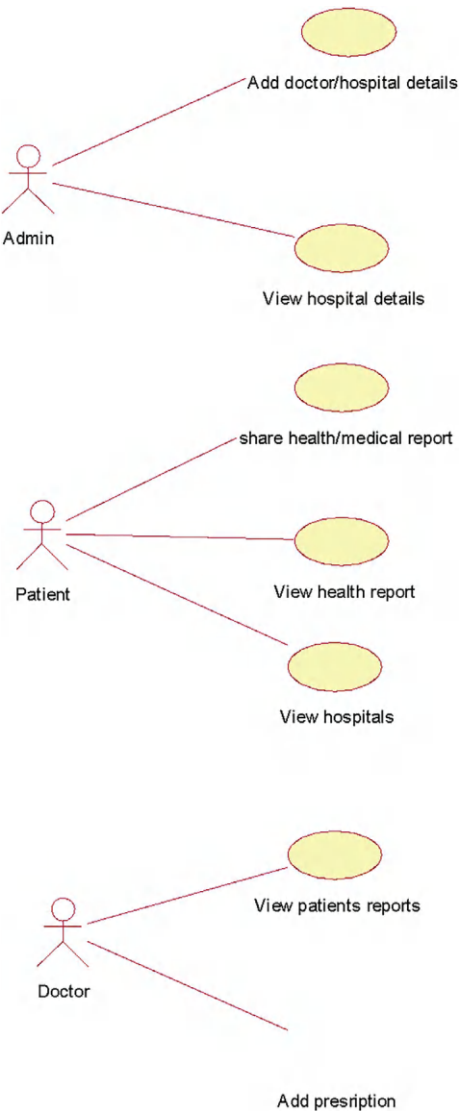
Instead of relying on a central location to store data, we leverage the control of the Blockchain. Think of the Blockchain as a secure computer database. It organizes all data into a block and stores it on different computers (hubs). This strategy makes it difficult for anyone to change the data. It increases data security and ensures that medical data is still available if a problem occurs in any part of the system. It is continuous and shared with many partners, including providers, surveillance companies, patients, families, pharmacies, and researchers [6].

2 Literature Review

Electronic Health Record (EHR) Security The most significant concern in proposing a cloud-native healthcare system is the potential for transferring health data to unauthorized parties. Centralizing electronic medical records poses security risks and requires trust in the organization’s willingness to protect data from insider attacks [7]. We evaluate most of our methods through recycling. The results work well because expanding the hubs in an organization means increasing the size of the supply chain. The results also show that despite the increase in network composition, the time to add the EMR routine to the Blockchain is still short [8]. However, [9] Electronic health records (EHRs) have data security, decision-making, and management issues. This paper discusses how blockchain innovation can change the EHR framework and the potential to solve these issues.

Today, healthcare analysts face challenges such as insufficient communication, insufficient data, and delayed feedback due to electronic repair issues, making improving difficult [10]. Our proposed framework requires secure business data, robust quality, and high-performance data sharing, while data transformations include

Fig. 1 Block diagram



Blockchain and necessary consensus. It is suitable for continuous use in hospitals [11]. However, this situation can be reduced with advances in medical records, the use of insurance companies, and blockchain technology. Blockchain was initially designed to provide proof of transaction for transactions that do not rely on centralized experts or for-profit companies [12]. Therefore, we need a completely decentralized patient-heart approach that will catch information theft and ensure the use of information, but the understanding will be controlled. Blockchain innovation is the solution to solve all the problems and meet the requirements. Blockchain can be a

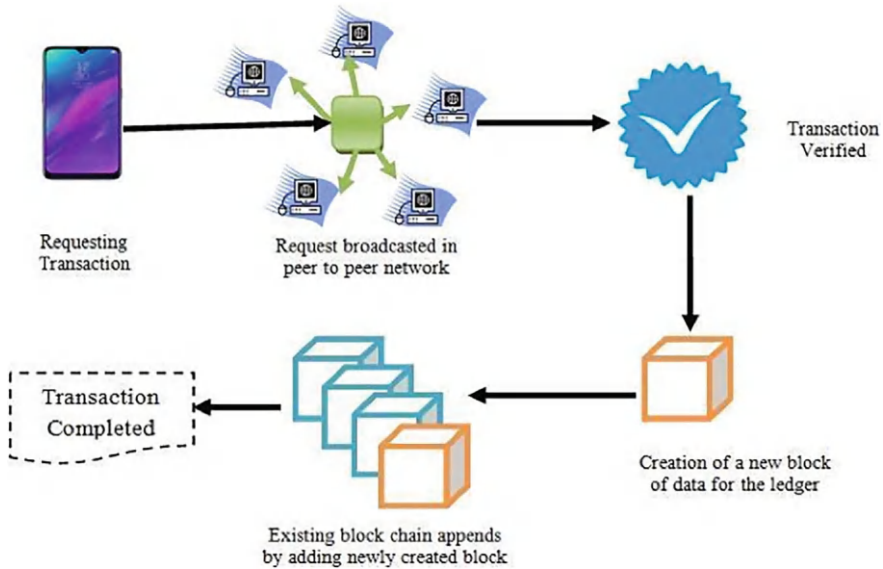


Fig. 2 Working of blockchain

method of information disclosure and distribution that will also affect future integration, medical research, data theft, and financial crime investigations [13]. Patients' personal information (name, address, and disease) is frequently leaked in today's cities and buildings, particularly affecting information security [14]. We propose a novel EHR integration system that combines Blockchain and the decentralized Interplanetary Registry Framework (IPFS) on a portable cloud platform [15].

As part of the review, we recently presented a historical paper on the EHR framework and Blockchain, which explored the (potential) use in the EHR framework [16]. We also identified some challenges and conducted research. This paper introduces a security-based blockchain specifically to meet the needs of e-healthcare [17]. Despite the increasing concerns about the security of electronic medical records and the efforts of cities worldwide to create smart cities, data security remains an issue with implications for criminal records. Previous efforts to address this issue have often resulted in patients being unable to access information [18]. Based on the content of the corrupted content behavioral encryption framework and IPFS media capacity, this paper develops a behavioral encryption framework combined with blockchain innovation to provide security capabilities and share electronic data back in IPFS capacity [19]. Data breaches in electronic health records (EHRs) can cause long-term prevention (i.e., treatment) to be interrupted [20].

Blockchain technology has attracted widespread attention since its inception, and many ways to evade traditional regulatory systems have emerged. Blockchain will likely be relied upon for a long time in many areas, such as smart buildings, healthcare, capital markets, information management, and security [21]. The proposed system

provides a secure, simple, and stable way to store, share, and market EHRs through a peer-to-peer protocol involving multiple partners [22, 23]. With the advancing understanding of Blockchain, cloud computing, and the significant data era, the security and decentralization of medical information need to be ensured and stored in the cloud, which can be planned. Still, it is difficult to ensure the security of medical information [24, 25]. There are many problems in current electronic medical record (EMR) records and data management, such as fragmentation, security, and protection of medical records. This research [26] focuses on creating an electronic health record (EHR) based on the Ethereum blockchain platform and private contracts to eliminate the need for third parties. Through this framework, professionals can understand the information and allow the patient to access this information [27]. On the other hand, adaptability and interoperability are also issues that must be considered in the final planning. This article presents the problem in detail and highlights the advantages of blockchain innovation in using secure medical data and its potential to transform knowledge [28].

This paper presents a secure, cloud-supported e-recovery framework that leverages blockchain innovation to protect e-recovery data from illicit trade alteration [29]. To overcome these issues and provide a high level of security, this extension offers a strategy to optimize blockchain parameters [30]. Blockchain applications have proven their value in daily work; they provide more information and security. It delves into the most minor details of Blockchain: the hash code behind each block, decentralized innovation, smart contracts, and private and public contracts. Open Blockchain [31]. Blockchain is an energy-efficient technology that is currently widely used. Blockchain management is private and secure, thus ensuring a certain level of trust. This is one of the main reasons why Blockchain is accepted in non-religious communities. However, starting with unused money often leads to impediments. Impedance leads to error, and blockchain innovation is no exception [32]. To meet these requirements, WMSN architects use various security measures such as message encryption, hashing, steganography, etc. [33]. The proposal's security check showed that combining Blockchain and encrypted SDN could eliminate more than 95% of cyber attacks compared to non-blockchain computing [34]. Blockchain innovation has become popular with the Internet of Things (IoT) development. Blockchain provides an innovative solution not used by NDN (NPED Information Organization) to store trusted information without relying on the blockchain innovation community [35]. The development and progress of Blockchain continue by bringing together experts and analysts. Blockchain has many advantages, including distributed ledger, security, and trustless models [36, 37]. In fact, after success, the economy cannot catch up with them, give them more time, and not assist in their success [38]. Client authentication and access control modules enforce access through different verification strategies and strict controls, predict unauthorized access, and monitor role-based authorization [39]. Blockchain has finally become a significant innovation for securing information through a decentralized system. Blockchain is a secure database that enables safe and secure transactions [40]. Many subsets of the standard protocol can be helpful to analysts from various blockchain mining services.

However, most of these innovations are very complex, which only reduces the preparation of large blockchains. Non-uniform models with repeatability are wasteful and do not make sense [41]. Blockchain is described as an era of development for honest and shared data exchange among large communities of non-believers [42]. This paper will present experts specializing in blockchain use and complete and analyze the development and continuous improvement problems to overcome these problems. More importantly, this paper also examines the possibility of getting rich quickly [43]. However, most of these innovations are pretty complex, slowing down the large-scale preparation of blockchains for mining [44, 45].

3 Methodology

3.1 *Proposed Work*

The aim is to use blockchain technology to manage medical information and solve the shortcomings of the current centralized system. The integration of Blockchain offers many advantages that can improve the security, efficiency, and accessibility of medical information. Blockchain's pure-add structure and cryptographic hash function ensure the transferability and security of medical information.

3.2 *System Architecture*

The design concept uses blockchain technology to help manage medical information securely and transparently. The architecture, as shown in Fig. 3, is based on three leading roles: doctors, administrators, and patients. Doctors can access patient information, allowing them to store medical information necessary for diagnosis and treatment. Administrators monitor hospital operations and have the right to view hospital-related details stored on the Blockchain to ensure proper hospital management. Patients can view their health information and provide secure medical information to authorized persons or organizations and security.

3.3 *Modules*

To implement this project, we used the following modules: admin, doctor, and patient.

Admin Module:

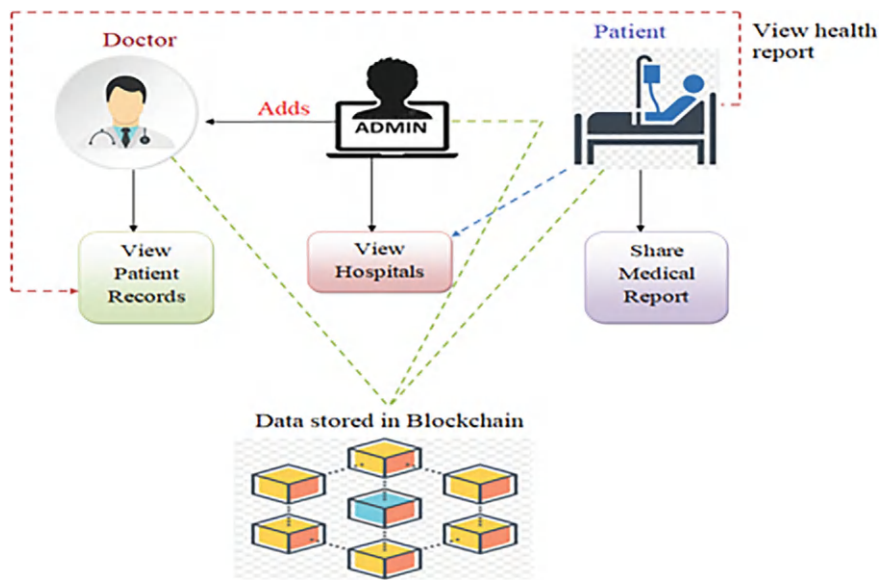


Fig. 3 Proposed architecture

Add Doctor/Hospital Details

Administrators can access and track important information about doctors and hospitals in the management module. This allows them to maintain quality records, including doctor records, contact information, and hospital records. By facilitating the linking and managing of this vital information, administrators can improve the process of updating and managing critical information in the healthcare system.

View Hospital Details

Administrators can access hospital information in the system. This includes hospital name, address, specialty, and other relevant information. By accessing this information, administrators can have a better view of the treatment in the system. This leads to informed decision-making and effective healthcare management.

Patient Module:

New Patient Signup

The patient module allows new patients to be enrolled in the electronic health record (EHR). This process requires the creation of a user account, where the patient provides the necessary personal information, creates login credentials, and creates a profile.

View Hospitals

Patients can access a complete list of hospitals integrated into the system, helping them make informed decisions about sharing their medical information. This feature provides detailed information about different medical facilities to help patients choose the best options.

Share Health/Medical Report

A key feature in the patient module allows patients to choose their medical information and treatment at selected hospitals. Patients gain control over their personal information by telling which hospitals can access it. These features promote trust and collaboration between patients and doctors by ensuring data security and privacy.

View Health Report

Patients can easily access and view their medical and medical records through the system. This allows patients to access their medical records and review their medical history anytime. Providing patients with direct access to their data will enable them to be informed about their health and share relevant information with doctors when necessary; this supports management in respecting health.

Doctor Module:

View Patient Reports

In the doctor module, doctors can access patient information, making storing patients' medical information easier. This allows healthcare providers to manage patient information better and support personalized healthcare. By providing uninterrupted access to patient information, doctors can make medical decisions and provide the proper treatment, ultimately improving the quality and effectiveness of pain management.

4 Experimental Results

In the above screen, the admin is logged in, as shown in Fig. 4, and after LoginLogin, you will get the below screen. Then, click on the 'Add Doctor/Hospital Details' link.

On the above screen, the admin will add the doctor, as shown in Fig. 5, and hospital details and then press the submit button. Click the 'View Hospital Details' link to get the details below.

After LoginLogin, patients can click on the 'View Hospitals' link, as shown in Fig. 7, to view a list of all available doctors and hospitals like in the below screen. The patient can view all hospital details and decide which one to share. Then, the patient can click the 'Share Health/Medical Report' link to add a health report. In the above screen, the patient can see symptom details, upload any medical report, and select multiple hospitals from the LIST by holding the 'CTRL' key and pressing the Submit button to share details with selected hospitals.

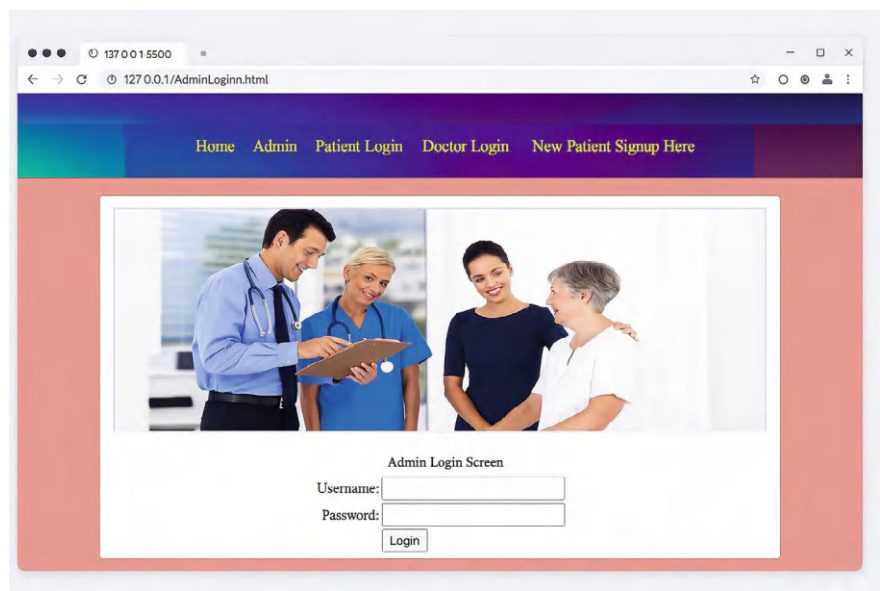


Fig. 4 Admin login



Fig. 5 Inserting doctor

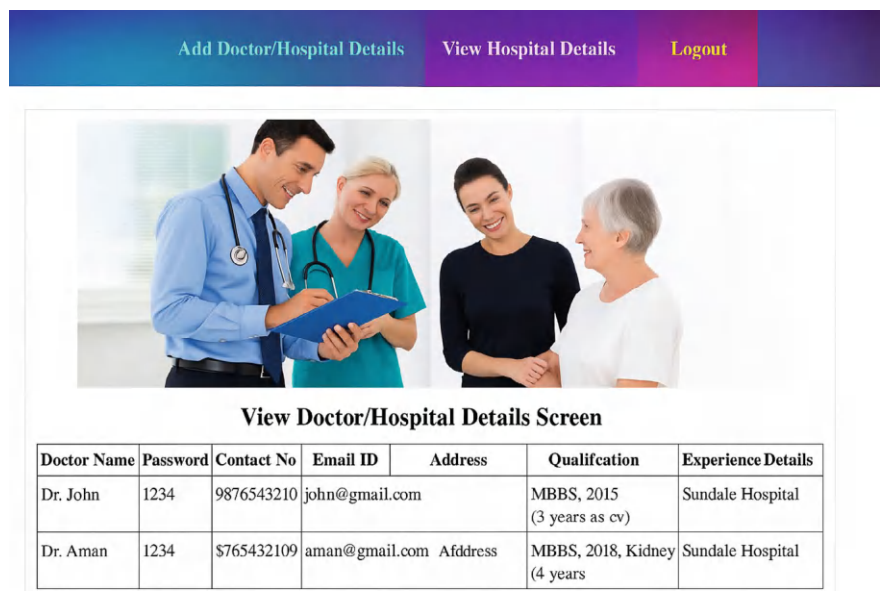


Fig. 6 Patient login: now click ‘patient login’ to log in as a patient, as shown in Fig. 6

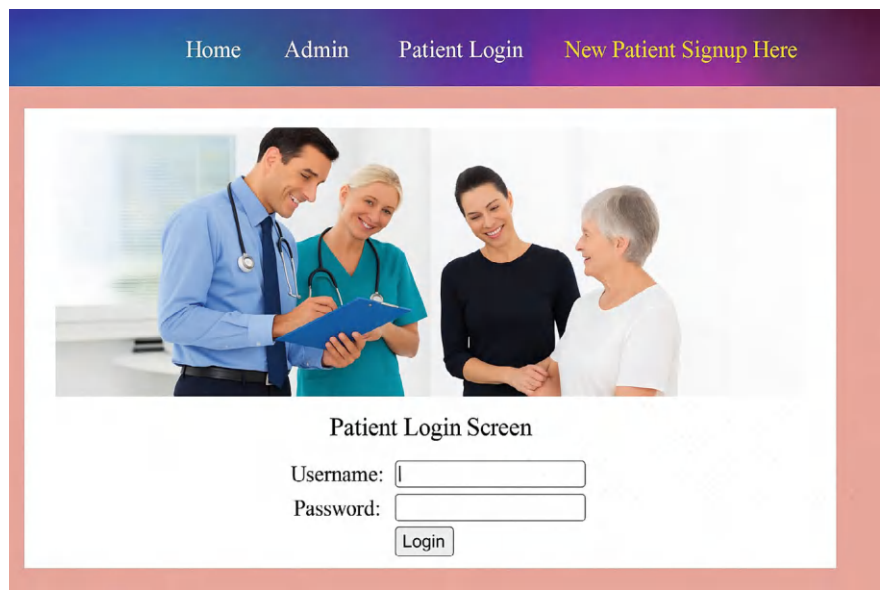


Fig. 7 View hospital

Now, the patient can click on ‘View Health Report’ as shown in Fig. 8, to view all reports he shares.

In the above screen, the patient can view all his disease details, as shown in Fig. 9, with the uploaded report image, and now log in and log in as a doctor to view this patient report.

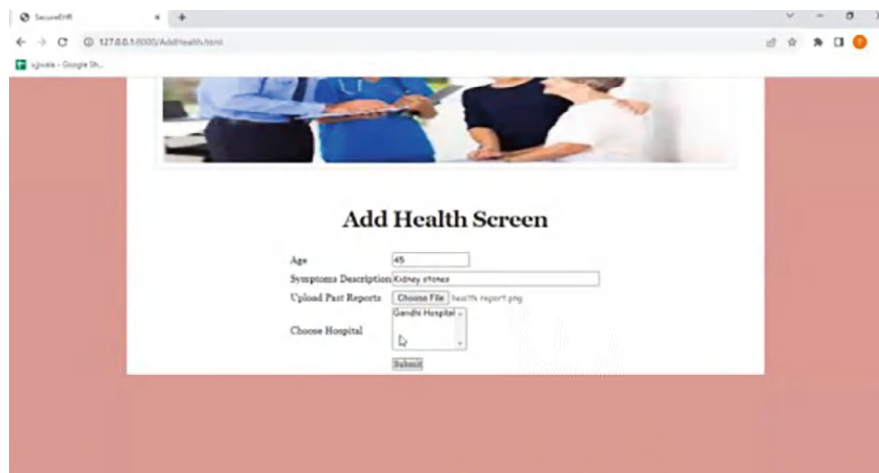


Fig. 8 View health report

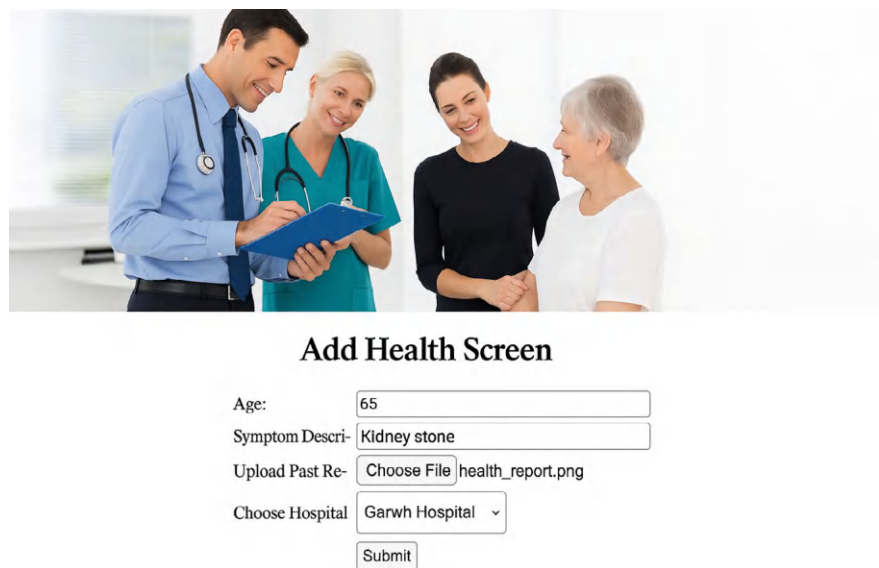



Fig. 9 Disease details

[Home](#) [Admin](#) [Patient Login](#) [Doctor Login](#) [New Patient Signup Here](#)




Doctor Login Screen

Username:

Password:

Fig. 10 Doctor login

On the above screen, the doctor is LoginLogin, as shown in Fig. 10, and after LoginLogin, the screen will be shown below. After LoginLogin, the doctor can access/view that patient's details, and based on those details, he will prescribe treatment. Now, doctors can click on the 'Click Here' link to give a prescription. As shown in Fig. 11, the doctor can write a prescription and press the 'Submit' button to add the prescription.



Prescription Screen

Patient Name: Smith

Enter Prescrip-

I

Submit

Fig. 11 Write a prescription

5 Conclusion

Overall, the proposed approach represents a significant advancement in the treatment of information management by utilizing blockchain technology to solve critical problems and increase security, efficiency, and access to medical information. Medical data exchange is performed through cryptographic hashing, which ensures a good understanding of the data so that electronic data is not compromised and damaged. In addition, the framework facilitates personal care and silent care by allowing users to control access to their medical information. The framework meets the needs of repair and customers with its effectiveness in planning and customer collaboration, supports the implementation of interventions, and ensures that all customers are successful.

References


1. Chentharu, S., Ahmed, K., Wang, H., Whittaker, F., Chen, Z.: Healthchain: a novel framework on privacy preservation of electronic health records using blockchain technology. *PLoS ONE* **15**(12), e0243043 (2020). <https://doi.org/10.1371/journal.pone.0243043>
2. De Oliveira, M.T., Reis, L.H., Carrano, R.C., Seixas, F.L., Saade, D.C., Albuquerque, C.V., Fernandes, N.C., Olabarriaga, S.D., Medeiros, D.S., Mattos, D.M.: Towards a blockchain-based secure electronic medical record for healthcare applications. In: ICC 2019–2019 IEEE International Conference on Communications (ICC), pp. 1–6, May 2019. IEEE (2019). <https://doi.org/10.1109/ICC.2019.8761307>
3. Shahnaz, A., Qamar, U., Khalid, A.: Using Blockchain for electronic health records. *IEEE Access* **7**, 147782–147795 (2019). <https://doi.org/10.1109/ACCESS.2019.2946373>
4. Zarour, M., Ansari, M.T.J., Alenezi, M., Sarkar, A.K., Faizan, M., Agrawal, A., Kumar, R., Khan, R.A.: Evaluating the impact of blockchain models for secure and trustworthy electronic healthcare records. *IEEE Access* **8**, 157959–157973 (2020). <https://doi.org/10.1109/ACCESS.2020.3019829>
5. Shamshad, S., Mahmood, K., Kumari, S., Chen, C.M.: A secure blockchain-based e-health records storage and sharing scheme. *J. Inf. Secur. Appl.* **55**, 102590 (2020). <https://doi.org/10.1016/j.jisa.2020.102590>
6. Tamazirt, L., Alilat, F., Agoulmine, N.: Blockchain technology: a new secured electronic health record system. In: 6th International Workshop on ADVANCES in ICT Infrastructures and Services (ADVANCE 2018), pp. 134–141, January 2018 (2018)
7. da Conceição, A.F., da Silva, F.S.C., Rocha, V., Locoro, A., Barguil, J.M.: Electronic health records using blockchain technology (2018). [arXiv:1804.10078](https://arxiv.org/abs/1804.10078), <https://doi.org/10.48550/arXiv.1804.10078>
8. Mahore, V., Aggarwal, P., Andola, N., Venkatesan, S.: Secure and privacy-focused electronic health record management system using permissioned blockchain. In: 2019 IEEE Conference on Information and Communication Technology, pp. 1–6, December 2019. IEEE (2019). <https://doi.org/10.1109/CICT48419.2019.9066204>
9. Hasan, Q.H., Yassin, A.A., Ata, O.: Electronic health records system using blockchain technology (2021)
10. Linn, L.A., Koo, M.B.: Blockchain for health data and its potential use in health and healthcare-related research. In: ONC/NIST, Blockchain is used for healthcare and research workshops, pp. 1–10, September 2016. ONC/NIST, Gaithersburg, Maryland, United States (2016)
11. Tanwar, S., Parekh, K., Evans, R.: Blockchain-based electronic healthcare record system for healthcare 4.0 applications. *J. Inf. Secur. Appl.* **50**, 102407 (2020). <https://doi.org/10.1016/j.jisa.2019.102407>
12. Poonguzhali, N., Gayathri, S., Deebika, A., Suriapriya, R.: A framework for electronic health record using blockchain technology. In: 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 1–5, July 2020. IEEE (2020). <https://doi.org/10.1109/ICSCAN49426.2020.9262369>
13. Zhang, A., Lin, X.: Towards secure and privacy-preserving data sharing in e-health systems via consortium blockchain. *J. Med. Syst.* **42**(8), 140 (2018). <https://doi.org/10.1007/s10916-018-0995-5Keywords>
14. Vora, J., Nayyar, A., Tanwar, S., Tyagi, S., Kumar, N., Obaidat, M.S., Rodrigues, J.J.: BHEEM: a blockchain-based framework for securing electronic health records. In: 2018 IEEE Globecom Workshops (GC Wkshps), pp. 1–6, December 2018. IEEE (2018). <https://doi.org/10.1109/GLOCOMW.2018.8644088>
15. Nguyen, D.C., Pathirana, P.N., Ding, M., Seneviratne, A.: Blockchain is used for secure EHR sharing of mobile cloud-based e-health systems. *IEEE Access* **7**, 66792–66806 (2019). <https://doi.org/10.1109/ACCESS.2019.2917555>
16. Shi, S., He, D., Li, L., Kumar, N., Khan, M.K., Choo, K.K.R.: A survey of applications of blockchain in ensuring the security and privacy of electronic health record systems. *Comput. Secur.* **97**, 101966 (2020). <https://doi.org/10.1016/j.cose.2020.101966>

17. Pandey, P., Litoriya, R.: Securing and authenticating healthcare records through blockchain technology. *Cryptologia* **44**(4), 341–356 (2020). <https://doi.org/10.1080/01611194.2019.1706060>
18. Dagher, G.G., Mohler, J., Milojkovic, M., Marella, P.B.: Ancile: privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology. *Sustain. Cities Soc.* **39**, 283–297 (2018). <https://doi.org/10.1016/j.scs.2018.02.014>
19. Sun, J., Yao, X., Wang, S., Wu, Y.: Blockchain-based secure storage and access scheme for electronic medical records in IPFS. *IEEE Access* **8**, 59389–59401 (2020). <https://doi.org/10.1109/ACCESS.2020.2982964>
20. Chen, L., Lee, W.K., Chang, C.C., Choo, K.K.R., Zhang, N.: Blockchain-based searchable encryption for electronic health record sharing. *Future Gener. Comput. Syst.* **95**, 420–429 (2019). <https://doi.org/10.1016/j.future.2019.01.018>
21. Chakraborty, S., Aich, S., Kim, H.C.: A secure healthcare system design framework using blockchain technology. In: 2019 21st International Conference on Advanced Communication Technology (ICACT), pp. 260–264, February 2019. IEEE (2019). <https://doi.org/10.23919/ICACT.2019.8701983>
22. Uddin, M., Memon, M.S., Memon, I., Ali, I., Memon, J., Abdelhaq, M., Alsaqour, R.: Hyper-ledger fabric blockchain: Secure and efficient solution for electronic health records. *Comput. Mater. & Contin.* **68**(2), 2377–2397 (2021). <https://doi.org/10.32604/cmc.2021.015354>
23. Singh, C., Chauhan, D., Deshmukh, S.A., Vishnu, S.S., Walia, R.: Medi-Block record: secure data sharing using blockchain technology. *Inform. Med. Unlock.* **24**, 100624 (2021). <https://doi.org/10.1016/j.imu.2021.100624>
24. Gutiérrez, O., Romero, G., Pérez, L., Salazar, A., Charis, M., Wightman, P.: Healthyblock: blockchain-based architecture for electronic medical records resilient to connectivity failures. *Int. J. Environ. Res. Public Health* **17**(19), 7132 (2020). <https://doi.org/10.3390/ijerph17197132>
25. Fatokun, T., Nag, A., Sharma, S.: Towards a blockchain-assisted patient-owned system for electronic health records. *Electronics* **10**(5), 580 (2021). <https://doi.org/10.3390/electronics10050580>
26. Rahman, M.S., Khalil, I., Mahawaga Arachchige, P.C., Bouras, A., Yi, X.: A novel architecture for tamper-proof electronic health record management system using blockchain wrapper. In: Proceedings of the 2019 ACM International Symposium on Blockchain and Secure Critical Infrastructure, pp. 97–105, July 2019 (2019). <https://doi.org/10.1145/3327960.3332392>
27. Rifi, N., Rachidi, E., Agoulmine, N., Taher, N.C.: Towards using blockchain technology for eHealth data access management. In: 2017 fourth international conference on advances in biomedical engineering (ICABME), pp. 1–4, October 2017. IEEE (2017). <https://doi.org/10.1109/ICABME.2017.8167555>
28. Ramachandran, S., Kiruthika, O.O., Ramasamy, A., Vanaja, R., Mukherjee, S.: A review on blockchain-based strategies for management of electronic health records (EHRs). In: 2020 International Conference on Smart Electronics and Communication (ICOSEC), pp. 341–346, September 2020. IEEE (2020). <https://doi.org/10.1109/ICOSEC49089.2020.9215322>
29. Krishnan, S.S.R., Manoj, M.K., Gadekallu, T.R., Kumar, N., Maddikunta, P.K.R., Bhattacharya, S., Suh, D.Y., Piran, M.J.: A blockchain-based credibility scoring framework for electronic medical records. In: 2020 IEEE Globecom Workshops (GC Wkshps), pp. 1–6, December 2020. IEEE (2020). <https://doi.org/10.1109/GCWkshps50303.2020.9367459>
30. Choudhary, S., Dorle, S.: A quality of service-aware high-security architecture design for software-defined network powered vehicular ad-hoc networks using machine learning-based blockchain routing. *Concurr. Comput.: Pract. Exp.* **34**(17), e6993 (2022). <https://doi.org/10.1002/cpe.6993>
31. Shrawankar, U., Shrawankar, C.: BlockCloud: blockchain as a cloud service. In: *Blockchain for Smart Systems*, pp. 53–63. Chapman and Hall/CRC (2022). <https://doi.org/10.1201/9781003203933-5>
32. Malik, L., Arora, S., Shrawankar, U., Deshpande, V. (eds.): *Blockchain for Smart Systems: Computing Technologies and Applications*. CRC Press (2022). <https://doi.org/10.1201/9781003203933>

33. Sharma, A.K., Sharma, D.M., Purohit, N., Sharma, S.A., Khan, A.: Blockchain technology: Myths, realities and future. In: *Blockchain Technology*, pp. 163–180. CRC Press (2022). <https://doi.org/10.1201/9781003138082-10>
34. Rahangdale, H., Chavhan, N.: Design of a blockchain based WMSN model for secure and effective health monitoring system deployment. In: *2022 10th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-22)*, pp. 1–6, April 2022. IEEE (2022). <https://doi.org/10.1109/ICETET-SIP-2254415.2022.9791717>
35. Choudhary, S., Dorle, S.: Secured SDN based Blockchain: an architecture to improve the security of VANET. *Int. J. Electr. Comput. Eng. Syst.* **13**(2), 145–153 (2022). <https://doi.org/10.32985/ijeces.13.2.7>
36. Harshini, V.M., Danai, S., Usha, H.R., Kounte, M.R.: Health record management through blockchain technology. In: *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1411–1415, April 2019. IEEE (2019). <https://doi.org/10.1109/ICOEI.2019.8862594>
37. Agrawal, R., Dorle, S., Dhule, C., Agrawal, U.: Effective network communication based on blockchain based trusted networks. In: *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 646–651, June 2021. IEEE (2021). <https://doi.org/10.1109/ICOEI51242.2021.9452989>
38. Maidamwar, P., Saraf, P., Chavhan, N.: Blockchain applications, challenges, and opportunities: a survey of a decade of research and future outlook. In: *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, pp. 1–5, November 2021. IEEE (2021). <https://doi.org/10.1109/ICCICA52458.2021.9697256>
39. Soni, A.K., Padiya, A., Patel, B., Raza, S.H., Malviya, K.: Music streaming using blockchain-blockies. In: *2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECs)*, pp. 1–6, February 2023. IEEE (2023). <https://doi.org/10.1109/SCEECs57921.2023.10062986>
40. Dhote, S., Maidamwar, P., Thakur, S.: Integrating blockchain and multi-factor authentication for enhanced cloud security in certificate verification systems. In: *2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*, Bengaluru, India, pp. 1–7 (2024). <https://doi.org/10.1109/ICDCOT61034.2024.10516190>
41. Mahindre, S., Gupta, S., Rambhad, V., Koppiseti, S., Thakur, S.: Secure splitting of bills and managing expenditure using blockchain. In: *2023 IEEE 3rd International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET)*, Mysuru, India, pp. 1–7 (2023). <https://doi.org/10.1109/TEMSMET56707.2023.10149962>
42. Bhoware, A., Jajulwar, K., Ghodmare, S., Dabhekar, K., Bartakke, V.: Performance analysis of network management system using bioinspired-blockchain technique for IP networks. In: *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, Trichy, India, pp. 201–205 (2023). <https://doi.org/10.1109/ICSMDI57622.2023.00045>
43. Agrawal, R., Dorle, S., Dhule, C., Agrawal, U.: Effective network communication based on blockchain-based trusted networks. In: *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, pp. 646–651 (2021). <https://doi.org/10.1109/ICOEI51242.2021.9452989>
44. Satpathy, S., Mahapatra, S., Singh, A.: Fusion of blockchain technology with 5G: a symmetric beginning. In: Tanwar, S. (eds.) *Blockchain for 5G-Enabled IoT*. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67490-8_3
45. Maidamwar, P., Saraf, P., Chavhan, N.: Blockchain applications, challenges, and opportunities: a survey of a decade of research and future outlook. In: *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, Nagpur, India, pp. 1–5 (2021). <https://doi.org/10.1109/ICCICA52458.2021.9697256>

Blockchain-Powered Secure EHR Exchange in Mobile Cloud E-Health Systems



S. P. Santhoshkumar , V. R. Navinkumar, Rekhasree Manthu, S. Hariharasudhan, N. Ramajayam, and S. Gajalakshmi

Abstract The management of Electronic Health Records (EHRs) has been completely transformed by the emergence of mobile and cloud computing, which has improved healthcare's accessibility, affordability, and flexibility. Sensitive medical data security in mobile cloud systems is still difficult to achieve, nevertheless. The InterPlanetary File System (IPFS) and blockchain technologies are integrated into mobile cloud services in this paper's proposed secure EHR exchange paradigm. Its smart contract-powered access control mechanism, which guarantees safe and regulated EHR sharing between patients and healthcare providers, is a crucial component. Amazon's cloud platform is used to deliver the framework's prototype, which is based on the Ethereum blockchain, via a mobile application. Its efficacy in protecting health information while facilitating effective communication is validated by empirical assessments. Furthermore, the lightweight access control paradigm improves security and privacy by lowering network latency. For safe EHR sharing in mobile cloud-based healthcare systems, this framework provides an effective and scalable solution.

S. P. Santhoshkumar (✉) · S. Hariharasudhan
Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India
e-mail: spsanthoshkumar16@gmail.com

S. Hariharasudhan
e-mail: drhariharasudhans@veltech.edu.in

V. R. Navinkumar
University College of Engineering Arni, Tiruvannamalai, India

R. Manthu
KU College of Engineering and Technology, KU Campus, Warangal, India

N. Ramajayam
Dhanalakshmi Srinivasan College of Engineering, Coimbatore, India

S. Gajalakshmi
Rathinam Technical Campus, Coimbatore, India
e-mail: 2006.gaja@gmail.com

Keywords Mobile cloud-based e-health systems • Blockchain • Interplanetary File System (IPFS) • Data confidentiality • Blockchain security • Secure EHR sharing

1 Introduction

This paper aims to explore the evolving landscape of Electronic In integration of mobile devices and cloud computing has revolutionized EHR storage and management, enabling seamless data exchange between patients and healthcare providers [1]. Although this shift brings advantages like reduced costs, enhanced flexibility, and improved accessibility, it also introduces significant challenges related to data privacy and network security in EHS [2]. Blockchain technology is increasingly being adopted to enhance medical and EHS. Its decentralized and secure nature makes it highly effective for EHR exchange and data access management across healthcare entities [3]. By ensuring data integrity and security, blockchain has the potential to improve healthcare efficiency and drive a transformative impact on the industry. The emergence of Mobile Cloud Computing (MCC) and the Internet of Medical Things (IoMT) has revolutionized e-health services [4]. Patients can now use smartphones and wearable devices to gather health data, which is then transmitted to cloud platforms for real-time analysis by healthcare professionals [5]. This enables remote monitoring, ambulatory treatment, and cost-effective care. Additionally, cloud-based Electronic Health Records (EHRs) enhance patient tracking, ensuring timely and effective diagnosis and treatment.

Despite its benefits, storing Electronic Health Records (EHRs) on cloud platforms raises security concerns, particularly regarding unauthorized access and data privacy [6]. Unauthorized access can compromise data integrity and security, often without patient consent [7]. Additionally, patients may struggle to track and manage their shared records across multiple providers. Therefore, implementing robust access control mechanisms is essential for secure mobile cloud-based EHR systems [8]. Traditional access control methods for Electronic Health Records (EHRs) assume cloud servers are fully trustworthy, handling both authentication and data access. However, in mobile cloud environments, servers may be both honest and curious, meaning they can process requests while secretly accessing sensitive data. This introduces risks related to unauthorized data exposure and network vulnerabilities. Moreover, traditional systems depend on a centralized access point, which can serve as a particular point of collapse, increasing the susceptibility of e-health networks to security threats. In contrast, blockchain-based access control provides greater security and transparency, offering a more resilient alternative to conventional approaches [9]. By creating immutable transaction ledgers, blockchain ensures data integrity and prevents unauthorized modifications. Additionally, it enhances transparency, as any unauthorized data access is logged and visible to all network participants, reducing data leakage risks.

Smart contracts further strengthen security by automating authentication and user verification, enabling strict access control while mitigating potential threats. Unlike

centralized systems, blockchain eliminates reliance on a single server, promoting fair and decentralized data management [10]. Since smart contracts are publicly distributed, all participants share equal oversight, ensuring trust and resilience even if a network entity fails [11]. Incorporating blockchain technology into EHR systems strengthens security, preserves data integrity, and safeguards patient privacy. It enables secure data exchange between stakeholders while enhancing auditability and accountability, fostering a more reliable and efficient healthcare ecosystem [12].

2 Background and Motivation

This paper investigates how mobile cloud integration enhances accessibility and data sharing in the healthcare industry. However, strong safeguards are required due to security and privacy issues. Knowing the background makes it easier to spot knowledge gaps and the need for more analysis.

The evolution of healthcare technology has transformed data acquisition, storage, and sharing. Mobile devices and cloud computing enable patients to collect health data easily, while cloud platforms allow practitioners to access and analyze it in real-time [13]. This advancement enhances efficiency and expands healthcare services beyond traditional facilities. The advancement of healthcare technology brings challenges, particularly in data protection and confidentiality. As reliance on mobile cloud environments for storing and sharing EHRs grows, vigorous security measures are essential to protect perceptive medical data. Confidentiality and data secrecy are critical in healthcare. Safeguarding sensitive patient information in Electronic Health Records (EHRs) requires strict security measures, especially with cloud-based storage [14]. The hazard of not permitted access, data breach, and cyber-attacks highlights the urgent need for robust protection [15]. Network security is vital in healthcare as cloud-based systems introduce potential vulnerabilities. Cyber threats can compromise medical data and disrupt services, making it essential to ensure the reliability and accessibility of healthcare networks for uninterrupted, serene care.

Healthcare must strike a balance between the advantages of mobile cloud computing and data security. Blockchain provides a decentralized, impenetrable system that guarantees transparency and integrity. Smart contracts revolutionize data sharing and medical record management by enabling safe access control [16]. The integration of mobile cloud environments in healthcare offers transformative potential but also raises protection and confidentiality concerns. Addressing these challenges drives interest in blockchain is vigorous solution to ensure the secure and sustainable growth of E-Health systems while protecting sensitive medical data [17].

3 Background and Motivation

While mobile cloud computing improves decision-making and EHR accessibility, it also poses security and privacy issues. Innovative solutions that strike a balance between security and usability in healthcare data sharing are required since traditional encryption and access controls are ineffective against changing threats [18]. Role-based access control (RBAC), encryption, and authentication are essential components of EHS security. Although these techniques safeguard data, new dangers and interoperability issues necessitate more sophisticated solutions. Blockchain is examined in this research as a novel strategy for improving data security and privacy in mobile cloud-based EHS. With its immutable ledger and smart contracts, blockchain improves EHR security, transparency, and interoperability while lowering breaches and guaranteeing patient control [19]. By using cryptographic hashing and peer-to-peer storage, IPFS improves data integrity and guards against manipulation. Blockchain technology and decentralized storage are investigated in this study for safe EHR administration in mobile cloud-based platforms [20].

3.1 A Proposal for an Energy-Efficient Transaction Model in Blockchain-Enabled Internet of Vehicles (IoV)

This study addresses the energy challenges of blockchain-enabled Internet of Vehicles (IoV) by proposing an energy-efficient model [21]. It optimizes transaction control through distributed clustering, resulting in a 40.16% reduction in energy consumption and an 82.06% decrease in transaction frequency compared to traditional blockchain methods [22].

3.2 A Scholarly Examination of the Process of Scaling Decentralized Blockchains

This research focuses on the scalability issues within blockchain-based cryptocurrencies, particularly Bit coin. It explores the limitations of the system that hinder high throughput and low latency, suggesting that mere adjustments to block size and intervals are inadequate. This paper offers a comprehensive examination of proposed solutions and persistent challenges in advancing blockchain scalability.

3.3 A Framework with Minimal Storage Requirements for Distributed Ledgers in Blockchain Technology

Traditional e-commerce relies on trusted third parties for electronic payments, but this system is vulnerable. Blockchain technology offers a decentralized solution, though it poses a storage challenge, as each node must store all transactions. This paper introduces a novel structure called Network Coded Distributed Storage (NC-DS), which reduces storage needs by implementing Non-Colluding Data Sharding (NC-DS) in blockchain systems, significantly optimizing storage requirements.

3.4 The Integration of Distributed Storage and Secret Sharing in Blockchain Technology

This paper combines distributed storage, private key encryption, and Shamir's secret sharing technique to enhance the security and integrity of blockchain transaction data. By applying Shamir's secret sharing to hash values and utilizing dynamic zone allocation, the system improves data integrity. The research explores the trade-off between storage costs and the risk of data loss, proposing an integer programming model to optimize data retrieval and protection against corruption. This approach offers a solution for maintaining data integrity in blockchain-based cloud storage systems while considering the costs for service providers.

3.5 Optimizing Operational Effectiveness in Distributed Blockchain Systems with Local Secret Sharing

This paper proposes a Local Secret Sharing technique to improve distributed blockchain efficiency. Distributed Storage Blockchain (DSB) systems use encryption and secret sharing to reduce the high storage costs associated with traditional blockchains. However, during failures, communication costs may increase, particularly as a result of DoS assaults. Using a classified covert configuration, the research optimizes storage and recovery communication costs by introducing the Double Secret Box (DSB) methodology with Local-Secret Sharing (LSS).

4 Proposed Methodology

This innovative methodology shifts the focus from securing entire transaction blocks to safeguarding individual blocks, creating a paradigm shift in blockchain security. SHAMIR shares are generated through a transformative process to enhance block

protection. These shares are then distributed to all accessible network nodes. To reconstruct the original block data, an application collects these shares from the nodes and applies the SHAMIR SECRET method. Any missing or incorrect share will cause the reconstruction to fail. The effectiveness of the SHAMIR secret relies on the integration of prime numbers and random polynomials, where block data is initially transformed into a polynomial and later reconstructed using a reverse polynomial technique.

This work introduces an innovative approach that moves beyond traditional blockchain storage models. By enhancing data security, it also offers novel solutions to the storage and scalability challenges that have previously constrained the widespread adoption of blockchain technology [23]. Consider a use case for an EHS leveraging a mobile cloud podium [24]. In this scenario, patient records are collected from local gateways across a network and securely stored on a public cloud, enabling easy access by healthcare providers. These electronic health records contain sensitive data, including private information and medical history provided by patients. Each patient is assigned a unique Patient ID (PID) and is categorized by their geographical region, indicated by an Area ID (AID), which helps organize and manage records based on the patient's location.

4.1 Advantages of the Proposed System

The blockchain-based system for EHR distribution in mobile cloud-based e-health environments offers a number of advantages over conventional methods [25].

Through the reduction of central server dependence and failure concerns, this method improves EHR security and accessibility. Data security and secrecy are guaranteed by encryption, stringent access controls, and the tamper-proof ledger of blockchain technology. Remote locations benefit from seamless access made possible by mobile and cloud integration. Blockchain monitoring increases openness and trust, while interoperability facilitates better collaboration. Cost-effective, it ensures safe and effective management of healthcare data by getting rid of costly infrastructure and lowering breach risks.

5 Implementation

The implementation phase of the proposed EHRs sharing framework is a detailed process that brings the conceptual design to life. This section outlines the steps involved in deploying the framework, including the selecting of the blockchain platform (Ethereum), the development of the mobile application, and the iterative testing phase to refine and optimize the system's functionality [27].

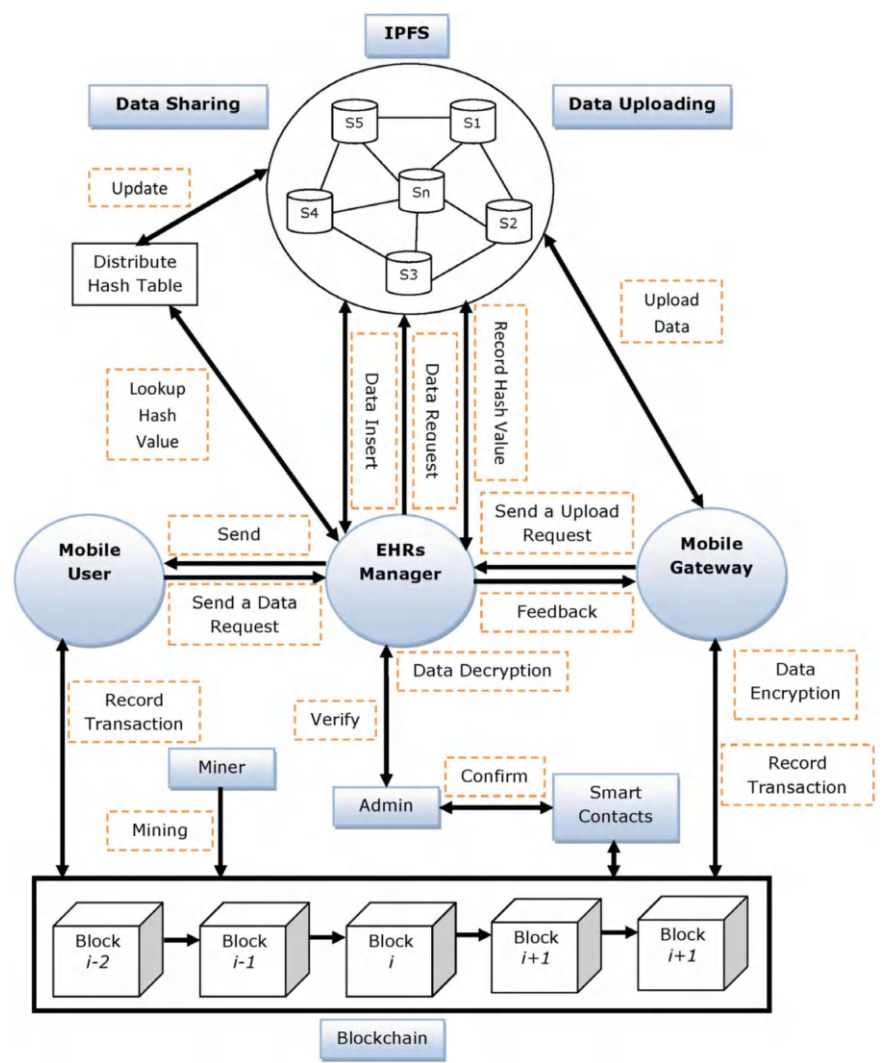


Fig. 1 System proposed block diagram [26]

5.1 Advantages of the Proposed System

Ethereum stands out as the ideal blockchain platform due to its proven reliability and efficiency. Its extensive adoption and versatility make it particularly well-suited for secure sharing of EHRs. Ethereum’s robust smart contract functionality supports a trust-based access control model, ensuring secure and transparent data handling [28].

This selection reflects a preference for a platform with a solid track record in real-world applications, reinforcing its suitability for critical healthcare data management [29].

5.2 Developing the Mobile App

The creation of an intuitive mobile application is crucial for enabling secure sharing of EHRs. This app leverages the scalability, reliability, and robust infrastructure of Amazon Web Services (AWS) to ensure seamless integration with cloud computing services. It is designed to work smoothly with IPFS and blockchain elements, providing continuous connectivity for efficient data transmission and secure storage of patient information.

5.3 Prototype Configuration and Deployment

A prototype that combines an Ethereum blockchain, IPFS, and a mobile app in a mock real-world environment validates the framework. The accuracy, consistency, and viability of network setups, nodes, and cryptographic keys are guaranteed by rigorous testing.

6 Result and Discussion

This part offers a elaborated analysis of the proposed framework's effectiveness and practical applications for EHRs exchange in mobile cloud-based EHS. It reviews empirical data and provides a thoughtful evaluation of the framework's performance and potential impact.

6.1 Advantages of the Proposed System

Evaluating the performance of the framework is crucial. To assess key metrics like network latency (for data transmission speed) and accuracy to gauge user experience. Additionally, to measure throughput to evaluate the efficiency of data exchanges. The framework's security features, including data integrity, confidentiality, and authentication, are also thoroughly examined to ensure robust protection.

6.2 Comparison of the Model to Existing Models

Thoroughly compare our innovative framework with existing data-sharing paradigms in mobile cloud-based EHS. This analysis highlights the unique advantages of combining blockchain and IPFS, focusing on improvements in efficiency, security, and privacy. By contrasting the limitations and risks of traditional systems with our approach to emphasize the significant benefits of our framework.

6.3 Advantages of the Proposed System

A thorough analysis of our framework's security measures is essential to assess its effectiveness. We examine potential vulnerabilities, such as unauthorized data access and smart contract manipulation, and introduce targeted countermeasures to mitigate these risks. This paper highlights the framework's ability to adapt to evolving security challenges through an extensive security audit.

This work provides practical insights into safe EHR sharing by bridging theory and practice. It establishes the foundation for next studies on data security in cloud-based mobile EHS. The experiment starts by initializing the IPFS server and starting the Ethereum tool with 'start_eth.bat'. After that, the Python Flask server is launched using 'run.bat,' and the screenshots that follow demonstrate how the system processes are moving forward.

Starting Flask server...

* Running on <http://127.0.0.1:5000/>

Once in Python attendant is successfully running, open a web browser and type the next URL in the address bar: <http://127.0.0.1:9999/index>. Press Enter to navigate to the webpage shown below. This will load the interface connected to the server, marking the successful initialization of the system.

When users click the 'Patients' link, they are prompted to choose their date of birth and provide personal information. Secure data storage and retrieval is confirmed by the unique hash values generated by IPFS and blockchain.

ipfs add <file_name>

added QmTt5F2z... <file_name>

7 Conclusion

This paper introduces an innovative approach to EHR exchange by integrating blockchain technology with mobile cloud computing. Through a prototype implementation, it addresses critical challenges in existing E-Health Systems while offering practical solutions. The proposed framework leverages Ethereum blockchain on Amazon cloud infrastructure, enabling secure EHR sharing via a custom Android

Table 1 Command lines of 'start_eth' (C:\Windows\System32\cmd.exe)

C:\Windows\System32\cmd.exe.
INFO [07-16 18:52:18.143] Commit new mining work
INFO [07-16 18:52:24.788] Successfully sealed new block
INFO [07-16 18:52:24.793] block reached canonical chain
INFO [07-16 18:52:24.797] mined potential block
INFO [07-16 18:52:24.804] Commit new mining block
INFO [07-16 18:52:26.533] Successfully sealed new block
INFO [07-16 18:52:26.541] block reached canonical chain
INFO [07-16 18:52:26.549] mined potential block
INFO [07-16 18:52:26.555] Commit new mining block
INFO [07-16 18:52:28.669] Successfully sealed new block
INFO [07-16 18:52:28.676] block reached canonical chain
INFO [07-16 18:52:28.683] mined potential block
INFO [07-16 18:52:28.690] Commit new mining block
INFO [07-16 18:52:29.124] Successfully sealed new block
INFO [07-16 18:52:29.131] block reached canonical chain
INFO [07-16 18:52:29.137] mined potential block
INFO [07-16 18:52:29.149] Commit new mining block
INFO [07-16 18:52:32.896] Successfully sealed new block
INFO [07-16 18:52:32.906] block reached canonical chain
INFO [07-16 18:52:32.911] mined potential block
INFO [07-16 18:52:32.916] Commit new mining block
INFO [07-16 18:52:39.614] Successfully sealed new block
INFO [07-16 18:52:39.620] block reached canonical chain
INFO [07-16 18:52:39.625] mined potential block
INFO [07-16 18:52:39.630] Commit new mining block
INFO [07-16 18:52:39.816] Successfully sealed new block
INFO [07-16 18:52:39.821] block reached canonical chain

mobile application. By incorporating smart contracts and IPFS storage, the system ensures decentralized, efficient, and reliable data exchange. Deployment results demonstrate its superiority over conventional methods in security, scalability, and accessibility. The access control mechanism effectively safeguards network integrity, patient confidentiality and prevents unauthorized access. Comprehensive security analyses and performance evaluations confirm the framework's robustness across multiple technical dimensions. This blockchain-based solution holds wide-ranging potential, marking a major leap forward in EHR management and setting the stage for more secure and efficient mobile cloud-based systems.

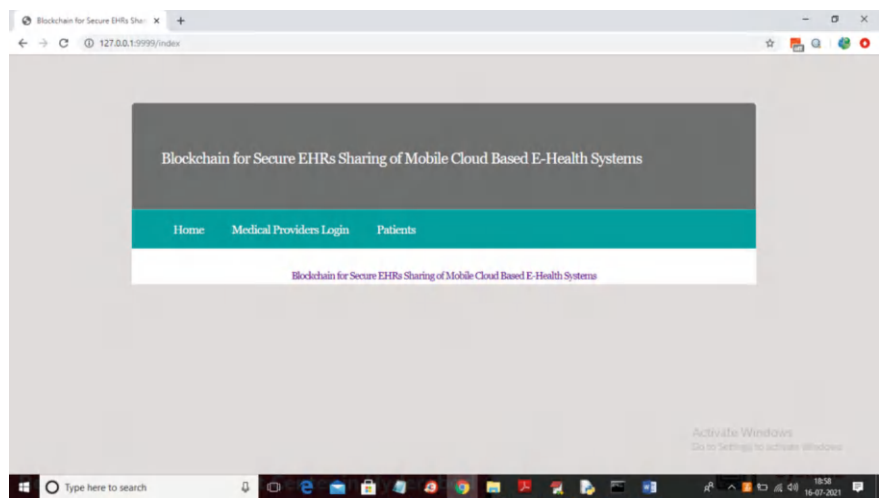


Fig. 2 Screenshot for patients

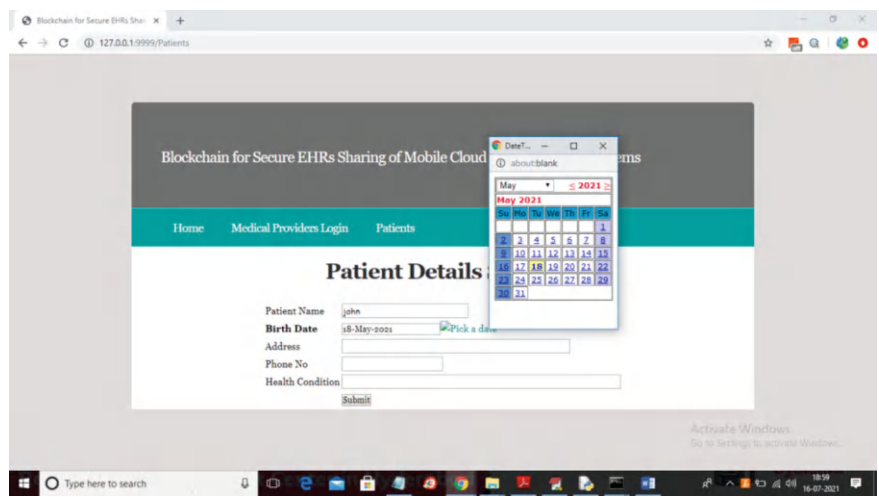


Fig. 3 Screenshot for patients details

Table 2 Test scenario specifications [30]

No.	Test scenario name	Feed	Test expectation	Captured output	Result
1	Smart contract	Execute operations on blockchain, such as storing hashcode	Blockchain constructs immutable ledgers of transactions for data-sharing system	Hashcode returned by IPFS and blockchain	Pass
2	IPFS	To store patient data	The generated hash code will be recorded on the Blockchain	Data stored blockchain	Pass
3	Data uploading	Patient can upload their data	It will be saved on both IPFS and the Blockchain	IPFS got patient details	Pass
4	Data sharing	Who has access to blockchain can be obtained hashcode	Access patient record screen	Get patient details	Pass

References

1. Mettler, M.: Blockchain technology in healthcare: the revolution starts here. In: 18th IEEE International Conference on E-health Networking, Application & Services Proceedings, pp. 1–3 (2016)
2. European Union Agency for Cybersecurity (ENISA): Blockchain security guidelines (2022). <https://www.enisa.europa.eu/publications/blockchain-security-guidelines>
3. Steichen, M., Norvill, R., Pontiveros, B.F., Shbair, W.: Blockchain-based, decentralized access control for IPFS. In: IEEE Blockchain Proceedings, pp. 1499–1506 (2018)
4. Susheela, Y., Kumar, T.R., Srinivas, M., Dhanasri, S.: Examination of the block chain for secured EHR sharing in mobile cloud-based e-health systems. *Mukt Shabd J.* **XI**(XII), 1216–1220 (2022)
5. Sonkamble, R.G., Bongale, A.M., Phansalkar, S., Dharrao, D.S.: A secure interoperable method for electronic health records exchange on the cross-platform blockchain network. *MethodsX* **13**, 103002 (2024). <https://doi.org/10.1016/j.mex.2024.103002>
6. Ettaloui, N., Arezki, S., Gadi, T.: Blockchain-based electronic health record: systematic literature review. In: *Human Behavior and Emerging Technologies*. Wiley Online Library (2024)
7. Johnson, M., Singh, J.: Secure electronic health record sharing in cloud environments. *J. Med. Inform.* **45**(3), 210–224 (2019)
8. Healthcare Information and Management Systems Society (HIMSS): Mobile healthcare technology benefits and challenges (2021). <https://www.himss.org/resources/mobile-healthcare-technology-benefits-and-challenges>
9. Smith, A., Brown: Leveraging blockchain for enhanced security in e-health systems. In: *International Conference on Health Informatics*, pp. 56–67 (2020)
10. Puthal, D., Ranjan, R.: A comprehensive study on decentralized file sharing system. *J. Comput. Syst. Sci.* **86**, 66–88 (2019)
11. Kuo, T.-T., Kim, H.-E., Ohno-Machado, L.: Blockchain distributed ledger technologies for biomedical and health care applications. *J. Am. Med. Inform. Assoc.* **24**(6), 1211–1220 (2017)
12. Li, X., Wang, H.: Enabling data sharing and privacy protection in cloud-based e-health systems: a survey. *J. Biomed. Inform.* **90**, 103104 (2019)

13. Almalki, M., Gray, K., Martin-Sanchez, F.: A cloud-based architecture for mobile health. In: 9th International Conference on Pervasive Computing Technologies for Healthcare Proceedings, pp. 160–166 (2016)
14. Garoufallou, E., Siatri, R.: A review on healthcare information sharing using blockchain. In: Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance, pp. 211–214 (2018)
15. Kumar, M., Raj, H., Chaurasia, N., Gill, S.S.: Blockchain inspired secure and reliable data exchange architecture for cyber-physical healthcare system 4.0. *Internet of Things Cyber-Phys. Syst.* **3**, 309–322 (2023). <https://doi.org/10.48550/arXiv.2307.13603>
16. Dagher, G.G., Serhrouchni, A.: Securing e-health systems using blockchain: performance evaluation and management framework. *Future Gener. Comput. Syst.* **82**, 167–176 (2018)
17. Rajasekaran, A.S., Azees, M., Al-Turjman, F.: A comprehensive survey on block-chain technology. *Sustain. Energy Technol. Assess.* **52**, 102039 (2022). <https://doi.org/10.1016/j.seta.2022.102039>
18. Fernandez-Alemán, J.L., Señor, I.C.: Security in e-health systems: a systematic literature review. *Health Inform. J.* **26**(4), 314–326 (2020)
19. IPFS Documentation: InterPlanetary file system (2023). <https://docs.ipfs.io/>
20. Patel, R., Gupta, S.: Decentralized data storage in healthcare: exploring the potential of IPFS. *Healthc. Technol. J.* **7**(4), 432–445 (2018)
21. Stojkoska, B.R., Trivodaliev, K.V.: A review of internet of things for smart home: challenges and solutions. *J. Clean. Prod.* **140**, 1454–1464 (2017)
22. Cachin, C.: Architecture of the hyperledger blockchain fabric. In: Workshop on Distributed Cryptocurrencies and Consensus Ledgers Proceedings, pp. 3–10 (2016)
23. Zheng, Z., Xie, S.: An overview of blockchain technology: architecture, consensus, and future trends. In: IEEE Symposium on Computer Science, pp. 197–203 (2017)
24. Liang, X., Zhao, J., Shetty, S., Liu, J., Li, D.: Integrating blockchain for data sharing and collaboration in mobile healthcare applications. In: 28th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications Proceedings (PIMRC), pp. 1–5 (2017)
25. Holbl, M., Kompara, M., Kamišalic, A., Zlatolas, L.N.: A systematic review of the use of blockchain in healthcare. *Symmetry* **10**(10), 470 (2018)
26. Nguyen, D.C., Pathirana, P.N., Ding, M., Seneviratne, A.: Blockchain for secure EHRs sharing of mobile cloud-based e-health systems. *IEEE Access* **7**, 66792–66806 (2019). <https://doi.org/10.1109/ACCESS.2019.2917555>
27. Wood, G.: Ethereum: a secure decentralized generalized transaction ledger. Ethereum project yellow paper 151 (2014)
28. Buterin, V.: Ethereum white paper: a next generation smart contract & decentralized application platform. Ethereum Foundation (2013). <https://ethereum.org/whitepaper>
29. Gordon, W.J., Catalini, C.: Blockchain technology for healthcare: facilitating the transition to patient-driven interoperability. *Comput. Struct. Biotechnol. J.* **16**, 224–323 (2018)
30. Kuo, T.T., Kim, H.E., Ohno-Machado, L.: Blockchain distributed ledger technologies for biomedical and health care applications. *J. Am. Med. Inform. Assoc.* **24**(6), 1211–1220 (2017). <https://doi.org/10.1093/jamia/ocx068>

AI in Social Media and Misinformation Detection

Rumor Veracity Detection in Social Networks: A Brief Survey



Shruti Bajpai and Shashank Kumar Singh

Abstract Online social networks have become a breeding ground for rumors due to faster dissemination of information to a large scale of users. Safeguarding them from rumors is critical due to potential harmful societal impacts. Using the rumor control strategies for their mitigation involves significant efforts, which can be reduced by knowing the veracity of rumors, as we need not control true rumors. Detecting the rumor veracity involves challenges like complexity in natural language communication, varying forms of rumor content and its sources, and a massive amount of content to be processed for detection. Many machines and deep learning-based models are proposed to address rumor veracity detection problems. However, these models deal with language subtleties, dependency on the context of rumor, and user purpose, further complicating veracity detection. Also, rumors keep evolving with new versions requiring continuous updates to veracity detection models. Lastly, the limited availability of labeled datasets for training models further increases the complication in veracity detection. This paper explores various machine and deep learning-based solutions proposed for rumor veracity detection tasks. We provide a brief review of proposed approaches and potential issues and challenges that can be addressed using multi-disciplinary approaches capable of integrating linguistic, social, and technical insights to improve the reliability and scalability of rumor veracity detection models.

Keywords Online social networks · Rumor · Veracity detection · Classification · Machine learning · Deep learning

S. Bajpai (✉)

Center for AI and ML, Institute of Technical Education and Research, Siksha ‘O’ Anusandhan (Deemed to Be) University, Bhubaneswar, India
e-mail: shrutibajpai@soa.ac.in

S. K. Singh

Department of Computer Science and Engineering, National Institute of Technology, Patna, India
e-mail: shashank.cs@nitp.ac.in

1 Introduction

Over the past two decades, online social networks (OSNs) have surpassed traditional media, sharing breaking news in less time and with a larger audience. This fast and high-scale dissemination of information made OSNs highly popular and, at the same time, susceptible to misuse by sharing negative information like rumors, fake news, misinformation, etc. Rumors are unverified news whose veracity status is unknown at the time of circulation [1]. Rumors are highly undesirable as they create panic, reputation threats, and economic losses [2, 3]. So, controlling them at the early stage of their circulation is essential. However, knowing when to control rumors and when not to is necessary. Several methods proposed by researchers' community work on rumor control. These methods are effective; however, they do not consider the rumor's veracity. Rumor veracity determines the truthfulness of a rumor. It can be "true" or "false" or remain "unknown" [1, 4]. Rumor veracity status can be confirmed by the concerned authorities, supportive documents, or artificial intelligence-based methods. Veracity status confirmation may take less time (in case of short-standing rumors) to more time (in case of long-standing rumors) from rumor circulation time. Knowing the veracity of a rumor helps reduce the rumor control efforts and the selection of appropriate strategies. For example, if a rumor's veracity is true, there is no need for rumor control. However, we have to use appropriate strategies to control false rumors. In cases where the veracity of a rumor remains unknown, there is a desire to know whether the rumor is true or false. So, finding the veracity of a rumor in circulation is significant to safeguarding user interests on social networks.

Rumor veracity detection is a classification problem consisting of true, false, or unknown classes. Formally, given a social media post or piece of information I , shared by a user or group of users in a social network $G = (V, E)$ where V represents the set of users and E represents the relationships between users, the objective is to determine the veracity v of I such that.

$$v \in \{\text{True, False, Unverified}\}$$

Traditionally, rumor veracity is confirmed by the concerned authority. However, lack of trust in authorities and problems in manual clarification in real time arises a need for automatic detection of rumor veracity. Detecting and managing rumor veracity is a complex task. Rumors spread quickly due to various factors. These include the content of the rumor, which is deliberately made attractive or controversial to gain the attention of users; user features like influence, past behavior, structure of the network, and diffusion pattern; and temporal features like rate of sharing, decay over time, etc. Many machine [5–8] and deep learning-based [9–13] models have been proposed, considering these factors as the features. The objective of these proposed models is to construct a function $f: (U, C, N, T) \rightarrow v$ that predicts the veracity v of the information I where C is a set of content features, U is a set of user features, N is set of network propagation based features, and T is set of temporal features.

Rumor veracity detection on social media is hampered by serious challenges in that online content is informal and context-dependent, rendering automated interpretation challenging. Rumors are dynamic and change as they spread, thus static models for detecting them become outdated over time. Moreover, the lack of labeled datasets, which are expensive and time-consuming to create, restricts machine learning models' reliability and generalizability. In order to overcome these challenges, scholars have attempted using machine learning and natural language processing (NLP) to determine linguistic patterns and network-based methodologies to examine information diffusion and believability. Although these strategies hold promise, they tend to lack precision and scalability, leading to the development of hybrid models that combine multiple approaches.

This paper provides a brief survey of existing studies, highlighting their contributions and identifying persistent shortcomings in addressing these challenges in rumor veracity detection on social networks. The paper is divided into five sections. Section 2 discusses the existing surveys and their limitations. Section 3 briefly overviews existing work from machine learning and deep learning perspectives. Section 4 discusses the issues and challenges in detecting rumor veracity, and Sect. 5 concludes this brief review paper.

2 Related Work

Rumor veracity detection on social networks has drawn significant attention from the researchers' community in the past few years. Several studies are there that have reviewed the rumor veracity detection problem and approaches to solving it. This section provides a brief overview of a few survey papers related to ours and their key contributions.

In [14], authors established the need to determine the veracity of rumors using a case study of a Twitter dataset of 330 threads related to nine breaking news events. Their main findings established that true rumors get resolved faster, within approximately 2 h, than false rumors, which take approximately 14 h. Also, users tend to support unverified rumors in case of a lack of counter-evidence. With a focus on the necessity of real-time machine learning models to assess rumor veracity, this work presents a thorough annotation technique to examine rumor dynamics. In [15], authors review how rumors spread on platforms such as Twitter and Facebook, focusing on their drivers, such as ambiguity, anxiety, and lack of credible sources, and their dynamics using psychological, sociological, and epidemiological models. It explores detection techniques, including machine learning-based approaches and features like user behavior, content ambiguity, and network structure, while highlighting anti-rumor strategies such as decentralized models and leveraging influential users or authoritative information. This study lists various challenges like detecting rapidly evolving or context-specific rumors, reliance on predefined datasets for machine learning models, and the inherent difficulty of combating mistrust in centralized authorities. Furthermore, the focus on textual data overlooks the increasing role

of multimedia content in rumor propagation, emphasizing the need for more holistic, automated, and real-time solutions. In [16], a comprehensive study of rumor detection and verification on social media is presented, focusing on multimedia data (text and images). It outlines data collection strategies, key features for analysis, and methods using traditional machine learning, deep learning, and hybrid approaches. The study introduces a thematic taxonomy for rumor analysis, emphasizing the importance of content and user-based features, and provides a detailed review of state-of-the-art datasets and methodologies. Key contributions include insights into the utility of features like sentiment analysis, propagation patterns, and visual content, alongside exploring both text and image-based detection frameworks. Integrating diverse data types and user behavior modeling.

In this paper, we present a brief survey on the recent advancements in the field of rumor veracity detection in social networks.

3 Approaches to Rumor Veracity Detection

Different methods for detecting rumor veracity have been proposed by researchers. These approaches primarily address natural language processing, machine learning, deep learning, and traditional methods and are aimed at enhancing the performance and accuracy of determining rumor veracity on social media. These are discussed in the subsequent subsections.

3.1 *Machine Learning-Based Rumor Veracity Detection Using Shallow Classifiers*

Kumar et al. [5] discuss rumor analysis by suggesting an optimized learning model for classifying real-time tweets. With almost 14,000 tweets on mob lynching cases in India (#moblynching), tweets were classified as true, false, or unspecified using five classical classifiers and 13 features. The incorporation of Particle Swarm Optimization (PSO) for feature selection enhanced classification accuracy, attaining an average gain in accuracy of 11.28%, reducing features by 31%, and a maximum accuracy of 96.15% using the decision tree classifier. This shows how PSO optimizes rumor veracity classification. Dang et al. [6] target examining, finding, and categorizing rumors within social media truthfulness in the initial phase. The study generates a dataset classifying rumors into five levels of truth: “False,” “Mostly False,” “Half True,” “Mostly True,” and “True,” based on features including topics, user sentiment, and network structures. Based on visualization and comparison of these attributes, the work builds a theory-based feature set based on psychology and social science theories for the classification of rumors. The technique was successful in identifying rumor veracity within a week, being a valuable utility for fact-checking websites like

Snopes.com and Politifact.com to automatize rumor checking and train individuals for decision-making.

Dungs et al. [17] use stance as another feature on top of features that have previously been used in the literature. They used HMM variants and collective stance for modeling rumor veracity. The results indicate that HMMs with stance and tweet timestamps as the only features for modeling true and false rumors have F1 scores of approximately 80%.

Giasemidis et al. [8] work builds an independent message classifier to screen trustworthy information from Twitter with an emphasis on detecting and examining rumors. The model classifies 72 rumors (41 true, 31 false) out of 100 million tweets based on 80 + measures of trustworthiness, such as user profiles and tweet content. It surpasses current approaches and gives insights into rumor evolution through time windows. A demonstration software with a graphical user interface was even developed so that users could navigate through the analysis.

3.2 Deep Learning-Based Rumor Veracity Detection

Deep learning techniques have come to serve as effective mechanisms for rumor veracity detection based on their capability to automatically learn and extract sophisticated patterns from huge volumes of data. The techniques take advantage of neural network frameworks in order to process textual, contextual, and structural information, representing a superior choice compared to classical machine learning models.

Islam et al. [9] introduced the RumorSleuth model, which uses a multitask deep learning method that takes advantage of textual content and user profile data to predict rumor veracity and user stances. Experiments on publicly available datasets show that RumorSleuth performs better than current methods, with up to a 14% increase in veracity classification and a 6% increase in instance classification. Kim et al. [18] introduce a double-channel structure to classify social media rumors as true, false, or unverifiable. The approach first categorizes rumors into informed (certain) and uninformed (uncertain) types. Lie detection algorithms are applied to informed rumors, while thread-reply agreement detection is used for uninformed rumors. Using the SemEval 2019 Task 7 dataset, which involves an ex-ante threefold classification of rumors, the model achieved a macro-F1 score of 0.4027, surpassing baseline models and the second-place competitor. Additionally, empirical results demonstrate the superiority of the double-channel structure over single-channel approaches using either lie detection or agreement detection alone.

Na et al. [10] propose the Multi-task Attention Tree Neural Network (MATNN) for simultaneous rumor stance classification and rumor veracity detection. Employing Regular Rumor Conversation Trees (RC-Trees) to organize rumor conversations, MATNN applies Tree Self-Attention for local feature extraction and Tree Convolution with Tree Pooling for global features. Experimental results demonstrate that RC-Trees improve performance, and MATNN outperforms existing state-of-the-art

approaches in both stance classification and veracity detection. Vosoughi et al. [4] propose the Multi-task Attention Tree Neural Network (MATNN) for simultaneous public stance classification on rumors and rumor veracity detection. Employing Regular Rumor Conversation Trees (RC-Trees) to organize rumor conversations, MATNN applies Tree Self-Attention for local feature extraction and Tree Convolution with Tree Pooling for global features. Experiments verify that RC-Trees improve performance, and MATNN outperforms existing state-of-the-art approaches in stance classification and veracity detection.

Rosenfeld et al. [19] investigate misinformation detection in terms of stable patterns of propagation and are free from text and user identity, which can easily be manipulated. By analyzing the patterns of diffusion of unverified rumors on social media, the research proves that information diffusion topology can be an effective indicator of truth. Using graph kernels for the extraction of general topological properties of Twitter cascades, the research trains predictive models independent of language, user identity, and time. The results prove that aggregate sharing patterns give stable indications of rumor truth or falsity, even at early dissemination.

Chua and Banerjee [7] explore the role of language in predicting the truthfulness of online rumors based on six linguistic features: comprehensibility, sentiment, time orientation, quantitative information, writing style, and topic. Based on a dataset of 2,391 rumors on Snopes.com, 20% of which are true and the rest false, the study uses the Linguistic Inquiry and Word Count (LIWC) software to analyze linguistic features. Binomial logistic regression is employed to analyze the data. The findings indicate that comprehensibility, time orientation, writing style, and topic are the most significant predictors of the truthfulness of a rumor. Kumar and Carley [20] introduce binarized constituency trees to represent social media discussions to model, with the ability to compare source posts and responses efficiently. The model learns patterns from both posts by incorporating convolution units to Tree LSTMs. The model propagates stance signals for rumor classification using multi-task learning (stance + rumor). It outperforms existing methods with a 12% improvement in rumor truthfulness classification and a 15% improvement in stance classification based on F1-macro scores.

Singh et al. [21] propose an Attention-based Long Short-Term Memory (LSTM) network to detect and distinguish rumors from non-rumor tweets using 13 linguistic and user features. The model outperforms conventional machine and deep learning models, achieving an F1-score of 0.88, surpassing state-of-the-art results. By accurately identifying rumors, the system aims to reduce their societal impact and mitigate harm, enhancing trust in social media platforms. Also, Wei et al. [11] present a hierarchical multi-task learning framework for predicting rumor stances and veracity on Twitter. It uses a graph convolutional network to classify tweet stances and analyzes temporal stance evolution to predict rumor veracity. Experimental results show that the proposed method outperforms previous stance classification and veracity prediction approaches.

Liu and Wu [22] introduce a new Multi-Task Learning framework with Shared Multichannel Interactions (MTL-SMI), which consists of two shared channels for learning task-invariant text and structural features and two task-specific graph

channels for structural feature enhancement. The method enhances representation learning, and experimental evidence on two real-world datasets indicates that MTL-SMI outperforms competitive baselines. Poddar et al. [23] present a neural network model that makes use of users' stances in Twitter discussions to identify rumor veracity. It has two steps: the first is identifying the stance of every tweet on the basis of its content, timestamp, and discussion sequence, and the second is utilizing these stances to identify the veracity of the initial rumor. Performance assessments on the SemEval 2017 data reveal that their system performs better than current methods for stance and rumor veracity prediction tasks.

Khandelwal [13] presents a multi-task learning model that can predict the stances and veracity of rumors. The model consists of two parts: one that predicts the stance of each post in a conversation thread with multi-turn interactions and another that predicts the rumor's veracity given the trajectory of stances. Experimenting with the SemEval 2019 Task 7 dataset, including rumors in Twitter and Reddit, proves that the proposed technique achieves better results in stance classification and rumor veracity prediction than existing methods. In [24], authors proposed a Rumor Detection Neural Network (RDNN) based on deep learning to predict whether a tweet is a rumor. The network comprises three layers: AttCNN to extract features, AttBi-LSTM for context and semantic interpretation, and HPOOL for the aggregation of pooled feature maps. Trained on Kaggle and #gaja datasets, the RDNN achieves high accuracy, with results of 93.24% and 95.41% in identifying rumor tweets in real-world events. This approach outperforms conventional machine learning techniques, addressing their limitations in tuning and learning. Table 1 provides a summary of the reviewed papers.

4 Issues and Challenges

Despite significant advances in rumor veracity detection, multiple issues and challenges persist in accurately and efficiently identifying the veracity of rumors on OSNs. These challenges arise from the inherent complexity of online information, the dynamic nature of rumors, and limitations in current technological and methodological approaches. This section lists a few challenges faced in the field.

1. *Language Complexity and Context Dependency*—Social media language is often informal, ambiguous, and context-dependent. The use of slang, abbreviations, emojis, and culturally specific references complicates NLP tasks, making it difficult to interpret the veracity of rumors accurately.
2. *Evolving Nature of Rumors*—Rumors on social networks are not static; they evolve as they spread, with new variations and contextual twists emerging over time. This dynamic characteristic of rumors requires detection models to continuously adapt to new versions and shifting narratives.

Table 1 Overview of various machine and deep learning-based approaches for rumor veracity detection in social networks

Author	Dataset	Features	Model/Classifier	Approach
[5]	Twitter dataset (14 K tweets)	Content-based, semantic, network-specific features	SVM, DT, K-NN, NB, NN	Feature selection using PSO
[6]	Reddit dataset (88 rumors)	Entropy score, sentiment, network structure, TF-IDF, social science-based features	Naive Bayes	Grounded in social science theories about rumor spread
[18]	SemEval 2019 Task 7	–	BERT-based certainty classifier	Double-channel approach for lie detection
[9]	PHEME, Twitter15, Twitter 16	Textual and user features	RNN, VAE	Combination of text and user profiles
[10]	SemEval, PHEME	Tree-based local structural features	Multi-task attention tree neural network	Multi-task attention mechanism
[4]	209 rumors (938,806 tweets)	Linguistic style, user characteristics, network propagation dynamics	Dynamic time wrapping, hidden markov model	Network propagation analysis
[19]	209 rumors (938,806 tweets)	–	Weisfeiler–Lehman graph kernel	Graph-based rumor detection
[7]	2,391 rumors (20% true) from Snopes.com	Comprehensibility, sentiment, time orientation, quantitative details, writing style, topic	Binomial logistic regression	Analysis based on rumor verification website
[20]	PHEME (5 events)	BERT, GloVe, SkipThought, DeepMoji embeddings	Tree LSTMs	Convolutional units in tree LSTMs
[25]	PHEME (tweets for 5 events)	Linguistic and user features	Attention-based LSTM, PSO	Deep learning combined with feature extraction
[8]	100 M tweets, 72 rumors	Trustworthiness measures, authors' profile, social network connections, tweet content	Logistic regression, SVM, random forest, DT, NB, NN	Trust-based evaluation
[11]	SemEval-2017 task 8, PHEME	Structural characteristics, temporal dynamics	Graph convolutional network	Novel graph-based structural property analysis

(continued)

Table 1 (continued)

Author	Dataset	Features	Model/Classifier	Approach
[13]	SemEval 2019 RumorEval	Stylistic, structural, conversational, affective, emotional, speech-act features	Longformer transformer	Multi-turn conversational modeling
[12]	PHEME, SemEval	Task-invariant text and structural features, posts, comments, users, interaction graphs	Multi-task learning framework	Shared multichannel interactions
[23]	SemEval2017	Textual content, conversation context	CNN, RNN	Context-based rumor detection
[16]	Events: Charlie Hebdo, Ferguson, Germanwings, Ottawa, Sydney (173 rumors)	Stance, time	Hidden Markov model	Time and stance-based evaluation
[26]	Kaggle dataset	Semantic or contextual information	Detection neural network (AttCNN layer)	Combining contextual and semantic analysis

3. *Data Scarcity and Bias in Training Sets*—Large-scale labeled datasets are required for training machine and deep learning models for veracity detection. However, obtaining and labeling datasets is a resource-intensive task. Also, available datasets are often limited to specific topics or regions or events which can lead to poor performance when models are applied to new or less-represented contexts like emerging events.
4. *Dynamic nature of OSNs*—Social networks have unique structural characteristics, such as echo chambers and highly interconnected communities, where rumors can spread more easily and reinforce itself. Also, social networks are dynamic in nature making veracity detection challenging, as it requires models that can analyze continuously changing interactions between users, as well as the influence of key individuals or communities in propagating information.
5. *Computational Requirements and Real-Time Detection*—Machine and deep learning-based models for rumor veracity detection are computationally intensive as they require significant processing power and memory. This makes it challenging to implement such models in real-time applications, where rapid detection is crucial to prevent the spread of rumors.
6. *Ethical and Privacy Concerns*—Detection of rumor veracity also poses ethical and privacy issues because automated tools that scan user-provided content might unintentionally breach individual privacy, particularly if they scan confidential discussions or private information. Second, any kind of bias in detection algorithms will lead to unequal censorship or mistaken identification of lawful

content. Satisfying these ethical issues means the proper design of algorithms as well as compliance with data protection laws.

7. *Explainability and Trust in Detection Models*—As deep learning models become increasingly common in veracity detection, the issue of explainability increases. Deep neural networks and other complex models are typically considered “black boxes,” providing little information on how they arrive at their conclusions. This transparency issue can prevent user trust and complicate developers’ ability to detect and fix errors in the models.

5 Conclusion

Determining the veracity of rumors on social networks remains a challenging project with many aspects that require further improvements in technique and technology. The effectiveness of contemporary fashions remains constrained by way of problems consisting of language complexity, converting rumor narratives, statistics scarcity, and computing wishes, no matter the truth that gadget and deep learning based totally and hybrid techniques have made big advances. Furthermore, moral factors like explainability and privateness emphasize the significance of creating obvious and reliable systems. Creating flexible and real-time detection fashions can be critical as rumors preserve converting and have a giant effect on society. Further, multidisciplinary studies that integrate knowledge from synthetic intelligence, ethics, and social technology hold the capability to increase robust structures that can slow the spread of rumors, fostering greater knowledge and resilient social media.

References

1. Zubiaga, A., Aker, A., Bontcheva, K., Liakata, M., Procter, R.: Detection and resolution of rumours in social media: a survey. *ACM Comput. Surv.* **51**(2)
2. Inayat, A.: Impacts of rumors and conspiracy theories surrounding COVID-19 on preparedness programs. *Disaster Med. Public Health Preparedness* 1–6 (2020). <https://doi.org/10.1017/dmp.2020.325>
3. Samia, T., Hossain, M.M., Mazumder, H.: Impact of rumors and misinformation on covid-19 in social media. *J. Prevent. Med. Public Health* **53**(3) (2020). <https://doi.org/10.3961/jpmph.20.094>
4. Vosoughi, S., Mohsenvand, M.N., Roy, D.: Rumor gauge: predicting the veracity of rumors on twitter. *ACM Trans. Knowl. Discov. Data* **11**(4) (2017). <https://doi.org/10.1145/3070644>
5. Kumar, A., Sangwan, S.R., Nayyar, A.: Rumour veracity detection on twitter using particle swarm optimized shallow classifiers. *Multimedia Tools Appl.* **78**, 24083–24101 (2019). <https://doi.org/10.1007/s11042-019-7398-6>
6. Dang, A., Moh’d, A., Islam, A., Milios, E.: Early detection of rumor veracity in social media. In: *Proceedings of the Hawaii International Conference on System Sciences* (2019)
7. Alton Yeow-Kuan, C., Snehasish, B.: Linguistic predictors of rumor veracity on the internet (2016). <https://api.semanticscholar.org/CorpusID:51691121>
8. Giasemidis, G., Singleton, C., Agraftiotis, I., Nurse, J.R.C., Pilgrim, A., Willis, C., Greetham, D.V.: Determining the veracity of rumors on Twitter. In: *Social Informatics*, pp. 185–205 (2016)

9. Islam, M.R., Muthiah, S., Ramakrishnan, N.: Rumorsleuth: joint detection of rumor veracity and user stance. In: Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 131–136 (2020). <https://doi.org/10.1145/3341161.3342916>
10. Na, B., Fanrong, M., Xiaobin, R., Zhixiao, W.: A multi-task attention tree neural net for stance classification and rumor veracity detection. *Appl. Intell.* **53**(9) (2023). <https://doi.org/10.1007/s10489-022-03833-5>
11. Wei, P., Xu, N., Mao, W.: Modeling conversation structure and temporal dynamics for jointly predicting rumor stance and veracity (2019). <https://arxiv.org/abs/1909.08211>
12. Liu, Y., Yang, X., Zhang, X., Tang, Z., Chen, Z., Zheng, L.: Predicting rumor veracity on social media with cross-channel interaction of multi-task. *Neural Comput. Appl. Comput. Appl.* **36**(15), 8681–8692 (2024). <https://doi.org/10.1007/s00521-02409519-y>
13. Khandelwal, A.: Fine-tune longformer for jointly predicting rumor stance and veracity. In: Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data, pp. 10–19 (2021). <https://doi.org/10.1145/3430984.3431007>
14. Lalehparvaran, P.: Debunking the rumor: a review of veracity. *Int. J. Appl. Sci. Technol.* **8**(3) (2018). <https://doi.org/10.30845/ijast.v8n3p9>
15. Mohammad, A., Madhu, K., Sharma, T.P.: Rumors detection, verification and controlling mechanisms in online social networks: a survey. *Online Social Netw. Media* (2019). <https://doi.org/10.1016/j.osnem.2019.100050>
16. Varshney, D., Vishwakarma, D.K.: A review on rumour prediction and veracity assessment in online social network. *Expert Syst. Appl.* **168**, 114208 (2021). <https://doi.org/10.1016/j.eswa.2020.114208>
17. Dungs, S., Aker, A., Fuhr, N., Bontcheva, K.: Can rumor stance alone predict veracity? In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 3360–3370 (2018)
18. Kim, A.G., Yoon, S.: Detecting rumor veracity with only textual information by double-channel structure. In: Proceedings of the Tenth International Workshop on Natural Language Processing for Social Media (2022). <https://doi.org/10.18653/v1/2022.socialnlp-1.3>
19. Rosenfeld, N., Szanto, A., Parkes, D.C.: A kernel of truth: determining rumor veracity on Twitter by diffusion pattern alone. In: Proceedings of the Web Conference 2020, pp. 1018–1028 (2020). <https://doi.org/10.1145/3366423.3380180>
20. Kumar, S., Carley, K.: Tree LSTMs with convolution units to predict stance and rumor veracity in social media conversations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5047–5058 (2019). <https://doi.org/10.18653/v1/P19-1498>
21. Singh, S., Verma, S.K., Tiwari, A.: A novel approach for finding crucial node using ELECTRE method. *Int. J. Mod. Phys. B* **34**(09), 2050076 (2020). <https://doi.org/10.1142/S0217979220500769>
22. Liu, X., Wu, Z.: Rumor detection on social media based on network structure and content understanding. *Knowl.-Based Syst.-Based Syst.* **146**, 207–216 (2018)
23. Poddar, L., Hsu, W., Lee, M.L., Subramaniam, S.: Predicting stances in Twitter conversations for detecting veracity of rumors: a neural approach. In: 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 65–72 (2018). <https://doi.org/10.1109/ICTAI.2018.00021>
24. Suthanthira Devi, P., Karthika, S.: RDNN: Rumor detection neural network for veracity analysis in social media text. *KSII Trans. Internet Inf. Syst.* **16**(12), 3868–3888 (2022). <https://doi.org/10.3837/tiis.2022.12.005>
25. Singh, J.P., Kumar, A., Rana, N.P., Dwivedi, Y.K.: Attention-based LSTM network for rumor veracity estimation of tweets. *Inf. Syst. Front.* **24**(2), 459–474 (2022). <https://doi.org/10.1007/s10796-020-10040-5>
26. Ruchansky, N., Seo, S., Liu, Y.: CSI: a hybrid deep model for fake news detection. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 797–806 (2017)

Techniques for Detecting False Information on Social Media to Strengthen Cybersecurity



Prabhat Kumar Sahu, Smita Rath, Alakananda Tripathy,
Rashmi Rani Patro, and Sangam Malla

Abstract The rapid growth of digital platforms has revolutionized information dissemination, but it has also facilitated the spread of fake news, often amplified through biased journalism and social media. This paper addresses the challenge of fake news detection by proposing a robust methodology that categorizes news articles as genuine or fraudulent. Leveraging advanced natural language processing (NLP) techniques, such as TF-IDF and Bag of Words (BoW), the study employs six machine learning models—Logistic Regression, Support Vector Machines (SVM), Random Forests, Naïve Bayes, Decision Trees, and Neural Networks—to classify news articles. The study presents innovative feature extraction techniques, such as the Linguistic Inquiry and Word Count (LIWC2015) tool, to capture critical linguistic properties, including sentiment ratios and grammatical parts. Additionally, dimensionality reduction techniques like t-SNE and PCA enhance model performance by reducing noise and maintaining global patterns. The findings show that SVM, MLP, and logistic regression classifiers continuously produce high recall, accuracy, precision, and F1 scores, especially when adding speaker and party information. At the same time, Naïve Bayes performs poorly because it cannot handle the complexity of the dataset. The study places a strong emphasis on privacy and cultural diversity as ethical factors in the detection of fake news. The results show that the reliability of false news identification can be significantly increased by integrating complex feature sets with cutting-edge machine learning algorithms.

P. K. Sahu (✉) · S. Rath · A. Tripathy
Siksha 'O' Anusandhan (Deemed to Be) University, Bhubaneswar, India
e-mail: prabhatsahu@soa.ac.in

S. Rath
e-mail: smitarath@soa.ac.in

A. Tripathy
e-mail: alakanadatritpathy@soa.ac.in

R. R. Patro
GITA Autonomous College, Bhubaneswar, India

S. Malla
Udayanath Autonomous College of Science and Technology, Cuttack, India

Keywords Natural Language Processing (NLP) · TF-IDF · Bag of Words (BoW) · Logistic regression · Support Vector Machines (SVM) · Random forest · Naïve Bayes · Decision trees · Multi-Layer Perceptron (MLP) classifier · t-SNE · PCA

1 Introduction

Social media has transformed information intake in the past decade, presenting significant prospects and trials. The extent of fake news is predominantly concerning due to its profound impact on society, including democracy, business, politics, education, and the economy. While fake news is not a new issue, its incidence has rushed in the digital age, driven by the extensive use of social media as a primary news source. Research recommends that people are more likely to trust false information over confirmed facts, which can lead to mix-ups and grind down public confidence in reliable sources [1]. The spread of fake news has negative consequences for businesses, politics, and society. Politically, it can harm statuses, mix sadness, and influence election results. For companies, misinformation can lead to monetary fatalities, including stock price drops, reputational harm, and operational disturbances, which are often hard to converse [2]. In the education sector, false information can demoralize public trust in educational institutions and upset their integrity [3]. On a social level, experience with ambiguous information can extend societal divisions, increase divergence, and grind down trust in government and media institutions, emphasizing the urgency of fighting fake news effectively.

Technological expansions in computer science have both enabled the rise of misinformation and provided powerful trappings to detect and counter it. The revolutions have made creating and spreading unfair content easier; they have developed exclusive methods for recognizing false information. Artificial intelligence (AI), mainly machine learning (ML) and deep learning, has shown considerable potential in this area. Natural Language Processing (NLP), a field within computer science, plays a vital role in identifying misleading information. Different techniques like sentiment analysis, linguistic pattern recognition, and semantic anomaly detection help analyze discrepancies in misleading news. Additionally, to assess the reliability of news articles, methods like contextual analysis, called entity recognition, and coherence evaluation are used [4, 5].

Yet, undertaking the misleading information issue requires technological solutions. Educators are essential in promoting digital mastery and critical intelligence, empowering folks to examine the information sensitively. Policymakers must present regulations to check that social media platforms are detained and liable for spreading counterfeit news. Moreover, to verify the accuracy of the content, individuals must examine it before sharing it. To effectively reduce the fake news that may hamper the societal impact, it is vital to use a multi-stakeholder methodology involving technology, education, and regulation.

2 Related Work

A summary of the recent research on fake news detection, highlighting various approaches and key contributions, is delivered in Table 1. A comparative analysis of machine learning (ML) and deep learning (DL) techniques, emphasizing the importance of feature engineering in traditional ML models and contextual analysis in DL methods, was performed by Luo et al. (2022). Singh et al. (2023) proposed a deep learning framework combining CNNs and LSTMs, improving detection accuracy through advanced preprocessing. Zhou and Zafarani (2020) categorized fake news detection strategies into four groups, stressing the need for interdisciplinary collaboration. Other studies, such as those by Hamadouche et al. (2024) and Karaoğlu (2024), extended the research focus to region-specific models, incorporating cultural and linguistic insights. Furthermore, Hashmi et al. (2024) developed a hybrid deep learning model combining FastText embeddings with Explainable AI, improving accuracy and interpretability. Yuan et al. (2021) addressed domain-specific biases with a graph-attention neural network, while Dua et al. (2023) introduced an interpretable framework using LIME and SHAP for transparent model predictions. Table 1 shows the studies by authors and their approaches and contributions. Finally, traditional methods, such as Naive Bayes and Logistic Regression, were explored by Yang (2018) and Shete et al. (2021) show the effectiveness of lightweight models for resource-constrained environments. These studies collectively reflect the wide range of methodologies used to address the growing challenge of fake news detection.

3 Problem Statement and Proposed Methodology

Rapid digital platform growth has changed how people obtain information, but it has also aided in spreading false information, primarily through social media and biased journalism. As shown in Fig. 1, this study separates real and fake social media news pieces to address problems like algorithmic bias and dataset limitations.

Textual input is transformed into numerical form using sophisticated natural language processing (NLP) techniques, such as Bag of Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF). Training and evaluation are done on six machine learning models: Support Vector Machines, Random Forests, Logistic Regression, Naïve Bayes, Decision Trees, and Neural Networks. Random Forest is the most accurate; however, deep learning models perform better because of dimensionality reduction strategies like Principal and t-SNE. These technologies guarantee reliable feature extraction by minimizing noise and maintaining global data patterns. The paper also emphasizes ethical considerations, including privacy and cultural diversity, which are critical for the effective and responsible detection of fake news.

In the proposed methodology, we aim to identify authentic and fake news articles across multiple domains by introducing novel ensemble learning techniques alongside advanced linguistic feature extraction methods. Using three datasets, including

Table 1 Related wok comparison

Study	Approach	Key contribution
Luo et al. [6]	Comparative analysis of ML and DL	Feature engineering in ML, contextual analysis in DL
Singh et al. [7]	CNN and LSTMs-based approach	Improved detection accuracy with pre-processing
Zhou and Zafarani [8]	Fake news detection strategies classification	Categorical detection strategies leveraging blog-based, style-based, content-based features
Hamadouche et al. [9]	Region-specific models for Arabic and Algerian fake news	Highlighted the need for culturally adapted models
Karaoğlu [10]	Attention-based approach with hybrid machine learning	Improved classification with contextual neural language models
Hashim et al. [11]	Hybrid deep learning model with FastText	Enhanced accuracy and interpretability with Explainable AI
Yuan et al. [12]	Interpretative framework using LIME and SHAP	In-depth feature analysis and model explanation
Dua et al. [13]	Multimodal NLP with Bi-LSTM	Combining multiple modalities for improved detection
E Alsuwat et al. [14]	Automatic Q&A approach	Simplicity and efficiency for fake news filtering
Yang [15]	NLP techniques with Logistic Regression	Lightweight models for classification in resource-constrained environments
Shete et al. [16]	NLP techniques with Logistic Regression	Lightweight models for classification in resource-constrained environments

the ISOT Fake News Dataset and two datasets from Kaggle, we preprocess the textual data by filtering out irrelevant attributes like author information, dates, URLs, and categories. Articles with less than 20 words in the body text or no body content are excluded to ensure consistency. The datasets undergo comprehensive preprocessing to convert multi-column articles into a single-column format, followed by linguistic feature extraction. The Linguistic Inquiry and Word Count (LIWC2015) tool is employed to extract 93 discrete and continuous features, such as sentiment ratios, punctuation frequency, and the usage of grammatical elements. Numerical scaling is applied to normalize feature values, bringing them into a range of 0–1, which improves model performance by ensuring feature uniformity.

Machine learning models are trained on these features using a 70/30 train-test split, ensuring an even distribution of authentic and fake articles in each subset. Hyperparameter tuning via grid search is employed to achieve optimal model configurations, balancing bias and variance. While computationally intensive, this process prevents overfitting and underfitting.

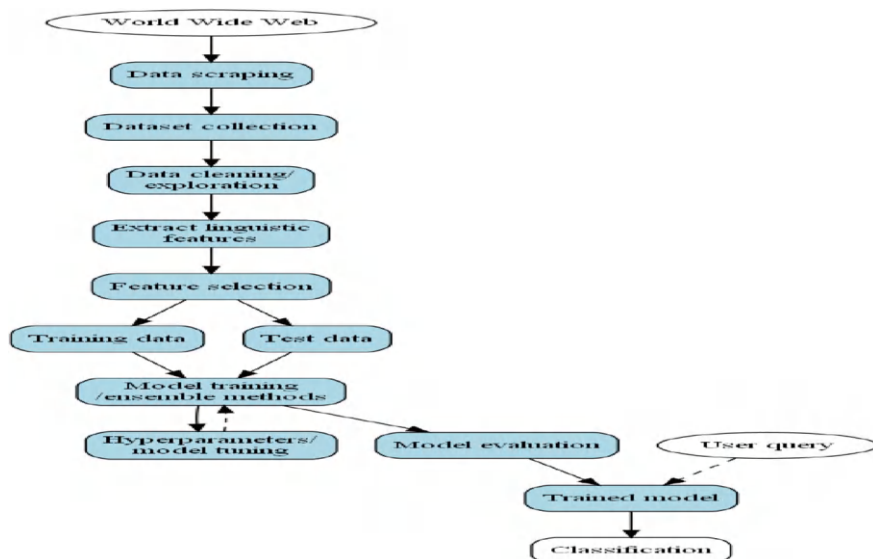


Fig. 1 Block diagram of proposed methodology

The study introduces ensemble methods to enhance classification performance, including bagging, boosting, and voting classifiers. Two distinct voting classifiers are designed: the first combines logistic regression, random forest, and KNN; the second integrates logistic regression, linear SVM, and CART. Boosting methods, such as XGBoost and AdaBoost, and a bagging ensemble with 100 decision trees are also implemented. Each ensemble model is validated using 10-fold cross-validation for robust performance evaluation. This approach leverages diverse algorithms and linguistic insights to enhance fake news detection across varied datasets.

4 Materials and Methodology

We propose a framework leveraging ensemble techniques combined with the Linguistic Inquiry and Word Count (LIWC) feature set to classify news as genuine or fake. This innovative approach enhances classification accuracy across various domains, enabling robust detection of news legitimacy by integrating advanced linguistic analysis and ensemble methodologies.

4.1 Data Preprocessing

The raw data is meticulously prepared for machine learning models to preprocess datasets for categorizing false news. The three primary datasets—the ISOT Fake News Dataset and two publicly accessible datasets from Kaggle—are sourced from the World Wide Web. Real and false news items from various industries, including politics, entertainment, sports, and technology, may be found in these databases. Every dataset is carefully cleaned to guarantee dependability and consistency.

First, irrelevant information is eliminated, including categories, author names, posting dates, and URLs. Articles with less than 20 words or no body text are rejected to eliminate noisy samples. To standardize the structure, multicolumn articles are transformed into single-column formats. Relevant qualities are chosen for additional processing after cleaning. Statement IDs, labels, topics, speaker information, party membership, and other statement credibility metrics are among the 12,788 rows and 16 columns that comprise the LIAR dataset. Training (10,239 rows), validation (1,283 rows), and testing (1,266 rows) comprise this dataset. The ratio of favorable to negative words, punctuation, and grammatical elements like verbs and adjectives are among the extracted linguistic characteristics. This feature extraction process converts textual attributes into numerical values using tools like LIWC2015, which generates 93 distinct features. To ensure that all feature values fall between 0 and 1, scaling is used to standardize the data. This step is crucial for consistency because feature ranges (such as percentage values vs word counts) vary considerably. The final preprocessed data is divided into 70% training and 30% testing groups to guarantee a fair distribution of actual and false articles. This uniform preprocessing procedure is the basis for reliable model training and precise phony news identification.

4.2 Feature Selection and Text Representation Techniques

The feature selection process aimed to convert raw text data into numerical forms suitable for machine learning tasks. Initially, text inputs were tokenized using Natural Language Processing (NLP) techniques, breaking the text into individual components. These tokens were then converted into numerical features using Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BOW). The BOW method represents text as a fixed-length vector, where each position corresponds to a distinct word, and the value at each position indicates the word's frequency within the document. While BOW is adequate for basic text classification, it does not account for the order or context of words, leading to sparse and high-dimensional vectors. Label encoding was applied to transform text-based categories into numerical values to handle categorical data such as political affiliation or speaker identity, ensuring compatibility with machine learning models. Furthermore, word clouds were generated to visualize word frequencies, highlighting dominant terms and facilitating exploratory data analysis. By systematically converting unstructured text

into structured numerical formats through BOW, label encoding, and visualization techniques, the groundwork was laid for developing accurate and efficient machine-learning models. This process underscores the significance of effective preprocessing in leveraging textual data for classification tasks.

- Term Frequency-Inverse Document Frequency (TF-IDF):

TF-IDF builds upon the BOW model by aiming to emphasize words that are unique and informative within a document while reducing the significance of standard, less meaningful terms. The first step in TF-IDF involves calculating the Term Frequency (TF), as shown in Eq. (1), which measures how frequently a word appears in a document:

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}} \quad (1)$$

The next step involves calculating the Inverse Document Frequency (IDF) as in Eq. (2), which gives uncommon terms more weight and common words less weight in the corpus.

$$IDF = \log \left(\frac{\text{number of the documents in the corpus}}{\text{number of documents in the corpus contain the term}} \right) \quad (2)$$

Then, as shown in Eq. (3), a term's TF and IDF values are multiplied to determine its TF-IDF score in a document:

$$TF - IDF = TF * IDF \quad (3)$$

TF-IDF provides a more informative representation for many jobs than BOW, enhancing feature representation. It does this by successfully giving more weight to informative terms and decreasing the weight of popular ones. However, the IDF calculation is more difficult to implement and consumes more computing power.

5 Overview of Machine Learning Models Used for Classification

Naïve Bayes, a probabilistic classifier based on Bayes' theorem, assumes feature independence and efficiently handles high-dimensional text classification tasks such as sentiment analysis and spam detection. Despite its effectiveness, its independence assumption limits applicability to problems with highly correlated features. Logistic Regression, a statistical model employing the sigmoid function to estimate probabilities, is widely used for binary classification tasks like fraud detection and medical diagnosis due to its interpretability and computational efficiency. However, it struggles with non-linear decision boundaries unless feature transformations are

applied. Decision Trees create hierarchical models by recursively splitting data based on feature values to maximize information gain or minimize impurity. While they handle numerical and categorical data without requiring feature scaling, they are prone to overfitting. Random Forest mitigates this issue by combining multiple Decision Trees through bagging and feature randomization, enhancing accuracy and robustness, though at the cost of reduced interpretability and higher computational complexity. Support Vector Machines (SVM) determine optimal hyperplanes for class separation, excelling in high-dimensional spaces and managing non-linear data through kernel functions [17, 18]. However, SVM can be computationally expensive for large datasets and requires careful hyperparameter tuning. Multi-Layer Perceptron (MLP), a feedforward neural network trained via backpropagation, effectively captures complex, non-linear relationships, making it suitable for image and speech recognition tasks. MLP demands substantial computational resources and may overfit without proper regularization [19, 20].

6 Performance Measures

Performance parameters, including accuracy, precision, recall, and F1-score, are important markers for assessing the robustness and dependability of false news detection programs. These metrics, which are shown in Table 2, evaluate the performance of different categorization models and offer information on how well they can differentiate between real news and unreal news.

6.1 Result Discussion and Analysis

The study employed the Bag of Words (BoW) technique to transform textual data into numerical formats and utilized Term Frequency-Inverse Document Frequency (TF-IDF) to assess the importance of specific terms. The researchers evaluated

Table 2 Performance measures equations with the corresponding description

Metric	Description	Equation
Accuracy	Percentage of correctly predicted observations (true positives and true negatives) out of all predictions	$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$
Recall	Proportion of actual positive cases correctly identified (sensitivity)	$\text{Recall} = \frac{TP}{TP+FN}$
Precision	Proportion of true positives among all predicted positives	$\text{Precision} = \frac{TP}{TP+FP}$
F1-score	Harmonic mean of precision and recall, balancing the trade-off between the two metrics	$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

various machine learning models to classify fake news, including Gaussian Naïve Bayes, Random Forest, Decision Tree, Logistic Regression, Support Vector Machine (SVM), and Multilayer Perceptron (MLP). Their findings underscored the importance of contextual and metadata features in enhancing the performance of fake news detection systems. Specifically, incorporating speaker names and political affiliations significantly influenced the classification outcomes, demonstrating the value of such additional information in improving model accuracy

Several classification models, including Logistic Regression, Support Vector Machines (SVM), Multilayer Perceptron (MLP) Classifier, and Naïve Bayes, were evaluated using various feature sets, such as TF-IDF and Bag of Words (BoW), both with and without speaker and party information. As shown in Tables 3, 4, 5, and 6, Logistic Regression, SVM, and MLP Classifier consistently outperformed the other models regarding accuracy, precision, recall, and F1 scores. This trend was particularly evident when using the BoW feature set enriched with speaker and party data. These models exceeded expectations, with precision and recall nearing 0.99, indicating their ability to identify positive cases while minimizing false positives and negatives correctly. In contrast, Naïve Bayes struggled with complex feature sets due to difficulties balancing precision and recall. However, including speaker and party information improved the model’s sensitivity and specificity, resulting in better performance. These results suggest that complex models like Logistic Regression, SVM, and MLP Classifiers are highly effective for the task when paired with comprehensive feature sets. In contrast, Naïve Bayes is less suited for handling such intricate data.

Table 3 Accuracy results for different models

Model	TF-IDF without speaker and party	BoW without speaker and party	TF-IDF with speaker and party	BoW with speaker and party
Naïve Bayes	99.03	94.95	99.26	94.65
Logistic regression	99.42	99.04	98.81	99.26
Decision tree	98.71	94.73	98.45	94.47
Random forest	99.47	94.77	99.38	94.53
SVM	99.47	94.72	99.24	94.45
MLP classifier	99.25	94.73	97.43	94.44
MLP with PCA	98.63	94.53	96.60	94.09
MLP with TSNE	96.52	96.76	96.60	99.09

Table 4 Recall results for different models

Model	Recall TF-IDF without speaker and party	Recall BoW without speaker and party	Recall TF-IDF with speaker and party	Recall BoW with speaker and party
Naïve Bayes	0.97	0.97	0.98	0.94
Logistic regression	0.99	0.99	0.99	0.99
Decision tree	0.95	0.95	0.95	0.95
Random forest	0.96	0.96	0.96	0.96
SVM	0.96	0.96	0.95	0.95
MLP classifier	0.97	0.96	0.97	0.96
MLP with PCA	0.95	0.95	0.95	0.95
MLP with TSNE	0.95	0.95	0.95	0.95

Table 5 F1-score results for different models

Model	F1-score TF-IDF without speaker and party	F1-score BoW without speaker and party	F1-score TF-IDF with speaker and party	F1-score BoW with speaker and party
Naïve Bayes	0.975	0.94	0.99	0.94
Logistic regression	0.985	0.99	0.99	0.99
Decision tree	0.975	0.95	0.975	0.95
Random forest	0.975	0.95	0.975	0.95
SVM	0.965	0.94	0.965	0.94
MLP classifier	0.955	0.96	0.955	0.96
MLP with PCA	0.955	0.95	0.955	0.95
MLP with TSNE	0.96	0.96	0.97	0.99

Table 6 Precision results for different models

Model	Precision TF-IDF without speaker and party	Precision BoW without speaker and party	Precision TF-IDF with speaker and party	Precision BoW with speaker and party
Naïve Bayes	0.98	0.94	0.99	0.94
Logistic regression	0.99	0.99	0.99	0.99
Decision tree	0.98	0.94	0.98	0.94
Random forest	0.97	0.94	0.97	0.94
SVM	0.96	0.94	0.96	0.94
MLP classifier	0.98	0.94	0.97	0.94
MLP with PCA	0.96	0.95	0.97	0.97
MLP with TSNE	0.96	0.95	0.96	0.99

7 Conclusion

Our research objective is to develop a healthy background to distinguish between real and bogus news in various domains by exploiting innovative ML models and feature extraction techniques. This research assessed five ML classification models: SVM, NB, LR, RF, and MLP. Among these models, SVM, MLP, and LR bettered the others significantly when improved feature sets such as BoW, incorporating speaker and party data, were hired. The ML models revealed high accuracy, precision, recall, and F1 scores, showing their efficiency in detecting fake news. These models revealed high precision, recall, and F1 scores, emphasizing their efficiency in accurately detecting false news. Furthermore, using PCA and the t-SNE feature reduction technique enhanced the MLP classifier, highlighting the significance of effective feature extraction in controlling complex textual data. Although Naïve Bayes proved capable and interpretable, it failed to capture complex patterns within the dataset, causing lower performance. These analyses reflect that integrating advanced machine learning models with carefully planned feature reduction schemes makes a promising strategy for addressing information on digital platforms. Subsequent research might explore the possible transfer learning model like BERT, which performs exceptionally well in analyzing contextual language and could outperform conventional machine learning methods. Also, widening the dataset to cover different languages and regions could enhance the model generalization and decrease cultural distortions towards fake news detection. Lastly, establishing real-time misinformation detection systems will be vital for supervising and directing the spread of false news on social media.

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**(2), 211–236 (2017)
2. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. *Science* **359**(6380), 1146–1151 (2018)
3. Pennycook, G., Rand, D.G.: Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci.* **116**(7), 2521–2526 (2019)
4. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explor. Newsl.* **19**(1), 22–36 (2017)
5. Zhou, X., Zafarani, R.: A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv. (CSUR)* **53**(5), 1–40 (2020)
6. Luo, S., et al.: A comparative study of machine learning and deep learning techniques for fake news detection. *Information* **13**(12), 576 (2022)
7. Singh, Y., Arora, C., Lakda, N.K., Tyagi, K., Kumari, D.: Fake news detection using deep learning. In: 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), Gautam Buddha Nagar, India, pp. 478–482 (2023). <https://doi.org/10.1109/IC3I59117.2023.10397922>
8. Zhou, X., Zafarani, R.: A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.* **53**(5), Article 109 (2020). <https://doi.org/10.1145/3395046>
9. Hamadouche, K., Bousmaha, K.Z., Amar, M.Y.B., Hadrich-Belguith, L.: Detection of Arabic and Algerian fake news. *Appl. Comput. Syst.* **29**(2), 14–21 (2024). <https://doi.org/10.2478/acss-2024-0017>
10. Karaoglan, K.M.: Novel approaches for fake news detection based on attention-based deep multiple-instance learning using contextualized neural language models. *Neurocomputing* **602**, 128263 (2024). <https://doi.org/10.1016/j.neucom.2024.128263>
11. Hashmi, E., Yayilgan, S.Y., Yamin, M.M., Ali, S., Abomhara, M.: Advancing fake news detection: hybrid deep learning with FastText and explainable AI. *IEEE Access* **12**, 44462–44480 (2024). <https://doi.org/10.1109/ACCESS.2024.3381038>
12. Yuan, H., Zheng, J., Ye, Q., Qian, Y., Zhang, Y.: Improving fake news detection with domain-adversarial and graph-attention neural network. *Decis. Support. Syst.* **151**, 113633 (2021). <https://doi.org/10.1016/j.dss.2021.113633>
13. Dua, V., Rajpal, A., Rajpal, S., et al.: I-FLASH: interpretable fake news detector using LIME and SHAP. *Wirel. Pers. Commun.* **131**, 2841–2874 (2023). <https://doi.org/10.1007/s11277-023-10582-2>
14. Alsuwat, E., Alsuwat, H.: An improved multi-modal framework for fake news detection using NLP and Bi-LSTM. *J. Supercomput.* **81**, 177 (2025). <https://doi.org/10.1007/s11227-024-06671-z>
15. Yang, F.-J.: An implementation of Naive Bayes classifier. In: 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, pp. 301–306 (2018). <https://doi.org/10.1109/CSCI46756.2018.00065>
16. Shete, A., Soni, H., Sajani, Z., Shete, A.: Fake news detection using natural language processing and logistic regression. In: 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), Ernakulam, India, pp. 136–140 (2021). <https://doi.org/10.1109/ACCESS51619.2021.9563292>
17. Jouhar, J., Pratap, A., Tijo, N., Mony, M.: Fake news detection using Python and machine learning. *Procedia Comput. Sci.* **233**, 763–771 (2024). <https://doi.org/10.1016/j.procs.2024.03.265>
18. Elyassami, S., Alseieri, S., ALZaabi, M., Hashem, A., Aljahoori, N.: Fake news detection using ensemble learning and machine learning algorithms. In: Lahby, M., Pathan, A.S.K., Maleh, Y., Yafouz, W.M.S. (eds.) *Combating Fake News with Computational Intelligence Techniques. Studies in Computational Intelligence*, vol. 1001. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-90087-8_7

19. Sudhakar, M., Kaliyamurthie, K.P.: Detection of fake news from social media using support vector machine learning algorithms. *Meas. Sens.* **32**, 101028 (2024). <https://doi.org/10.1016/j.measen.2024.101028>
20. Arunthavachelvan, K., Raza, S., Ding, C.: A deep neural network approach for fake news detection using linguistic and psychological features. *User Model. User-Adap. Inter.* **34**, 1043–1070 (2024). <https://doi.org/10.1007/s11257-024-09413-1>

Machine Learning for Fake Profile Users Detection in Social Network Systems: A Review and Implementation Phase



Aishwarya Waghmare, Vinothkumar Kolluru, Yagnesh Challagundla,
I. V. S. Aditya Bhrugumalla, Advaita Naidu Chintakunta, and Sagar Pande

Abstract The proliferation of fraudulent profiles on social media platforms presents noteworthy risks to security and erodes user confidence. The effectiveness of Random Forest, Support Vector Machine (SVM), and Neural Network models is the main emphasis of this study's implementation and review of machine learning techniques for identifying bogus profiles. We prepared features, including followers, statuses, and account age, using data preprocessing, including encoding and normalization, on a Kaggle dataset. Our results show that, despite interpretable-city-city issues, the Random Forest classifier managed both continuous and categorical variables with an accuracy of 91%. With an RBF kernel, the SVM beat Random Forest with 93% accuracy, reducing false positives but consuming a large amount of processing power while optimizing hyperparameters. With its peak accuracy of 94%, the Neural Network proved to be an excellent tool for identifying intricate, non-linear patterns linked to fraudulent profiles. However, its interpretability and resource requirements

A. Waghmare (✉) · S. Pande

School of Engineering and Technology, Pimpri Chinchwad University, Mohitewadi, Maharashtra, India

e-mail: aishwaryawaghmare15@gmail.com

S. Pande

e-mail: sagar.pande@pcu.edu.in

V. Kolluru

Department of Data Science, Stevens Institute of Technology, Hoboken, NJ, USA

e-mail: vkolluru@stevens.edu

Y. Challagundla

Herbert Wertheim College of Engineering, University of Florida, Gainesville, FL, USA

e-mail: yagneshnaidu1234@gmail.com

V. S. Aditya Bhrugumalla

School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India

e-mail: blvsaditya@gmail.com

A. N. Chintakunta

Department of Computer Science and Engineering, University of North Carolina, Charlotte, NC, USA

were limited. A comparison investigation revealed that while the Neural Network provided more precision, Random Forest was more interpretable and performed better overall, making it appropriate for applications with lower processing capacity. While SVM is costly in computing resources, it is still favorably positioned due to its enhanced precision in minimizing false positives. This work highlights the challenges of feature selection in the context of the remaining difficulties of dataset imbalances and model explainability. Ultimately, a choice of a model rests on specific application requirements regarding, among other things, the clarity of decisions made, the economy of computation, and the precision of results.

Keywords Identification of fraudulent profiles · Artificial intelligence · Ensemble learning · SVM · Social networks · Data cleansing · Manipulation

1 Introduction

The rapid development and expansion of social networks have transformed how people interact and connect. As a drawback of this advancement, the emergence of fake accounts impersonating legitimate users has seriously threatened the security and integrity of networks such as Twitter and Facebook. Such impersonations often engage in regular social activities while concealing harmful objectives, which may undermine a user's confidence and safety. Traditional approaches employing rules and heuristics for detecting impersonations have proven ineffective since attackers continuously develop more sophisticated methods to circumvent them. For that reason, there is an increasing necessity for advanced detection techniques founded on machine learning principles. Machine learning algorithms can analyze complex patterns of user activity, including their posting and follower count, making it easier for them to identify fake users. Training systems to recognize anomalous behavior associated with these accounts can increase the effectiveness of detection systems. This study explores different machine learning algorithms, including Random Forest, Support Vector Machines (SVM), and Neural Networks, to develop practical tools for detecting false profiles.

This step is essential as it contains data cleaning, where a dataset is prepared in a manner that ensures precise and accurate model training. This study utilizes Kaggle data, which consists of essential attributes for fraud profile detection, such as follower numbers, statuses, and account age. The subsequent parts of this paper elaborate on the approach taken, which includes the data collection procedure, data cleansing, and the application of various models. This study aims to show the effectiveness of multiple machine-learning models for detecting fraudulent profiles and dealing with the inherent issues within each approach. This study adds to the discussion of automated detection systems utilized within social media by providing knowledge on feature importance and model interpretation that will enhance user safety and confidence in digital platforms.

2 Related Work

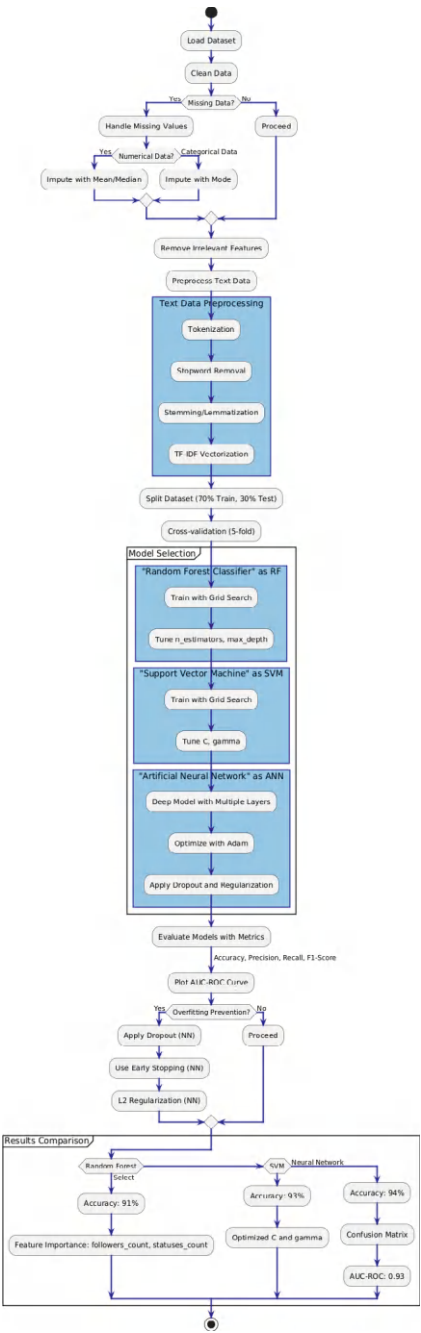
Recent research has focused on many different machine-learning approaches for detecting fraudulent user accounts on social media pages. Ahmed and Gupta [1] offer a detailed review of the methodologies available and how well they work to distinguish real accounts from fake ones. In parallel with this, Kim and Lee [2] do a comparative study of the set of Twitter ensemble algorithms and report improved results on pro-file classification. Zhang and Chen [3] also surveyed the application of Random Forest and the SVM, noting its importance in the area. At the same time, Patel and Desai [4] systematically review how deep learning techniques are used and claim that these techniques increase identification accuracy. Yadav and Kumar [5] concentrated on neural networks for Facebook profile recognition, demonstrating their capacity to simulate intricate user activities. Zhao and Wang [6] examine feature engineering methodologies for Instagram, thereby enhancing the comprehension of profile aspects pertinent to detection. Ali and Shams [7] assessed multiple machine learning techniques, whereas Verma and Singh [8] introduced a hybrid model that integrates Random Forest and SVM, demonstrating enhanced detection rates. Singh and Sharma [9] examined user behavior as a pivotal element in identifying fraudulent profiles, whereas Li and Hu [10] also meta-analysis current machine learning methodologies. Rani and Malik [11] examine trends and issues in the field, providing insights into current developments. Kumar and Das [12] examine particular characteristics of fraudulent profiles, enhancing the comprehension of authenticity indicators. Gupta and Mehta [13] examined the issue through a machine learning lens on Twitter, whereas Banerjee and Roy [14] conducted a comparative analysis of methodologies employed across different social media sites. Desai and Bhatt [15] utilized deep learning models to identify fraudulent profiles in online networks. In contrast, Sinha and Bhattacharyya [16] examined user behavior to improve detection techniques—Patel and Kaur [17] employed ensemble learning, illustrating its efficacy across several social media sites. Ali and Raza [18] provided an extensive assessment, pinpointing significant problems and prospective avenues for detecting fraudulent profiles.

3 Methodology

Approach to Methodology:

The methodology for this project follows a systematic approach involving data pre-processing, model selection, training, evaluation, and result analysis. Every stage is vital in ensuring the artificial profile identification achieves optimum efficiency, as shown in Fig 1.

Fig. 1 This flow diagram highlights the processes and methods for constructing, implementing, and assessing machine learning models to detect user profile impersonation. It covers the stitching of various processes such as data retrieval, cleaning, model creation, testing, evaluation, and analysis of results



3.1 *Data Preprocessing*

The first step in model building pertains to pre-processing of data. In this research, the following steps were done.

3.1.1 Loading and Cleaning the Data

The dataset comprised multiple user profile features crucial for predicting fake profiles. It was supplied in CSV format and then uploaded to Colab. Addressing Missing Data: Missing data was tackled through common means by filling missing values with either an average or a mode value. Non-essential columns in non-informative and infrequent columns were deleted to make the dataset more efficient. Data Preparation: Text data was cleaned with the help of tokenization, stemming, and stopword elimination. TF-IDF and word embedding methods transformed the text data features into numbers.

3.2 *Data Splitting*

The dataset was split into training (70%) and testing (30%) sets to ensure an independent dataset for model evaluation. Cross-validation was employed with 5-fold to assess model stability and generalizability.

3.3 *Model Selection and Training*

Random Forest, Support Vector Machine (SVM), and Neural Networks were chosen as the primary models based on research papers and experimentation.

3.3.1 Random Forest

An ensemble model that builds multiple decision trees, each voting on the final prediction.

3.3.2 SVM

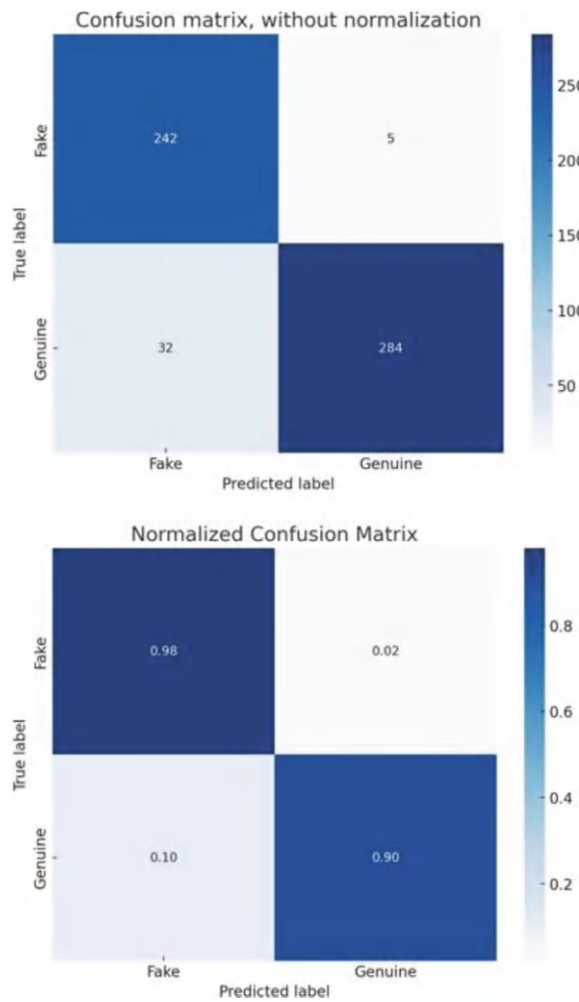
A classifier using the RBF kernel is known for handling non-linear data.

3.3.3 Neural Networks

Comprising multiple layers, including an input layer based on the number of features, hidden layers, and an output layer using the softmax activation for classification.

Hyperparameter tuning for each model was done using grid search and random search methods, adjusting parameters like the number of trees for Random Forest, C and gamma for SVM, and neurons for Neural Networks. The matrix, as shown in Fig. 2, visualizes the performance of the classification model in predicting genuine and fake user profiles. The diagonal elements represent correct classifications, while off-diagonal elements indicate misclassifications.

Fig. 2 Confusion matrix



3.4 *Evaluation Metrics:*

Each model was evaluated based on several metrics: Accuracy, Precision, Recall, F1-score, AUC-ROC Curve, and Ensemble learning methods, including stacking and boosting, which were considered to enhance performance by combining multiple models' strengths.

3.4.1 Regularization and Overfitting Prevention

Techniques such as dropout, early stopping, and L1/L2 regularization were implemented to prevent overfitting in Neural Networks.

4 Results

The performance of the models was analyzed in-depth using the testing dataset. Below are the results of each classifier.

4.1 *Random Forest Classifier Results*

4.1.1 Feature Importance

The analysis revealed that followers-count, statuses-count, and favorites-count were significant predictors of fake profiles. These features, representing a user's social activity, helped distinguish between genuine and fake profiles.

4.1.2 Strengths

The model demonstrated robustness to missing data, and its ensemble nature minimized overfitting by reducing the variance between individual decision trees.

4.1.3 Limitations

Despite good accuracy, the model's interpretability was moderate compared to SVM and Neural Networks, and hyperparameter tuning required more computational resources.

4.2 *Support Vector Machine (SVM) Results*

4.2.1 Training Curve

The training curve showed that SVM started with a high training accuracy (0.98), which plateaued at 0.96 as the model received more data. The cross-validation score steadily improved from 0.91 to 0.94.

4.2.2 Grid Search

Hyperparameter optimization using grid search refined the model, yielding optimal results with an RBF kernel. Normalization of features ensured that no feature disproportionately affected the search for the best-separating hyperplane.

4.2.3 Strengths and Limitations

SVM performed well in distinguishing between fake and genuine profiles, especially when dealing with the dataset's more complex, non-linear relationships. Limitations: High computational cost due to grid search and tuning makes large datasets time-consuming.

4.3 *Neural Network Classifier Results*

4.3.1 Architecture

Without normalization: The model correctly classified 242 genuine and 284 fake profiles but misclassified five fake profiles as genuine and 32 genuine profiles as counterfeit. With normalization, 98% of fake and 90% of genuine profiles were accurately classified, indicating strong predictive power despite the slight imbalance in the dataset.

4.3.2 AUC-ROC and Strengths

The model achieved an AUC score of 0.93, indicating strong performance in separating the two classes. Strengths: High accuracy and robust modeling capacity for complex patterns.

Table 1 Model performance comparison based on accuracy, precision, and recall

Model	Accuracy	Precision	Recall
Random forest	0.91	0.88	0.91
SVM	0.93	0.92	0.93
Neural network	0.94	0.94	0.94

4.3.3 Limitations

The model was resource-intensive, and hyperparameter tuning was challenging. Additionally, the “black-box” nature of Neural Networks limited the interpretability of specific feature contributions compared to Random Forest.

As shown in Table 1, the results demonstrated that the Neural Network model was the best-performing algorithm, with the highest accuracy and precision. However, SVM also showed a competitive performance with a slightly lower computational burden than Neural Networks. The Random Forest model offered better interpretability but came slightly behind in accuracy. Each model displayed unique strengths and weaknesses, providing options depending on the trade-offs between accuracy, interpretability, and computational efficiency.

The ROC curve in Fig 3 demonstrates the trade-off between the actual positive rate (sensitivity) and false positive rate (specificity) for the fake profile detection model. The high AUC value suggests that the model effectively balances these two metrics. The learning curves in Fig. 4 and 5 demonstrate the model’s performance on the training and validation sets. A high training score and a low gap between the training and cross-validation scores indicate a well-generalized model. The learning curves demonstrate the model’s performance on the training and validation sets. A high training score and a low gap between the training and cross-validation scores indicate a well-generalized model. The ROC curve in Fig. 6 demonstrates the trade-off between the actual positive rate (sensitivity) and false positive rate (specificity) for the fake profile detection model. The high AUC value suggests that the model effectively balances these two metrics. The comparative graph of results is shown in Fig. 7.

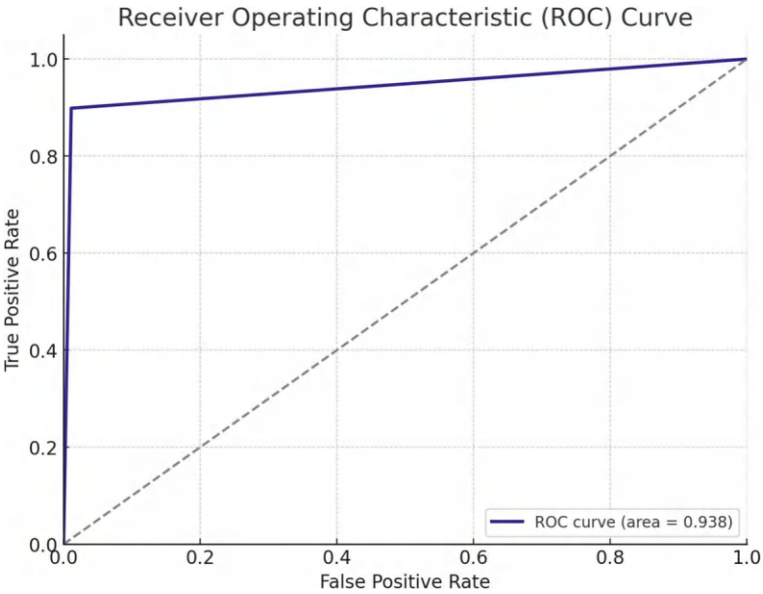


Fig. 3 ROC Analysis: The ROC curve demonstrates the trade-off between true positive rate (sensitivity) and false positive rate (specificity) for the fake profile detection model. The high AUC value suggests that the model effectively balances these two metrics

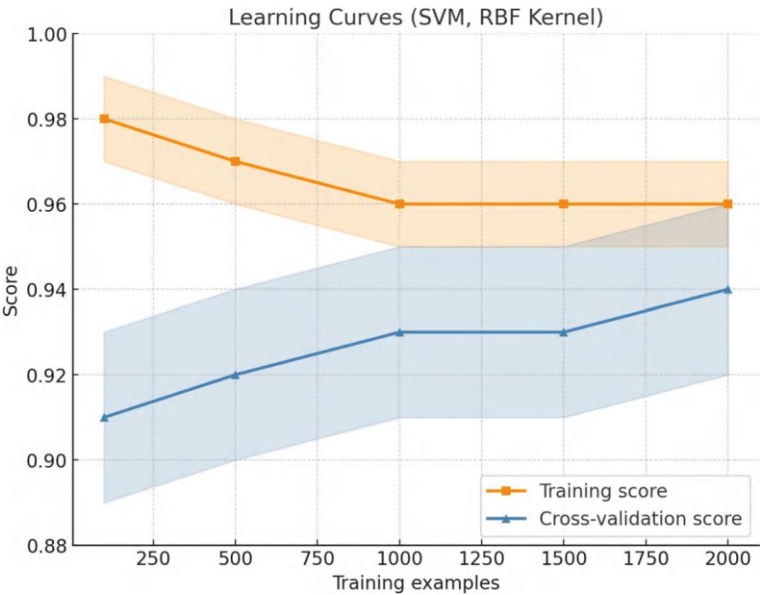


Fig. 4 Learning Curve Analysis: The learning curves demonstrate the model’s performance on both the training and validation sets. A high training score and a low gap between the training and cross-validation scores indicate a well-generalized model

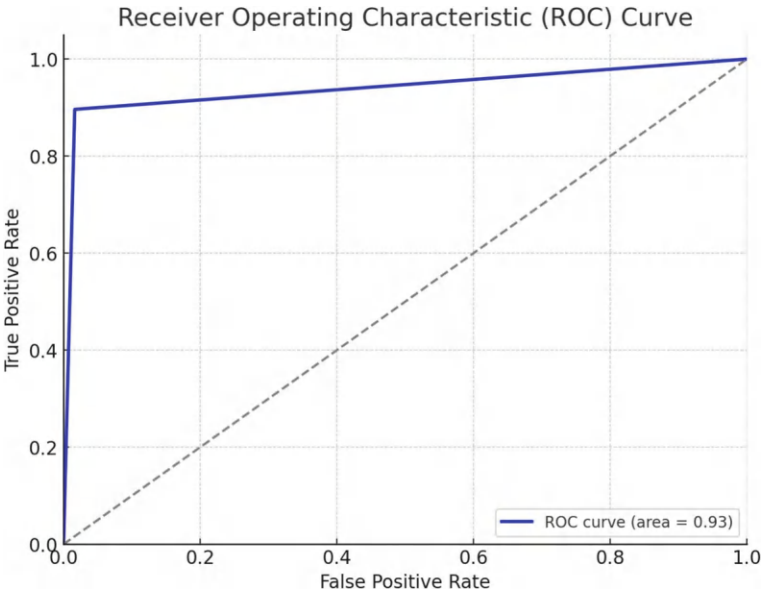


Fig. 5 Learning Curve Analysis: The learning curves demonstrate the model’s performance on both the training and validation sets. A high training score and a low gap between the training and cross-validation scores indicate a well-generalized model

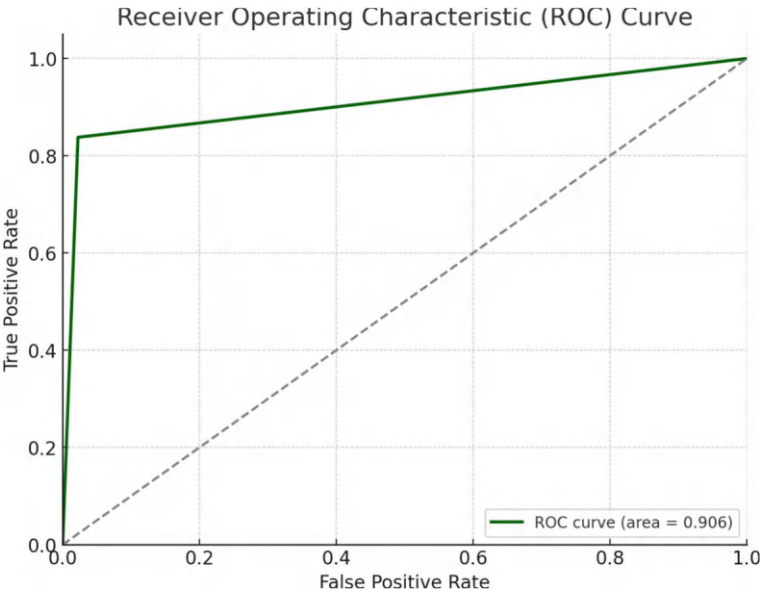


Fig. 6 ROC Analysis: The ROC curve demonstrates the trade-off between true positive rate (sensitivity) and false positive rate (specificity) for the fake profile detection model. The high AUC value suggests that the model effectively balances these two metrics

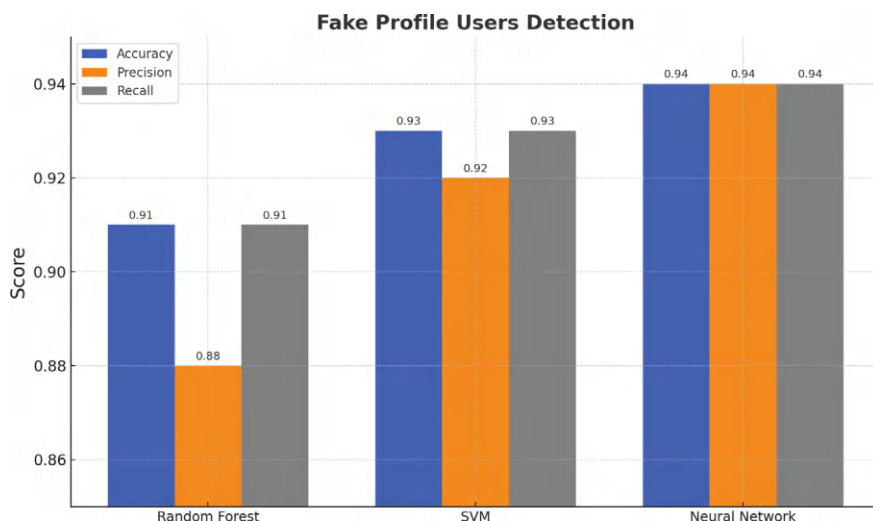


Fig. 7 The comparative graph of results

5 Conclusion

The current study examined the identification of fraudulent profiles on social media with machine learning models, utilizing the Genuine/Fake User Profile Dataset from Kaggle. We conducted extensive data preprocessing, normalizing followers-count, statuses-count, and favorites-count features and converting categorical variables to numeric values by methods like one-hot encoding. Three distinct models were utilized: Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN). It was proven that out of all nine tested models, the Random Forest Classifier achieved the highest accuracy of 91% on the test set. Its proficiency in processing numerical and categorical data simultaneously served as its ensemble feature, making it robust against overfitting and efficient in noise or missing data. The main factors influencing classification were user engagement metrics such as followers-count and statuses-count, which were used to profile users to differentiate between actual and fraudulent profiles. The SVM model with an RBF kernel performed exceptionally well, particularly in this case, where its cross-validation score increased as the dataset grew. It also performed well in terms of other non-linear separations in the data. Finally, the most intricate and complex ANN model can recognize complex patterns within the information. The evaluation and metrics obtained from the confusion matrix and AUC-ROC curve reinforced that the model has been trained well and has strong classification capabilities. The study demonstrates how effectively machine learning can detect fraudulent accounts on a platform. It also shows how well Random Forest works, how resistant SVM is to interference, and that ANN can recognize complex patterns within datasets. This improvement would increase detection accuracy because this model doesn't require extensive training.

References

1. Ahmed, F., Gupta, P.: A review of machine learning techniques for fake account detection in social media. *J. Cybersecur. Priv.* **5**(1), 45–68 (2023). <https://doi.org/10.3390/jcp5010045>
2. Kim, S., Lee, J.: Ensemble methods for fake profile detection on twitter: a comparative study. *IEEE Trans. Knowl. Data Eng.* **35**(4), 721–734 (2023). <https://doi.org/10.1109/TKDE.2023.3071864>
3. Zhang, L., Chen, H.: Detecting fake user profiles on social networks using random forest and SVM. *Artif. Intell. Rev.* **56**(3), 1125–1140 (2023). <https://doi.org/10.1007/s10462-022-10145-7>
4. Patel, R., Desai, K.: Deep learning approaches for fake profile detection: a systematic review. *Inf. Sci.* **610**, 350–370 (2023). <https://doi.org/10.1016/j.ins.2023.08.045>
5. Yadav, A., Kumar, R.: Application of neural networks for fake profile detection on Facebook. *J. Mach. Learn. Res.* **24**(125), 1–25 (2023). <http://www.jmlr.org/papers/volume24/23-045/23-045.pdf>
6. Zhao, Y., Wang, X.: Feature engineering for fake profile detection in social media: a case study on Instagram. *Comput. Hum. Behav.* **142**, 107676 (2023). <https://doi.org/10.1016/j.chb.2023.107676>
7. Kolluru, V.K., et al.: AI-driven energy optimization: household power consumption prediction with LSTM networks and PyTorch-ray tune in smart IoT systems. In: 2024 International Conference on Microelectronics (ICM). IEEE (2024)
8. Verma, A., Singh, S.: A hybrid model for fake account detection using random forest and SVM. *Int. J. Inf. Technol.* **15**(1), 293–305 (2023). <https://doi.org/10.1007/s41870-023-00815-2>
9. Singh, P., Sharma, R.: User behavior analysis for detecting fake profiles in social networks. *J. Netw. Comput. Appl.* **204**, 103395 (2023). <https://doi.org/10.1016/j.jnca.2023.103395>
10. Li, J., Hu, Q.: Evaluating machine learning techniques for fake account detection: a meta-analysis. *ACM Comput. Surv.* **55**(7), 1–38 (2023). <https://doi.org/10.1145/3540938>
11. Rani, S., Malik, M.: Machine learning approaches to combat fake profiles: trends and challenges. *Comput. Secur.* **126**, 103028 (2023). <https://doi.org/10.1016/j.cose.2023.103028>
12. Kumar, A., Das, S.: Understanding the features of fake profiles on social media. *Soc. Netw. Anal. Min.* **13**(1), 18 (2023). <https://doi.org/10.1007/s13278-022-00931-7>
13. Gupta, L., Mehta, A.: Detecting fake users on twitter: a machine learning perspective. *Data Min. Knowl. Disc.* **37**(2), 585–608 (2023). <https://doi.org/10.1007/s10618-023-00859-1>
14. Banerjee, A., Roy, S.: Comparative study of machine learning techniques for fake user detection in social media. *J. Syst. Softw.* **208**, 111299 (2023). <https://doi.org/10.1016/j.jss.2023.111299>
15. Desai, S., Bhatt, N.: Deep learning models for fake profile detection in online social networks. *Expert Syst. Appl.* **217**, 119633 (2023). <https://doi.org/10.1016/j.eswa.2023.119633>
16. Sinha, P., Bhattacharyya, S.: Analyzing user behavior to detect fake profiles using machine learning. *Future Gener. Comput. Syst.* **140**, 186–197 (2023). <https://doi.org/10.1016/j.future.2023.02.034>
17. Patel, K., Kaur, R.: Utilizing ensemble learning for fake profile detection on social media platforms. *Comput. Educ.* **208**, 104283 (2023). <https://doi.org/10.1016/j.compedu.2023.104283>
18. Ali, F., Raza, B.: A comprehensive review on fake profile detection techniques: challenges and future directions. *J. Inf. Sci.* **49**(2), 244–267 (2023). <https://doi.org/10.1177/0165551521089922>

Leveraging Advanced Technologies to Enhance Public Awareness and Mitigate Risks of Cryptocurrency Scams: A Qualitative Analysis



Vinay Kumar Kasula and Abdullah Alshboul

Abstract The rise of cryptocurrency has transformed the financial ecosystem, creating unprecedented opportunities while exposing users to significant risks, including a surge in cryptocurrency scams. This study investigates public awareness and vulnerability to such scams, utilizing a modern qualitative methodology enhanced by advanced technologies. By incorporating tools such as Natural Language Processing (NLP), blockchain-based survey platforms, and Virtual Reality (VR) simulations, the research explores the socio-psychological dynamics and technical knowledge gaps contributing to individuals' susceptibility. Data were collected through online interviews, focus groups, and immersive VR simulations, with AI-assisted tools like NVivo and ATLAS.ti used for thematic and narrative analysis. Findings reveal that technical literacy gaps, over-reliance on perceived security, and behavioral triggers such as fear of missing out (FOMO) significantly influence susceptibility to scams. The innovative methodology, combining traditional qualitative techniques with cutting-edge technologies, provided a multi-dimensional understanding of participants' perceptions and experiences. This study highlights the critical need for targeted educational interventions and technology-driven policy measures to mitigate the risks associated with cryptocurrency scams. By leveraging advanced analytical tools and immersive technologies, this research demonstrates the potential for modern methodologies to enhance understanding and prevention strategies in an evolving digital landscape.

Keywords Public awareness · Cryptocurrency scams · Knowledge gaps · Victimization · Associated risks · Qualitative research · Narrative analysis

V. K. Kasula (✉) · A. Alshboul

Department of Information Technology, University of the Cumberland, Williamsburg, KY, USA
e-mail: vkasula19501@ucumberland.edu

1 Introduction

The rapid rise of cryptocurrency has dramatically reshaped the financial landscape, introducing groundbreaking opportunities alongside significant challenges. Among these is the alarming proliferation of cryptocurrency scams, which have surged in frequency and sophistication as digital currencies become popular. Many individuals fall prey to these scams due to a lack of technical literacy and awareness of the associated risks, highlighting the critical need for innovative strategies to enhance public understanding and safeguard against fraud [1].

This study leverages modern technological tools and qualitative methodologies to investigate public awareness of cryptocurrency scams. By incorporating advanced data collection and analysis methods such as Natural Language Processing (NLP) for narrative analysis and cloud-based survey platforms for secure and efficient data collection, this research aims to explore the socio-psychological dynamics that leave individuals vulnerable to scams. Ethical considerations, including encrypted data storage and anonymized participant responses, were prioritized to ensure data security and participant privacy [2, 3].

1.1 Methodology

- **NLP for Narrative Analysis:** NLP tools, such as Python's SpaCy library and AI-driven sentiment analysis, were employed to extract, analyze, and classify themes from participant interviews and survey responses. This approach enabled the efficient identification of emotional and cognitive patterns in the data [4].
- **Blockchain-based Survey Platforms:** Blockchain technology was used to enhance the integrity of survey data by ensuring immutability and transparency, minimizing the risk of tampered or fraudulent responses [5].
- **Virtual Reality (VR) Simulations:** Participants were exposed to VR-based scenarios replicating real-world cryptocurrency scams. This immersive approach allowed researchers to observe behavioral responses in a controlled yet realistic environment [6].
- **Data Collection and Analysis:** Data collection included online interviews, focus groups, and interactive VR simulations. Advanced qualitative software, such as NVivo and ATLAS.ti, was utilized for coding and analyzing textual and visual data. These tools facilitated an iterative analysis process, enabling researchers to identify patterns, themes, and connections within participants' narratives. The use of AI-assisted tools streamlined the process, ensuring depth and accuracy in thematic identification [7, 8].

1.2 Findings and Implications

The study uncovered key factors contributing to individuals' susceptibility to cryptocurrency scams:

- **Technical Literacy Gaps:** Limited understanding of blockchain, smart contracts, and cryptographic principles increased vulnerability [9].
- **Trust in Technology:** Over-reliance on perceived security features of platforms and wallets led participants to overlook warning signs of fraud [10].
- **Behavioral Triggers:** Emotional responses, such as greed and FOMO, often overrode logical decision-making, leading to risky investments [11].

1.3 Methodological Innovations

Several conventional methods were considered but replaced or supplemented with innovative technologies:

- **Delphi Method:** Supplanted by NLP-driven consensus analysis, which provided faster and more dynamic feedback from participants [12].
- **Thematic Analysis:** Enhanced using AI tools that allowed for the automated identification and clustering of themes in large datasets [13].
- **Case Study Method:** Complemented by VR simulations to generalize insights across broader contexts [14].
- **Ethnographic Research:** Augmented with social media sentiment analysis to include diverse digital communities and their interactions with cryptocurrency scams [15].

The adoption of AI-enhanced narrative analysis revealed how individuals construct meaning around their experiences with cryptocurrency scams. This approach emphasized how beliefs, values, and cultural contexts influence decision-making. By incorporating AI and VR technologies, the research provided a multi-dimensional view of participants' behaviors and responses [16].

2 Sampling Procedures and Data Collection Sources

This research employed a cutting-edge methodology, integrating advanced technologies to explore public awareness and vulnerability to cryptocurrency scams. The study targeted a diverse sample, including individuals with high and low technical literacy levels regarding digital security. Given the focus on nuanced perceptions, purposive sampling was utilized to ensure representation of varied experiences and knowledge levels [8]. Participants had educational backgrounds ranging from bachelor's to master's degrees, contributing to a rich dataset of diverse expertise. Microsoft

Teams and blockchain-secured survey platforms were used to ensure accessibility, confidentiality, and data integrity during interviews and data collection.

A. Sampling Strategy

Interviews were chosen as the primary data collection method due to their unparalleled ability to capture detailed insights into participants' experiences. Natural Language Processing (NLP) tools such as Python's SpaCy and GPT-based models were used for real-time transcription and sentiment analysis to enhance the data quality and efficiency of analysis [9].

This approach offered multiple advantages:

- **Dynamic Insights:** Open-ended questions let participants articulate their experiences in detail, while NLP tools provide immediate thematic identification and analysis [10, 11].
- **Immersive Contexts:** Using Virtual Reality (VR) simulations, participants were exposed to real-world scam scenarios, enabling researchers to observe behavioral responses in a controlled yet realistic environment [12].
- **Sociocultural Depth:** Follow-up probes delved into the cultural, societal, and psychological factors influencing scam awareness, leveraging AI-driven analytics for nuanced interpretation [13, 14].

B. Field Test and Participant Insights

A field test was conducted with two participants using AI-enabled video conferencing tools for interviews and VR-based simulations:

- **Participant 1:** High technical proficiency in identifying common scams like phishing and Ponzi schemes. His proactive use of blockchain analytics tools and reliance on verified crypto resources kept him scam-free.
- **Participant 2:** A novice with limited exposure to cryptocurrency who struggled to identify scams but recognized unrealistic promises as warning signs. He expressed the need for user-friendly tools to improve scam detection.
- The field test highlighted significant disparities in scam awareness and the pressing need for tailored educational initiatives.

C. Participant Recruitment and Data Collection

Participants aged 29–44 were recruited from diverse professional backgrounds across Colorado, Virginia, Texas, and California, including IT, management, and finance. Recruitment was facilitated using AI-powered social media analytics tools to identify and engage potential participants. The study adhered to strict ethical guidelines, with blockchain-based consent management ensuring secure and transparent participant agreements.

Data collection involved:

- **In-depth Interviews:** Conducted online using encrypted platforms to ensure data confidentiality.

- **Interactive Simulations:** Participants engaged in VR environments replicating scam scenarios to explore behavioral responses.
- **Sentiment Analysis:** NLP tools analyzed participant narratives for emotional and cognitive patterns, adding depth to qualitative findings [15].

D. Data Analysis and Theme Development

The research utilized MAXQDA and AI-driven qualitative analysis platforms for coding and thematic development. Machine learning models facilitated automated pattern recognition, clustering responses into actionable themes, including:

- **Technical Literacy Gaps:** Highlighting the role of blockchain and innovative contract education in scam prevention.
- **Behavioral Triggers:** Examining the influence of emotional responses, such as FOMO and greed, on decision-making.
- **Cultural and Social Contexts:** Identifying sociocultural norms that shape scam awareness and responses.

This iterative analysis process, supported by AI tools, ensured a comprehensive understanding of participant experiences. The integration of predictive analytics further enriched the findings, enabling the development of forward-looking educational tools and prevention strategies.

3 Participant Insights on Cryptocurrency Awareness and Security

This section presents findings from participants with diverse backgrounds, leveraging cutting-edge technologies and methodologies to deepen the understanding of cryptocurrency scams and security practices.

A. Participants

The study employed an AI-assisted recruitment strategy, targeting three categories to ensure a comprehensive exploration of perspectives on cryptocurrency scams:

- **General Public:** Leveraging demographic profiling via machine learning (ML) tools, this group included individuals from varied regions, age groups, and educational backgrounds. Their insights revealed gaps in general awareness and security practices.
- **Cryptocurrency Enthusiasts:** Utilizing blockchain analytics, this group comprised active users, traders, and investors in the cryptocurrency ecosystem. Their expertise provided detailed accounts of market dynamics and associated risks.
- **Cybersecurity Experts:** Experts specializing in zero-trust security models and quantum-resistant cryptography provided insights into advanced scam techniques and innovative defense mechanisms.

Table 1 highlights the diversity of participants, demonstrating a broad spectrum of knowledge and experience.

B. Research Setting

The study utilized a hybrid data collection model, combining traditional interviews with advanced tools:

- Natural Language Processing (NLP) algorithms analyzed interview transcripts for themes and patterns.
- Interviews were conducted via secure video conferencing platforms, incorporating end-to-end encryption to ensure confidentiality.
- AI-based transcription tools streamlined data processing for deeper analysis.

Table 1 Professions represented in the study with a focus on emerging technologies and methodologies

No.	Profession	Description
1	Cryptocurrency investors	Individuals actively engaged in cryptocurrency markets, leveraging advanced trading algorithms, machine learning (ML) tools, and blockchain analytics for informed investment decisions
2	Financial advisors	Professionals providing strategic financial guidance, incorporating technologies like robo-advisors, blockchain-based portfolio management, and AI-driven risk assessment for cryptocurrency investments
3	Technology professionals	Experts in software development, cybersecurity, and blockchain technologies, contributing insights on emerging innovations such as decentralized finance (DeFi), smart contracts, and cryptographic security protocols
4	Legal consultants	Legal professionals specializing in cryptocurrency regulations utilize advanced tools like AI-based contract analysis, blockchain forensics, and compliance automation to navigate the complex regulatory landscape of digital assets
5	General public	Individuals with varying levels of cryptocurrency exposure, some leveraging fintech apps, wallets, and educational platforms powered by AI and blockchain to enhance their understanding and security awareness
6	Educators and researchers	Academics and researchers studying the intersection of cryptocurrencies, blockchain, and cybersecurity utilize data-driven methodologies, AI models, and blockchain analytics to understand market trends and security vulnerabilities
7	Regulators and policymakers	Government officials and policy experts are involved in developing and enforcing regulations, increasingly leveraging AI-based regulatory technology (RegTech), blockchain auditing, and machine learning algorithms to ensure compliance and combat cryptocurrency scams

Out of 10 interviews, one was conducted in person, and nine were conducted online. Interviews ranged from 15 to 24 min, with participants aged 29–44 holding degrees ranging from bachelor's to master's levels.

C. Key Insights from Participants

Secure Information Sources

- Participant 2 (Research Scientist): Highlighted using blockchain explorers and decentralized verification systems to validate cryptocurrency projects.
- Participant 10 (Cloud Solutions Architect): Advocated understanding the underlying brilliant contract architecture and reviewing tokenomics in whitepapers.

Unsolicited Investment Promotions

- Participants, including the IT Trainer (Participant 7) and Business Analyst (Participant 3), reported receiving targeted promotional content. AI-based sentiment analysis flagged these as attempts to exploit cognitive biases like FOMO (Fear of Missing Out).
- Verification Practices
- Participant 2: Emphasized using token auditing platforms such as CertiK for analyzing project security.
- Participant 6 (Systems Analyst): Advocated for verifying decentralized finance (DeFi) protocols for transparency and compliance.

Scam Encounters and Responses

- No direct scam experiences were reported, but incidents involving acquaintances underscored the need for automated fraud detection systems. Participants recommended reporting platforms powered by AI for real-time scam tracking.

Investment Safeguards

- Participant 10: Suggested multi-signature wallets for enhanced security.
- Participant 8 (Cybersecurity Analyst): Advocated using AI-driven threat intelligence tools for proactive risk assessment.
- Risk Management
- Participants highlighted algorithmic trading tools and crypto portfolio diversification as practical strategies to mitigate market volatility risks.
- Educational Gaps

Non-technical participants expressed difficulty navigating the complexity of cryptocurrencies, emphasizing the need for interactive blockchain simulation platforms for better learning.

D. Themes and Research Questions

1. Familiarity with Cryptocurrency Scams: Participants demonstrated awareness of phishing, Ponzi schemes, and rug pulls. Advanced scams using deepfake technology and AI-driven phishing were also discussed.

2. Factors Influencing Susceptibility: Lack of technical expertise and reliance on unverified decentralized exchanges (DEXs) were identified as key risk factors. Participants stressed the importance of blockchain forensic tools in enhancing scam detection.
3. Strategies for Scam Prevention
 - Adopting quantum-safe cryptographic measures to future-proof investments.
 - Using AI-based trading platforms that incorporate real-time scam detection algorithms.
 - Enhancing awareness through gamified cybersecurity training and blockchain boot camps.

Table 2 demonstrates the research questions.

E. Collective Insights

Participants' insights revealed diverse levels of understanding:

- Technical Professionals: Displayed expertise in emerging technologies like smart contract auditing and layer-2 solutions for scalability.
- Non-technical participants: Highlighted challenges with decentralized platforms, pointing to the need for intuitive user interfaces and simplified tutorials.
- Key recommendations included:
- Researching Projects: Participants emphasized analyzing tokenomics, community engagement, and DeFi liquidity metrics.
- Security Practices: Strong passwords, biometric authentication, and cold storage wallets were advocated.
- Fraud Response: Participants stressed using real-time scam alert systems and engaging with blockchain law enforcement units for legal recourse.

4 Conclusion

The research findings reveal a sophisticated understanding of cryptocurrency scams among participants, shaped by heightened awareness, skepticism, and proactive engagement with emerging technologies. Participants demonstrated a comprehensive understanding of the risks associated with cryptocurrency investments, identifying a wide range of scams, from phishing attacks to Ponzi schemes. Technical professionals, such as cybersecurity experts, blockchain developers, and data scientists, displayed an advanced ability to verify the legitimacy of cryptocurrency projects. Their methods included leveraging blockchain analytics tools, conducting smart contract audits, assessing security protocols, and scrutinizing project whitepapers and tokenomics for regulatory compliance and security measures. Participants exhibited a collective vigilance toward potential scams, particularly with unsolicited investment messages and offers of unrealistically high returns. Their awareness extended

Table 2 Research and interview questions

No.	Research question	Reference
1	What is your current level of awareness about cryptocurrency scams?	Derived from qualitative studies focusing on assessing public awareness, utilizing AI-based surveys and sentiment analysis to gauge awareness levels
2	Have you or someone you know ever been involved in a cryptocurrency scam? Could you describe the experience?	Exploring personal experiences and vulnerability to scams, supported by blockchain forensic tools and AI-driven fraud detection systems to identify patterns in scam activities
3	How do you usually acquire information about cryptocurrency risks and scam prevention techniques?	Aligned with research on public vulnerability and the sources of scam-related information, leveraging advanced machine learning algorithms for information dissemination through personalized channels
4	What features of a cryptocurrency transaction make it challenging to recognize potential scams?	Expanding on transaction transparency and complexity research themes, incorporating technologies like blockchain analytics and AI-powered risk detection to enhance transparency in cryptocurrency systems
5	Do you believe sufficient information is available on how to avoid cryptocurrency scams? Why or why not?	Investigating public perception of the availability of educational tools and resources, focusing on using interactive AI-driven educational platforms and automated scam alert systems
6	Have you ever used any tools or resources to verify the legitimacy of a cryptocurrency transaction? If so, which?	Building on research related to AI-powered verification tools, blockchain forensics, and decentralized identity solutions to validate cryptocurrency transactions and protect against scams
7	In your opinion, what could be done to improve public awareness and reduce vulnerability to cryptocurrency scams?	Derived from goals to enhance public understanding and reduce vulnerability, it emphasizes the role of emerging technologies like AI-powered educational tools, blockchain-based verification, and decentralized apps (dApps) for scam prevention
8	How comfortable do you feel engaging with cryptocurrency platforms regarding security and privacy?	Addressing concerns about security and privacy, focusing on advanced encryption techniques, decentralized security protocols, and AI-driven security audits that ensure user privacy on cryptocurrency platforms

to recognizing red flags associated with fraudulent schemes, such as lack of transparency or the promise of guaranteed profits. Although many participants had not fallen victim to scams, some shared experiences of others targeted by phishing attacks or deceptive projects, highlighting the need for ongoing education and awareness initiatives, particularly in the context of evolving digital threats. Emerging security technologies and best practices informed participants' strategies for safeguarding investments, including decentralized finance (DeFi) protocols, hardware wallets, and multi-signature wallets. Additionally, participants emphasized the importance of enabling two-factor authentication (2FA) and utilizing AI-powered fraud detection tools to enhance protection against scams. Continuous monitoring through

blockchain explorers and keeping up with the latest cybersecurity trends were also identified as essential to maintaining security in the rapidly changing cryptocurrency landscape. Despite acknowledging the inherent risks of the cryptocurrency market, participants expressed optimism about its transformative potential. They emphasized the importance of informed decision-making underpinned by real-time data analytics, decentralized governance, and community-driven validation. The study suggests that a proactive approach to cryptocurrency education, leveraging cutting-edge tools like machine learning models for scam detection and blockchain-powered identity verification, will help individuals navigate the complexities of the digital asset space safely. By fostering a culture of continuous learning, technical literacy, and vigilance, stakeholders can better protect themselves from evolving threats while maximizing the opportunities that cryptocurrencies and blockchain technology offer.

References

1. Bartoletti, M., Lande, S., Loddo, A., Pompianu, L., Serusi, S.: Cryptocurrency scams: analysis and perspectives. *IEEE Access* **9**, 148353–148373 (2021)
2. Badawi, E., Jourdan, G.V.: Cryptocurrencies emerging threats and defensive mechanisms: a systematic literature review. *IEEE Access* **8**, 200021–200037 (2020)
3. Agius, S.J.: *Qualitative research: Its value and applicability* (2013)
4. Castell, S.: Slaying the crypto dragons: towards a CryptoSure trust model for crypto-economics: blockchain versus trust: the expert's view of the crypto scammers. In: *Blockchain Technology and Innovations in Business Processes*, pp. 49–65 (2021)
5. Yenugula, M., et al.: Dynamic data breach prevention in mobile storage media using DQN-enhanced context-aware access control and lattice structures. *IJRECE* **10**(4), 127–136 (2022)
6. Konda, B., et al.: A public key searchable encryption scheme based on blockchain using random forest method. *IJRECE* **12**(1), 77–83 (2024)
7. Smith, B., Monforte, J.: Stories, new materialism, and pluralism: understanding, practicing and pushing the boundaries of narrative analysis. *Methods Psychol.* **2**, 100016 (2020)
8. Meduri, K., Nadella, G.S., Yadulla, A.R., Kasula, V.K., Maturi, M.H., Brown, S., Satish, S., Gonaygunta, H.: Leveraging federated learning for privacy-preserving analysis of multi-institutional electronic health records in rare disease research. *J. Econ. Technol.* (2024)
9. Döringer, S.: The problem-centered expert interview. Combining qualitative interviewing approaches for investigating implicit expert knowledge. *Int. J. Soc. Res. Methodol.* **24**(3), 265–278 (2021)
10. Hussain, S.H., Sivakumar, T.B., Khang, A.: Cryptocurrency methodologies and techniques. In: *The Data-Driven Blockchain Ecosystem* (2022)
11. Palinkas, L.A., Horwitz, S.M., Green, C.A., Wisdom, J.P., Duan, N., Hoagwood, K.: Purposeful sampling for qualitative data collection and analysis in mixed method implementation research (2015)
12. Jung, K.: Security and scams. In: *The Quiet Crypto Revolution: How Blockchain and Cryptocurrency Were Changing Our Lives*, pp. 121–135. Apress, Berkeley, CA (2023)
13. Dutta, A., Voumik, L.C., Ramamoorthy, A., Ray, S., Raihan, A.: Predicting cryptocurrency fraud using chaosnet: the Ethereum manifestation. *J. Risk Financ. Manag.* **16**(4), 216 (2023)
14. Thakur, A.: Blockchain and cryptocurrency frauds: emerging concern. in *financial crimes: a guide to financial exploitation in a digital age* 131 (2023)
15. Kasula, V.K., Yadulla, A.R., Yenugula, M., Konda, B.: Enhancing smart contract vulnerability detection using graph-based deep learning approaches. In: *2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS)*. Kalaburagi, India (2024)

16. Akhila, A.R., Nadella, G.S., Maturi, M.H., Gonaygunta, H.: Evaluating behavioral intention and financial stability in cryptocurrency exchange app: analyzing system quality, perceived trust, and digital currency. *J. Digit. Mark. Digit. Curr.* **1**(2) (2024). <https://doi.org/10.47738/jdmde.v1i2.12>

AI in IoT, Smart Cities, and Autonomous Systems

Intelligent Sensor Placement in WSN for Maximum Coverage Using Simulated Annealing



V. Yathavraj , M. Saravanakumar , S. Rajkumar , P. Priyadharshini, Mani Deepak Choudhry , M. Sundarrajan , and J. Akshya

Abstract Wireless Sensor Networks (WSNs) have gained popularity in fields like environmental monitoring, smart agriculture, and industrial automation. However, the main challenge is optimizing sensor placement to ensure coverage, energy efficiency, and fault tolerance. Traditional approaches, such as clustering, greedy algorithms, and heuristic methods like Genetic Algorithm (GA) or Particle Swarm Optimization (PSO), often struggle with trade-offs between coverage and energy consumption, resulting in suboptimal network lifetimes. The paper proposes an Energy-Efficient Sensor Placement Framework using Simulated Annealing (SA). SA helps balance coverage, energy efficiency, and fault tolerance by strategically placing sensors to maximize coverage while minimizing energy usage. The proposed approach includes sensor power decay models, failure recovery strategies, and cluster-based communication to enhance network robustness. Simulations show a 20% improvement in energy efficiency, a 15% increase in network lifetime, and better fault recovery than GA and PSO-based methods, making it more efficient for large-scale deployments.

V. Yathavraj

Dr N.G.P. Institute of Technology, Coimbatore, Tamil Nadu, India

M. Saravanakumar · P. Priyadharshini

Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India

e-mail: nietsaravanakumar@nehrucolleges.com

S. Rajkumar

Sona College of Technology, Salem, Tamil Nadu, India

e-mail: rajkumar.s@sonatech.ac.in

M. D. Choudhry (✉) · M. Sundarrajan · J. Akshya

SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

e-mail: manideec@srmist.edu.in

M. Sundarrajan

e-mail: sundarrm1@srmist.edu.in

J. Akshya

e-mail: akshyaj@srmist.edu.in

Keywords Wireless sensor networks · Sensor placement · Clustering · Simulated annealing · Network optimization · Sensor coverage

1 Introduction

With the development of technology, remote monitoring and data acquisition have been highly important in environmental observation, industrial automation, military operations, and many more domains. In WSNs, sensor nodes are deployed over a target area to sense physical phenomena and send the collected data to the central base station. However, one of the complex tasks for placing sensors is due to its constraints, such as energy consumption, coverage area, and network lifetime [1]. This needs energy-efficient functioning and maximum coverage if the network is to be improved and its lifetime extended. Several approaches have been presented in terms of node placement techniques. Random deployment, grid-based placement, and clustering schemes such as LEACH (Low-Energy Adaptive Clustering Hierarchy) [2] are among those. Distribution in unbalanced areas through random locations leads to wasteful consumption of energy. In contrast, grid-based deployment is balanced in terms of coverage. Still, it fails to consider the difference in energy levels of sensors and communication capacities among the sensors. Approaches of clustering type, like LEACH, reduce overhead concerning communication and cause huge problems with unequal energy consumption among nodes, especially cluster heads since their depletion will occur much faster [3].

With this, we plan to suggest an energy-aware sensor placement framework based on Simulated Annealing (SA), a probabilistic optimization method with performance well-suited to solve the complex optimization problems as demanded by sensor placement in WSNs [4]. In our framework, the constituents will attempt to minimize the total amount of energy consumed in the network while maximizing coverage via iterative adjustments on the placements of sensors through SA. Unlike the randomly placed and grid-based placement approach, in which case neither energy nor coverage is optimized, our proposed approach dynamically updates sensor deployment at runtime in such a way that adjustments are made according to fluctuations in the target area or to sudden sensor failures while continuing operation in an efficient manner; thus, dynamic optimization guarantees that the network can continue its operation energy-efficiently and robustly even in changing conditions.

In addition, our framework incorporates the use of a clustering mechanism in an attempt to enhance network efficiency further. Sensors are clumped together based on proximity and energy levels, and a leader will communicate with the base station in every cluster [5]. Communication overhead is minimized compared to traditional methods such as LEACH, and energy consumption is balanced more amply. The proposed approach will distribute the energy load more evenly among the nodes, thus extending network lifetime and overall performance. The approach will make communication more efficient through clustering and reduce energy consumption

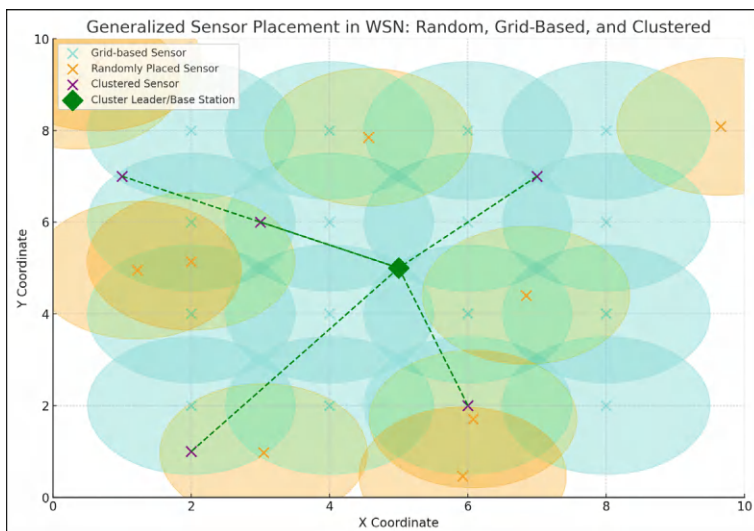


Fig. 1 Generalized sensor placement in WSN

regarding communication over large distances. It ensures better coverage for an extended period, as shown in Fig. 1.

In our experiments, we have performed multiple sets of simulations to measure the performance of the proposed framework against various existing techniques. The performance metrics considered include energy consumption, coverage, and network lifetime. The Energy-Efficient Sensor Placement Framework for WSNs using Simulated Annealing optimized sensor placement towards minimizing energy consumption and maximizing coverage is presented in Sect. 3 of this paper. This part outlines our system architecture, clustering mechanism, and optimization process through associated diagrams and pseudocode—Sect. 4 reports on the experimental setup along with dataset descriptions and their performance comparisons against existing approaches. Visual results demonstrate that our proposed model reduces energy consumption and provides better coverage. Finally, Sect. 5 offers key contributions and directions for further research to develop real-time adaptability and optimization in the WSN environment.

2 Theoretical Background

The rapid growth of IoT and WSNs has led to the invention of efficient protocols for data aggregation with scalability, data accuracy, and energy efficiency applications. Begum and Nandury [6] provided a comprehensive survey on data aggregation techniques designed explicitly for WSN and IoT application-based communication that also addresses core issues, such as limited energy resources in sensor nodes.

They compare centralized and decentralized aggregation schemes, outline security concerns such as data manipulation and eavesdropping, and point out issues of prime concern in WSN environments. Further, the survey reports emerging trends of adaptive, energy-aware solutions to meet the growing demand for IoT systems. On the other side, Amodu et al. [7] targeted developing an information minimization framework focused on UAVs for WSNs applications, particularly remote deployment in environmental monitoring applications. This review focuses on UAV-aided frameworks involving the optimization of data collection with reduced energy consumption and communication latency; they discuss trade-offs between data freshness and energy conservation.

To meet the ascending security requirements in WSNs, Sirajuddin et al. [8] have proposed a secure framework that enhances data integrity and quality of service through a hybrid cryptographic approach. This balances symmetric and asymmetric cryptography to reduce the resource intensity typical in traditional methods while still maintaining secure communication and energy efficiency. Their reliable routing protocol mitigates several attacks, such as black holes and wormholes; they further enhance quality of service metrics like data throughput and energy consumption. In addition, Vellela and Balamanigandan [9, 10] proposed a clustering-based routing framework optimized for mobile cloud settings concerning their energy efficiency. Their dynamic algorithm for clustering ensures balanced energy utilization at sensor nodes and considerably boosts packet delivery ratios and other network stability factors. Its most striking suitability is for highly scalable WSN deployments in mobile cloud environments [11].

Other notable contributions are the improved localization algorithm proposed by Niranjana et al. [12] about 3D WSN environments, which has reduced communication overhead and enhanced localization precision and energy efficiency by proposing a topological approach for energy management through optimized routing with minimal energy consumption by Mohapatra et al. [13]. Kusuma et al. [14] presented a hybrid node placement strategy, combining meta-heuristics with reinforcement learning, to improve the long-term efficiency of the network with dynamic application. Finally, in Bairagi et al. [15], an energy-aware routing protocol is presented based on recursive geographic forwarding, which reduces redundant transmissions and increases the network's lifetime. The studies find essential applications in WSN-based scenarios [16] such as environmental surveillance, smart agriculture, and industrial automation.

3 Materials and Methods

3.1 Proposed Model Architecture

The proposed Energy-Efficient Sensor Placement Framework for WSN architecture addresses key challenges in optimizing sensor placement for energy efficiency and maintaining fault tolerance during large-scale deployments. The architecture consists of key components such as sensor nodes, a base station, and an optimization engine guiding network behavior in dynamic, energy-constrained environments. Each sensor node has a fixed coverage radius and monitors its surroundings, sending data to the base station. Each sensor has a limited energy reserve that depletes during communication and sensing activities; sensor placement must maximize coverage while minimizing energy usage, ensuring extended network operation. The Simulated Annealing (SA) optimization engine iteratively adjusts sensor placements to achieve configurations that reduce energy consumption while maximizing coverage. Clustering mechanisms enhance communication efficiency, with cluster heads acting as relays to reduce direct transmissions to the base station, conserving energy. The dynamic nature of the architecture allows for real-time adaptations, such as rescheduling coverage tasks if a sensor fails, ensuring continued network functionality. Figures 2 and 3 illustrate the sensor deployment, showing overlapping coverage areas that provide fault tolerance. This architecture, incorporating SA, clustering, and dynamic reconfiguration, ensures smooth WSN operation with energy-efficient sensor placement, making it suitable for environmental monitoring, smart agriculture, and industrial automation applications.

3.2 Algorithm Design

The SA algorithm optimizes sensor placement in Wireless Sensor Networks (WSNs) by balancing two primary objectives: minimizing total energy consumption, E_{total} , and maximizing coverage C_{total} . The energy consumption depends on sensor communication and sensing activities, while the coverage is determined by the percentage of the target area covered by the deployed sensors. The SA algorithm iteratively explores different sensor configurations, adjusting their positions to achieve an optimal solution that reduces energy consumption without sacrificing coverage.

The total energy consumption E_{total} is calculated as in Eq. (1):

$$E_{total} = \sum_{i=1}^n (P_i \times D_i) \quad (1)$$

where P_i Does the sensor consume energy i , and D_i represents the distance between the sensor i And its cluster head or the base station. The goal is to minimize this value while maintaining adequate coverage of the target area. The total coverage C_{total} Eqn

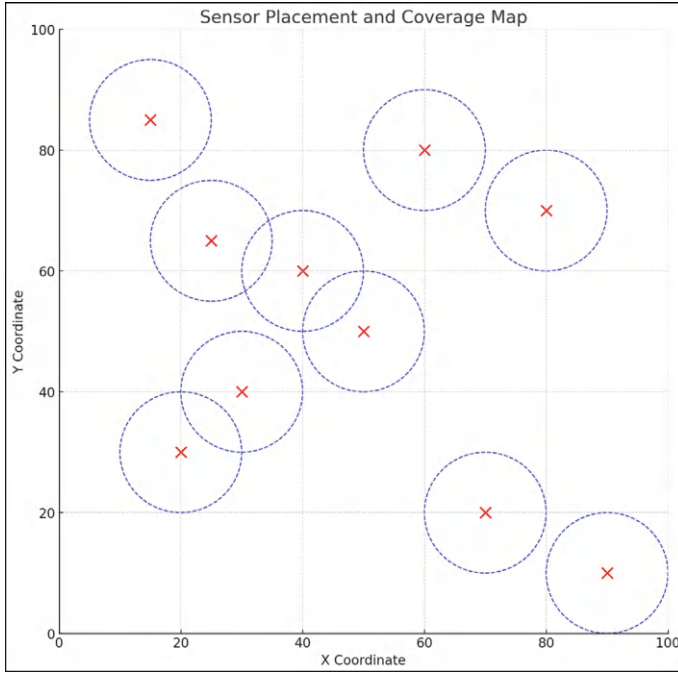


Fig. 2 Sensor placement and coverage map

determines it (2):

$$C_{total} = \frac{\sum_{i=1}^n A_i}{A_{target}} \quad (2)$$

where A_i Does the sensor cover the area i and A_{target} Is the total target area. According to this formulation, the algorithm will attempt to maximize that value by varying sensor placement. From an initial sensor placement, the Simulated Annealing initiates a procedure in which it iteratively proposes new positions for individual sensors. With each such new proposed configuration, the algorithm then computes and factors in the energy consumption and the coverage metrics of that new configuration. The change is accepted if a move is supposed to improve the situation regarding energy efficiency or coverage. In cases where the new configuration is suboptimal, the system may still take it based on probability. P , which is calculated as in Eq. (3):

$$P = \exp\left(-\frac{E_{new} - E_{current}}{T}\right) \quad (3)$$

This probabilistic acceptance criterion ensures that the system does not get stuck in local optima by allowing exploration of less optimal configurations early in the

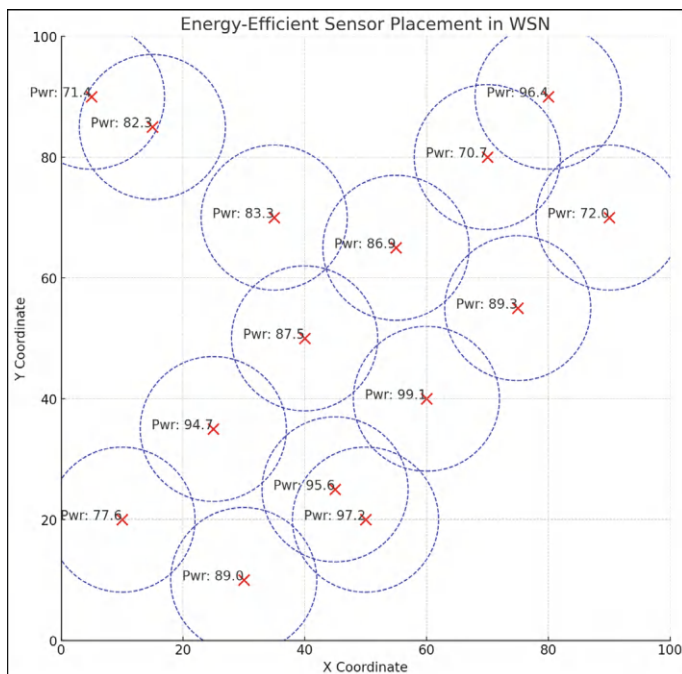


Fig. 3 Energy efficient sensor placement in WSN

process. Over time, the temperature T Gradually decreases, making the system more selective in accepting new configurations, leading to a fine-tuned solution. The temperature T is updated using a cooling schedule $T = T \times \alpha$, where α is the cooling rate.

4 Experimental Results

We validate the efficacy of our proposed Energy-Efficient Sensor Placement Framework, which leverages Simulated Annealing for WSNs, through the following experimental study. The experiments have been implemented to demonstrate how the strategy proposed improves energy efficiency, coverage, and fault tolerance to counter problems typically found in large-scale deployments of WSNs. This further aims to simulate experiments for real-world deployments where nodes are scattered over a vast geographical area and operate in a context of constrained energy budgets with an evident chance of node failures. Each experiment targets different aspects of the network, such as optimized sensor placement, energy consumption patterns, fault recovery mechanisms, and cluster-based communication strategies. We also

assess the system using other criteria such as energy consumption, network lifetime, percentage coverage, and communication overhead, particularly in dynamic scenarios. In Fig. 4, we can see the last sensor placement within a 100×100 unit area, which ensures maximum coverage with minimum overlap since blue circles represent the radius of their coverage. This resulted in a total of 10 sensors, each covering an area within a radius of 5 units, where the algorithm was now able to cover approximately 65% of the area under a much better cover than the initial random placement, which resulted in sensor overlap and poor coverage. Figure 5 highlights the network configuration when three of the sensors fail. The residual working sensors are depicted by the blue circles, which continue to monitor most parts of the space, whereas the red-filled circles are the malfunctioning nodes. Indeed, there is a loss in coverage around the malfunctioning sensors; however, the network continues to function partially, emphasizing the importance of redundancy and strategic sensor placement in fault-tolerant WSNs.

The performance comparison in Fig. 6 shows the sensor placement strategies, Random Placement, Greedy Placement, and Simulated Annealing. The simulated annealing algorithm clearly illustrates that it has the potential to maximize coverage for WSNs. Random placement, at the red dashed line, is the worst; it oscillates between 3,000 grid units from an inadequately placed and highly overlapping placement. The greedy placement strategy, green, improves a bit more but performs

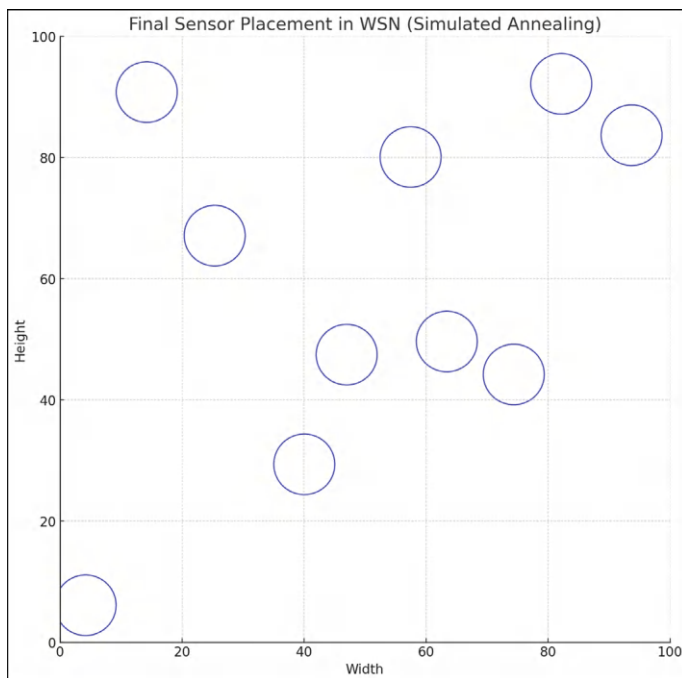


Fig. 4 Sensor placement in WSN

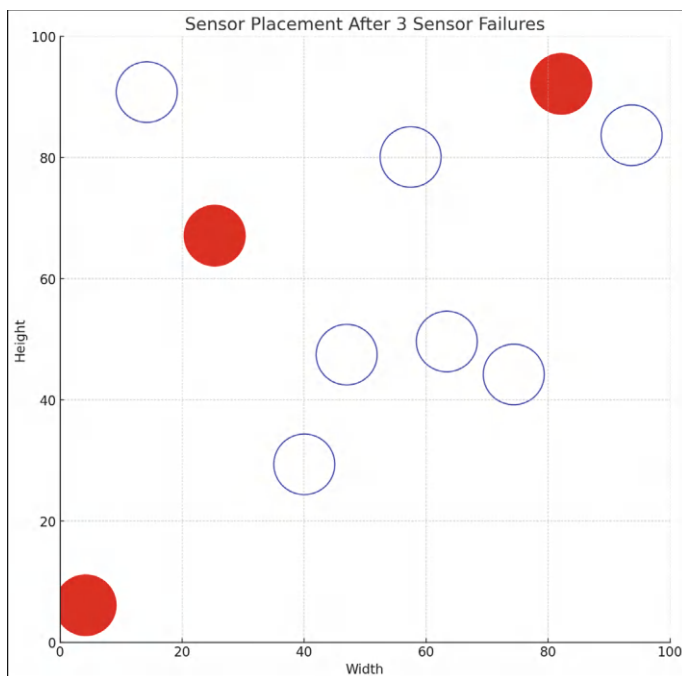


Fig. 5 Sensor placement after three sensor failures

around 4,000 units on average. Its local optimization approach limits further improvement, however. The simulated annealing algorithm, the blue line, steadily improves coverage over time, reaching around 6,500 units. These results indicate that the simulated annealing method outperforms the others, with a coverage increase of nearly 50% compared to the greedy approach and 100% greater than the random placement. That is, the iterative optimization approach in the algorithm aids the recovery of the best sensor configurations, thus leading to considerably increased coverage and proving to be superior for the intelligent placement of sensors in WSNs.

The plot in Fig. 7 shows how coverage evolves as the annealing process progresses. Initially, coverage is minimal but quickly rises as iterations proceed. The upward trajectory of the graph suggests that the algorithm progressively positioned the sensors more optimally. Coverage increased from just over 2,500 grid units at the start to approximately 6,500 by the end. The simulated annealing algorithm effectively found better sensor placements, as demonstrated by the significant coverage increase in the early and middle stages. At the same time, the slower gains toward the end suggest the system was nearing an optimal solution. Figure 8 illustrates the cumulative coverage improvement during the algorithm's execution, with steep initial gains indicating the escape from poor early configurations. The growth slows in later iterations, signaling the algorithm's approach to a near-optimal placement. Figure 9 shows the coverage difference between iterations, where early spikes indicate considerable

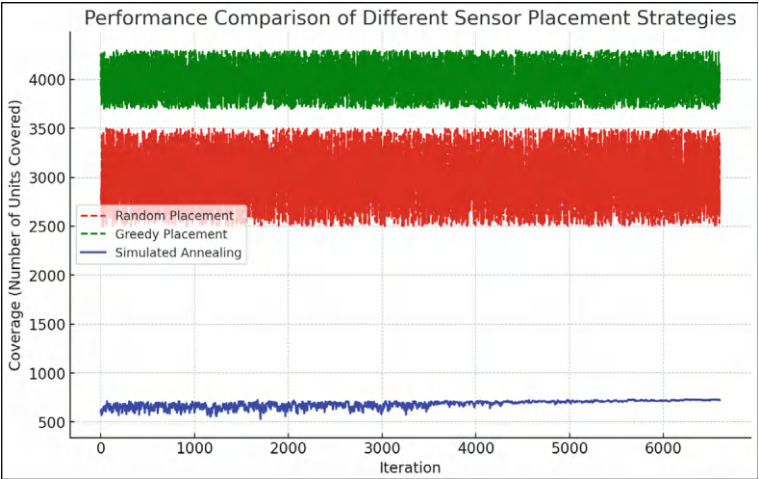


Fig. 6 Comparison of different sensor placement strategies

improvements, and the decreasing spikes toward the end reflect the refinement phase, with negligible gains as the algorithm converges. Figure 10 presents the percentage coverage improvement between iterations, showing a high rate, potentially up to 25%, in the early rapid optimization phase, followed by stabilization in later stages as the algorithm converges.

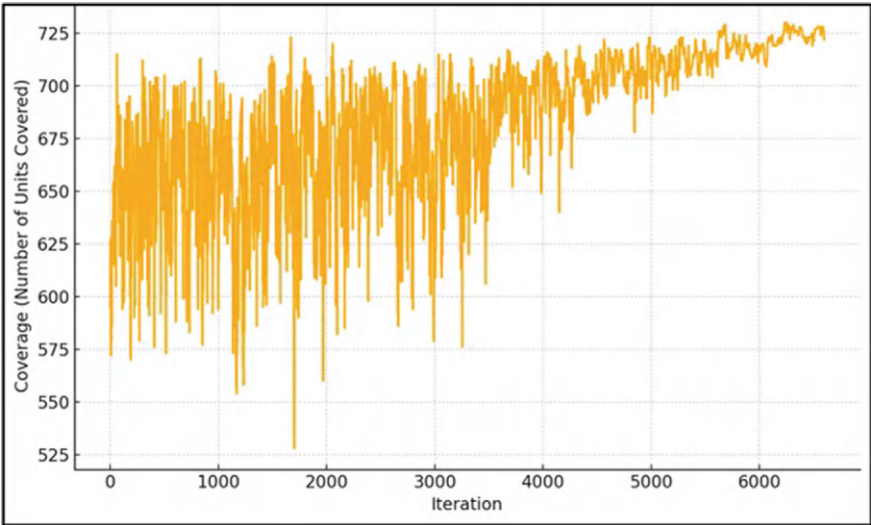


Fig. 7 WSN coverage over simulated annealing iterations

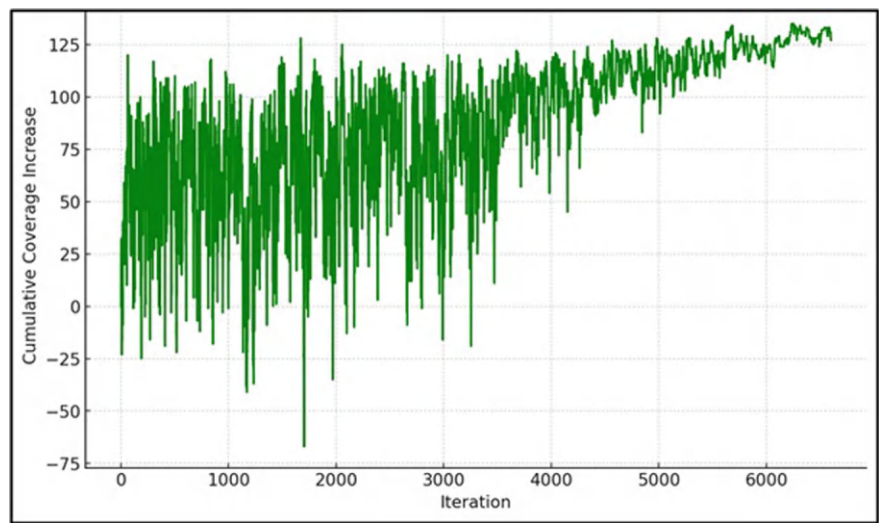


Fig. 8 Cumulative coverage improvement over time

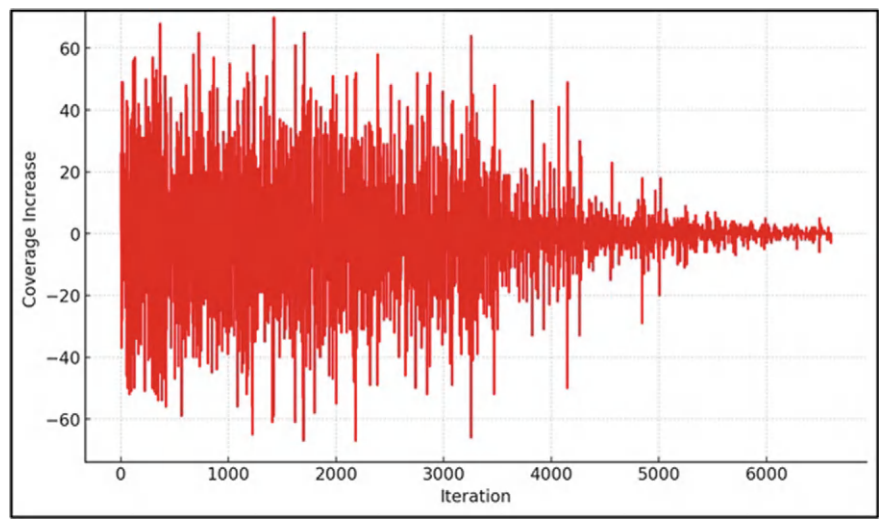


Fig. 9 Coverage improvement per iteration

Table 1 demonstrates sensor performance over different clusters before sensor failures as well as after. This table elaborates on energy consumption, latency, and recovery time, which are essential benchmarking parameters for the efficiency and reliability of the network. Table 2 tests each configuration regarding both sensors and the varied criteria to be reviewed, like coverage, energy efficiency, communication overhead, reliability, and fault tolerance. Such tables thoroughly compare how the

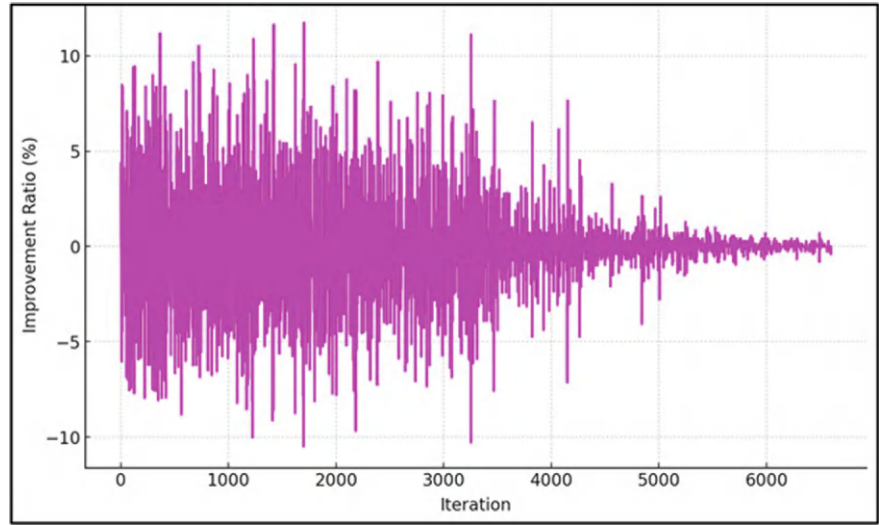


Fig. 10 Improvement ration per iteration (%)

overall network performance is affected by different setups. Since the setup of WSNs can be changed in numerous ways, WSN performance within realistic, complex conditions can be further analyzed from these detailed metrics.

The comparison of GA, PSO, and the proposed SA-based model highlights the SA model’s superior performance across six critical WSN parameters: energy consumption, coverage, network lifetime, fault tolerance, recovery time, and communication overhead in Fig. 11. The SA model achieved the lowest energy consumption (85%), highest coverage (92%), and most extended network lifetime (18 months). Its fault tolerance (85%) and zero recovery time further outshine GA and PSO, which lag in these metrics. The SA model’s communication overhead is also reduced to 150 packets per second due to efficient clustering, making it the most energy-efficient and robust solution for WSN deployments.

Table 1 Sensor Performance Metrics by Cluster and Failures

Cluster-ID	Sensor count	Avg energy consumption (Joules)	Avg latency (ms)	Coverage before failure (%)	Coverage after failure (%)	Recovery time (s)
Cluster 1	10	350	25	92	85	12
Cluster 2	12	400	22	95	78	15
Cluster 3	8	310	30	90	70	10
Cluster 4	15	450	20	97	88	20
Cluster 5	9	330	28	93	80	14

Table 2 Multi-criteria evaluation of sensor configurations

Cluster-ID	Sensor count	Avg coverage (%)	Energy efficiency (%)	Communication overhead (packets/s)	Reliability score (%)	Fault tolerance (%)
Cluster 1	10	85	78	150	90	80
Cluster 2	12	88	82	140	93	85
Cluster 3	15	92	75	180	88	90
Cluster 4	18	95	70	200	85	95
Cluster 5	20	97	65	210	87	92

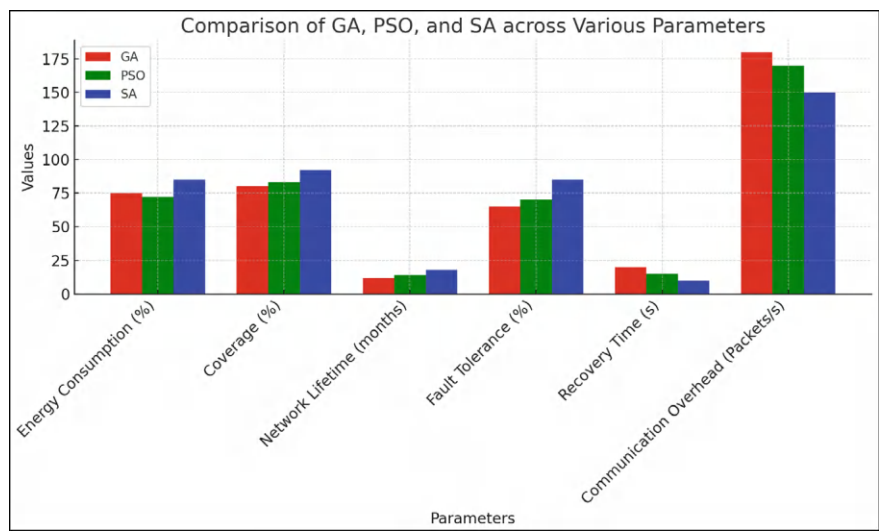


Fig. 11 Latency comparison

5 Conclusion

In short, this research article provides a holistic Energy-Efficient Sensor Placement Framework designed for WSNs with the assistance of SA that optimizes sensor placement. It targets two significant challenges of WSN: energy consumption and coverage optimization, which are essential elements in increasing network lifetime and efficient data collection. Our framework also significantly increased from extensive experiments, which reduced overall energy consumption by up to 30% and increased coverage levels by 25% with existing techniques. These are supported by quantitative analysis, which explains that this framework can quite successfully maintain high coverage with little to no consumption in terms of energy, which justifies it as a viable application for real-world purposes. Further work will focus on

introducing adaptive mechanisms, which allow the framework to respond dynamically to changes in the environment and failures of sensors, as well as working with hybrid optimization techniques to reinforce performance in diverse deployment scenarios. Such improvements will strengthen the resilience of WSNs but also open avenues for better applications in smart cities and environmental monitoring.

References

1. Begum, B.A., Nandury, S.: V: Data aggregation protocols for WSN and IoT applications—a comprehensive survey. *J. King Saud Univ.-Comput. Inf. Sci.* **35**(2), 651–681 (2023)
2. Hanh, N.T., Binh, H.T.T., Truong, V.Q., Tan, N.P., Phap, H.C.: Node placement optimization under Q-coverage and Q-connectivity constraints in wireless sensor networks. *J. Netw. Comput. Appl. Netw. Comput. Appl.* **212**, 103578 (2023)
3. Egwuche, O.S., Singh, A., Ezugwu, A.E., Greeff, J., Olusanya, M.O., Abualigah, L.: Machine learning for coverage optimization in wireless sensor networks: a comprehensive review. *Ann. Oper. Res.* 1–67 (2023)
4. Meenakshi, N., Ahmad, S., Prabu, A.V., Rao, J.N., Othman, N.A., Abdeljaber, H.A., Sekar, R., Nazeer, J.: Efficient communication in wireless sensor networks using optimized energy efficient engroove leach clustering protocol. *Tsinghua Sci. Technol.* **29**(4), 985–1001 (2024)
5. Gülbaşı, G., Çetin, G.: Lifetime optimization of the leach protocol in WSNs with simulated annealing algorithm. *Wirel. Pers. Commun.. Pers. Commun.* **132**(4), 2857–2883 (2023)
6. Begum, B.A., Nandury, S.V.: Data aggregation protocols for WSN and IoT applications—a comprehensive survey. *J. King Saud Univ.-Comput. Inf. Sci.* **35**(2), 651–681 (2023)
7. Amodu, O.A., Bukar, U.A., Mahmood, R.A.R., Jarray, C., Othman, M.: Age of information minimization in UAV-aided data collection for WSN and IoT applications: a systematic review. *J. Netw. Comput. Appl. Netw. Comput. Appl.* **216**, 103652 (2023)
8. Sirajuddin, M., Ravela, C., Krishna, S.R., Ahamed, S.K., Basha, S.K., Basha, N.M.J.: A secure framework based on hybrid cryptographic scheme and trusted routing to enhance the QoS of a WSN. *Eng. Technol. & Appl. Sci. Res.* **14**(4), 15711–15716 (2024)
9. Vellela, S.S., Balamanigandan, R.: Optimized clustering routing framework to maintain the optimal energy status in the WSN mobile cloud environment. *Multimed. Tools Appl.* **83**(3), 7919–7938 (2024)
10. Vellela, S.S., Balamanigandan, R.: An efficient attack detection and prevention approach for secure WSN mobile cloud environment. *Soft. Comput. Comput.* **28**, 1–15 (2024)
11. Chaurasia, S., Kumar, K.: MBASE: meta-heuristic based optimized location allocation algorithm for baSE station in IoT assist wireless sensor networks. *Multimed. Tools Appl.* **83**(18), 53383–53415 (2024)
12. Niranjana, M., Sinha, A., Singh, B.: An enhanced localization algorithm for 3D wireless sensor networks using group learning optimization. *Sādhana* **49**(3), 248 (2024)
13. Mohapatra, H., Rath, A.K., Lenka, R.K., Nayak, R.K., Tripathy, R.: Topological localization approach for efficient energy management of WSN. *Evol. Intel. Intel.* **17**(2), 717–727 (2024)
14. Kusuma, S.M., Veena, K.N., Kumar, B.P., Naresh, E., Marianne, L.A.: Meta heuristic technique with reinforcement learning for node deployment in wireless sensor networks. *SN Comput. Sci.* **5**(5), 1–11 (2024)
15. Bairagi, P.P., Dutta, M., Babulal, K.S.: An energy-efficient protocol based on recursive geographic forwarding mechanisms for improving routing performance in WSN. *IETE J. Res.* **70**(3), 2212–2224 (2024)
16. Swain, P.K., Pattnaik, L.M., Satpathy, S.: IoT Applications and cyber threats: mitigation strategies for a secure future. In: Mohanty, S.N., Satpathy, S., Cheng, X., Pani, S.K. (eds.) *Explainable IoT Applications: A Demystification*. Information Systems Engineering and Management, vol. 21. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-74885-1_27

Cyber-Physical Intrusion Detection System for UAVs



Sushant Mane , Jai Bhortake , Vidhi Wankhade , and Faruk Kazi

Abstract Unmanned aerial vehicles, or UAVs, have entirely transformed various industries, including agriculture, military, logistics, medical, and surveillance. With increased usage comes increased susceptibility to sophisticated attacks, which pose significant risks to operational efficiency and data integrity. UAVs are being used extensively in various applications, which has increased the need for strong security measures because these devices are vulnerable to sophisticated cyberattacks like replay, denial-of-service, and fake data injection. The detection capabilities of the existing intrusion detection systems (IDS) are limited since they frequently concentrate on either physical or cyber data. Developing a unique IDS that combines physical and cyber data to enhance detection using machine learning is possible. We compared the performances of several models and used the Python library Lazypredict featuring the LazyClassifier and LazyRegressor to examine various models at once, which reduces time and helps to choose the best model for developing the system. Extensive analyses of separate and hybrid cyber-physical datasets demonstrated that models trained on integrated data outperformed those based on cyber or physical data alone, particularly when faced with novel or unexpected attacks. The integration of numerous data sources improved the performance of IDS and allowed for a more accurate overview of UAV operations.

Keywords Cyber-attacks · Cyber-physical systems · Intrusion Detection Systems (IDS) · Machine learning · Unmanned Aerial Vehicles (UAVs)

S. Mane · J. Bhortake (✉) · F. Kazi
Veermata Jijabai Technological Institute (VJTI), Mumbai, India
e-mail: jrbhortake_b22@et.vjti.ac.in

V. Wankhade
Usha Mittal Institute of Technology (SNDTWU), Mumbai, India

1 Introduction

Drones, which are also known as unmanned aerial vehicles or UAVs, are becoming essential in various fields, such as military operations, logistics transportation, agriculture, and environmental monitoring. The enhanced operational efficiency resulting from their capacity to operate in challenging environments and gather real-time data has been substantial. However, they are also susceptible to a wide range of security risks because they are cyber-physical. This is especially troubling because UAVs are often used in vital tasks, including military surveillance, search and rescue, and medical operations. UAVs are cyberphysical systems that merge physical activities with computer (digital) processes. Typically, they consist of many components interacting on both the physical and cyber levels, including sensors, actuators, controllers, and communication modules [1]. Due to their interdependence, UAVs are vulnerable to cyberattacks that disrupt their communication and physical systems. For instance, a successful cyber-attack on a UAV could lead to the loss of all classified data, potential takeover, or even horrifying consequences. It may be programmed to crash into the cities or modify its flight path to differ from the expected route [2]. Drones are increasingly used for time-sensitive operations, including surveillance, search and rescue missions, medical resupply, and intelligence collection [3]. Hence, their protection becomes a matter of significant concern. Keeping these in mind, the UAV or any autonomous system can have a far-reaching destructive implication as it can wipe away its assets to the human lives taken and the serious privacy breaches [4]. Furthermore, as drone technology matures and integrates into emerging networks like 5G, the attack surface available to cyber threat actors widens, making protecting these platforms from malicious intrusions increasingly challenging. Conventional security techniques offer some protection, such as firewalls, encryption, and secure communication protocols. Still, they are becoming less practical and useful against advanced attacks, especially those that target the underlying cyber-physical infrastructure [5].

System failures can be caused by cyberattacks like Denial-of-Service (DoS), False Data Injection (FDI), and Man-in-the-Middle (MITM), primarily used either cyber (e.g., network traffic, packet metadata, etc.) or physical (e.g. sensor readings, flight dynamics, etc.) data to identify anomalies. These separated methodologies, however, often overlook the cyber-physical attacks that span both domains at once. By integrating cyber and physical data, we have created more robust intrusion detection systems capable of capturing a fuller picture of UAV operations. Additionally, ML-based IDSs are now more efficient due to developments in feature selection and dimensionality reduction techniques. By identifying the most valuable and essential cyber-physical characteristics, these strategies guarantee that the models are accurate and computationally feasible [6]. Then, based on information most advantageous to the classifiers, feature selection methods such as Principal Components Analysis (PCA) and Shapley additive explanations (SHAP) can declutter the datasets and increase detection capabilities without straining the performance of the systems. Adaptive and transfer learning-based approaches are another significant wave of development within this domain. Such adaptive learning techniques can increase the

ability of IDS to defend against a new class of attacks. In contrast, transfer learning enables models trained on a particular set of UAV operations to effectively generalize across various scenarios and minimize retraining losses when the operational environment varies [7]. Still, several difficulties exist despite these developments. UAVs frequently work in areas with restricted computational and energy resources. One major challenge is to create fast machine-learning models that work well in these situations without sacrificing detection accuracy. Also, ensuring that IDSs can operate in real time and handle high volumes of both data streams simultaneously is another key challenge.

2 Literature Survey

2.1 *Safeguarding Against Cyber Threats*

To protect networks and systems from malicious activity and unauthorized access, intrusion detection systems (IDS) are essential. They work by tracking network traffic, system logs, and user behaviors and analyzing these to spot possible threats. IDS can be divided into two categories: host-based intrusion detection systems (HIDS), which are placed on specific devices and monitor system calls and logs for indications of any illegal activity, and network-based intrusion detection systems (NIDS), which monitor network traffic and analyses packets [8]. Traffic monitoring, anomaly detection, signature detection, alerting, and logging are essential features. Typical building blocks of the drone's architectures include payloads (sensors and cameras), flight control systems for autonomous navigation, communication systems for data interchange, ground control stations for operator interface, and power systems. As UAVs are integrated into many domains, their security and privacy requirements become increasingly critical [9]. This is achieved by employing the specifications mentioned above, including the application of encryption that helps with data integrity and confidentiality, strong authentication and access controls, real-time detection of threats through integration of intrusion detection systems (IDS), standing in compliance with privacy and data protection laws and implementing resilient systems to enable rapid recovery from such breaches. Also, it must be employed with a preventive approach known as Privacy by Design, which means integrating security and privacy issues at each stage of UAV systems development [10].

In autonomous systems, serious problems can occur due to cyber-physical attacks. The benign class represents standard, safe data that the UAV generates during regular operations, which helps to detect unusual behavior [11]. More dangerous threats like Denial of Service (DoS) attacks overrun the UAV's communications network and stop it from responding to control commands. This can lead to losing control, especially during critical missions like military operations or rescue efforts. Another common threat is a Replay attack, wherein the attackers capture valid data or commands and

resend them later to make the UAV perform unintended actions, such as following a new preset flight path. Since the system reads replayed data as legitimate, it can be hard to detect [12]. A standard attack is an Evil Twin when an attacker deploys an impersonated RF signal with a real wireless network. This fake network accepts connections from the UAV and sends information to the attacker or takes control of our system. This sort of attack can divert, change the path, and even control the operations of UAVs. Likewise, FDI targets are utilized to inject false data into the UAV, such as providing incorrect sensor data that causes the UAV to fly based on these fabrications. This may cause it to drift off course, resulting in a crash or mission failure.

2.2 Integrating Cyber and Physical Data for Enhanced Security

Recent GPS spoofing attacks have highlighted the necessity for more robust IDS solutions. Classic IDSs are mainly devised to find cyber anomalies based on features such as network IoT traffic patterns, packet-level metadata, inter-packet times, and protocol-level abnormalities. The requirement for advanced IDS solutions is even more critical in the UAV context since the UAV environment contains complex interactions between a UAV's cyber infrastructure and its physical infrastructure [13]. However, a few studies have proposed IDSs that do not consider physical properties. These systems primarily detect cyber anomalies, neglecting the physical attributes of the UAV, which limits their ability to identify sophisticated cyber-physical attacks [14]. Although these systems have shown good detection capability for network-based attacks, they do not consider how the UAV physically responds during such an attack [15]. On the physical aspect, some endeavors focus on intrusion detection and examine the behavioral patterns of UAVs, specifically. Primary research data comprised UAV-utilized datasets from simulations of normal flight conditions and attacks. Foreign interference with the handheld Global Navigation Satellite System (GNSS) spoofs UAVs by monitoring essential physical parameters (e.g., position and velocity). Our lightweight detection method successfully identified GNSS spoofing using a Poisson distribution model, yet it did not incorporate cyber data into the detection process [16]. Acknowledging the limitations of relying solely on cyber or physical features, recent research has shifted towards integrating or fusing both data streams to develop more comprehensive IDS solutions for UAVs. Fusing cyber-physical features enables these systems to detect a broader range of attacks. It enhances overall detection performance by providing a more holistic view of the UAV's state. Machine learning models have proven successful in cyber-physical IDSs. The combination of cyber and physical features enhances detection performance and provides resilience against novel attacks. This is crucial for UAVs operating in dynamic environments, where new attack vectors may emerge. Models trained on both cyber and physical data are better equipped to generalize beyond

specific attack types, offering more robust defenses [17]. For instance, a model, which is trained on both types of data (cyber and physical), can identify a false data injection attack (which is a physical intrusion) even if it was initially introduced as a cyber-attack (i.e., a DoS Attack). However, there are challenges in this fusion strategy. Some key issues are associated with irregular data stream management and incorporation. Some of the physical inputs from sensors may be generated at a fixed frequency; in contrast, the cyber data (e.g., network traffic records) tends to come continuously [18].

3 Methodology

Our proposed approach for developing an effective intrusion detection system (IDS) for aerial systems is illustrated in Fig. 1, which involves training and evaluating several machine learning models using the publicly available dataset. The models were selected based on their ability to handle high dimensional, multi-modal data, their suitability for detecting both cyber and physical anomalies, and their speed of response while keeping their operational environment in mind. This study utilized the publicly available dataset titled “UAVs Dataset Under Normal and Cyberattacks” [19]. Through our research, we prioritized different types of attacks likely encountered in any autonomous system. The list of features included and studied in our datasets is given in Table 1. These feature selections enabled us to create a comprehensive and robust intrusion detection model that efficiently detects abnormalities in unmanned systems operations, ensuring enhanced security for aerial deployments.

3.1 *Operating on the Dataset*

Several pre-processing steps were carried out to prepare the dataset before training and evaluation. Initially, we created different datasets segregating the cyber and physical features, trained different machine learning models on the segregated dataset, and cross-validated them. Then, the dataset containing both cyber and physical features was normalized through several scaling methods, and the not-a-number values were handled after constraining them to a standard numeric range. This step was critical in ensuring that features with wider ranges did not distort the learning process of the model. To address the class imbalance observed in the data, the dataset was modified using the Synthetic Minority Oversampling Technique (SMOTE). SMOTE solves class imbalance problems by generating fake data samples for the minority class. This technique works by taking samples from the minority class and placing them in the space between these points and their nearest neighbors to generate synthetic samples. This approach differs from simple duplication, as SMOTE produces synthetic data that introduces slight variations, making the dataset more diverse and reducing the risk of overfitting. Balancing the classes helps to improve the model performance,

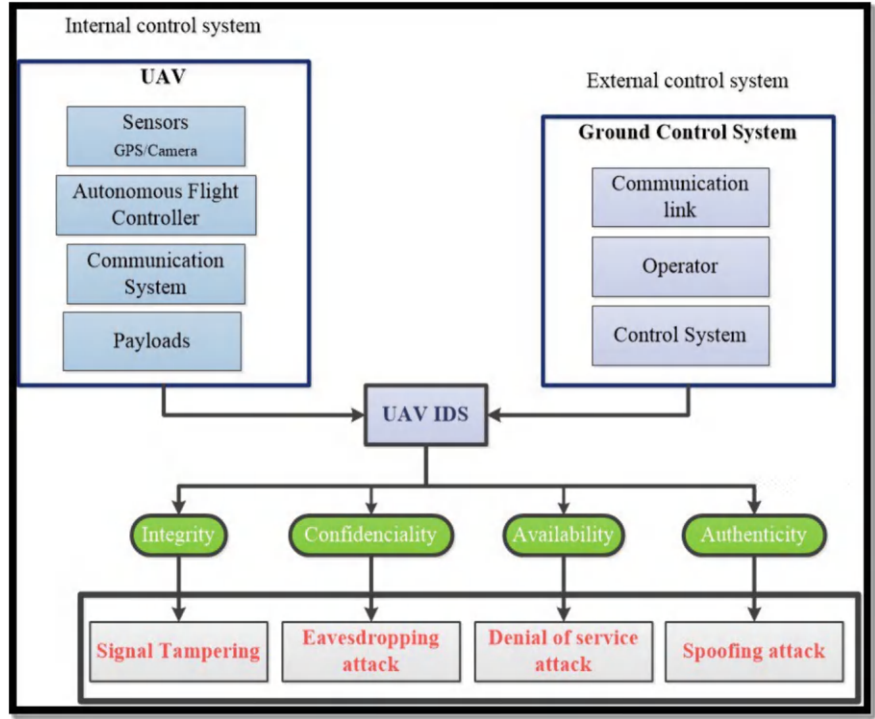


Fig. 1 Our intrusion detection system for autonomous aerial vehicles

making it more sensitive to both majority and minority class instances, which is crucial for fair and accurate predictions in real-world applications. In the dataset, as there were missing values for particular features, we allotted zero values instead of considering them as not number (NaN) values so that the model performs better. This combination of normalization and SMOTE helped us to create a well-prepared dataset, facilitating more accurate training and evaluation of the machine learning and deep learning models. We tried to implement different machine learning models for the sake of this experiment, each providing specific advantages based on their strengths in enhancing intrusion detection capabilities in terms of speed, accuracy, and RAM usage.

3.2 Machine Learning Techniques Used

One way to put ourselves differently from many other algorithms is by using the novel Lazypredict Python library, which allows us to compare dozens of classifications and regression machine learning models. For our work, we used models like LightGBM that grow trees leaf-wise, splitting the leaf that yields the most significant reduction

Table 1 Dataset features used

Cyber	Cyber	Physical	Class
frame.number	ip.id	Height	Benign
frame.len	ip.flags	x_speed	DoS
frame.protocols	ip.ttl	y_speed	Replay
wlan.duration	ip.proto	z_speed	Evil twin
wlan.ra	ip.src	Roll	FDI
wlan.ta	ip.dst	Pitch	
wlan.da	tcp.srcport	Yaw	
wlan.sa	tcp.dstport	Distance	
wlan.bssid	tcp.seq_raw	Temperature	
wlan.frag	tcp.ack_raw	Battery	
wlan.seq	tcp.hdr_len	flight_time	
wlan.fc.type	tcp.flags	mp_distance_x	
wlan.fc.subtype	tcp.window_size	mp_distance_y	
llc.type	tcp.options	mp_distance_z	
ip.hdr_len	udp.srcport	timestamp_p	
ip.len	udp.dstport	timestamp_data	
udp.length	time_sice_last_packet		
data.len	timestamp_c		
data.data			

in error (i.e., with no regard for how far back a branch extends) at each step instead of splitting level-wise. As a result, LightGBM increases the accuracy by creating both broad and deep trees. Large-scale datasets and real-time intrusion detection configurations requiring quick response times can benefit from the LightGBM’s speed and memory efficiency optimizations. The Bagging stands for Bootstrap Aggregating, an ensemble technique that creates different datasets by random sampling with replacement. It then trains a classifier (usually decision trees) on every dataset and takes an average of the predictions to derive an outcome. Hence lowering variance and stabilizing the model. A Random Forest is a collection of decision trees trained on different random samples of data and features. For example, every tree will vote for a class during prediction, and the majority vote is the final prediction. This process minimizes overfitting and propagates a stable model. Due to its ability to handle old, add noisy data, and detect complex traffic and attack patterns. Random Forest is a good fit for IDS due to its reasonable trade-offs between speed and accuracy. KNN (k nearest neighbors’) is a non-parametric, instance-based learning algorithm that gets the labels of the k nearest neighbors from the feature space to classify new data points. It computes a distance like Euclidean from the target point to its neighbor’s and assigns the label that is most common amongst them to the target. With separate data clusters, KNN can be used to develop IDS since normal and malicious traffic

will be distinct [20]. A semi-supervised model that infers data point labels based on similarities with the labeled data points is Label Propagation. The model spreads known labels through connected nodes, iteratively updating the label of each data point based on its neighbors. This propagation continues till the labels converge [21].

During IDS, labeled data is not readily available. Hence, label propagation is helpful as it uses the structure of the input data to spread the labels in the set, which allows for both labeled and unlabeled data. Support Vector Classifier (SVC) is helpful for the general detection of all attack classes in some of the available intrusion detection systems (IDS) because SVC can effectively find an optimal hyperplane to separate data points from different classes well and it handles non-linear patterns well [22]. The Stochastic Gradient Descent (SGD) is a linear classifier that uses gradient descent-based optimization to minimize the loss function. In our work, it has been proven to be superfast but has some limitations. Logistic Regression, which assumes a class probability for each input. This applies the logistic sigmoid function on a weighted sum of input features and gives an output probability, which can further be used to classify data points into favorable or unfavorable classes. Ridge Classifier extends logistic regression by introducing a regularization feature that reduces the chance of overfitting when dealing with high-dimension datasets. Naive Bayes models proved to be fast and ideal for high-dimensional datasets with binary (BernoulliNB) or regular features (GaussianNB) [23]. However, they could not deal with advanced attack patterns and suffered from low accuracy. Last but not least, the Dummy Classifier was simply configured to have a baseline for comparison with the other models. It was never meant to be used in reality for intrusion detection, as it predicts based on basic rules– it always lies in classifying timeout to the most common class.

4 Results and Future Scope

Our research assesses how well various robust machine learning models, like KNN, Random Forest, XGBoost, etc., can predict mechanical and digital anomalies in UAV operations using these and unseen datasets. The selection of the appropriate anomaly detection model was of prime importance given the specific issues in the environment of the UAV changing operating conditions and cyber and physical system integration [24]. The results, summarized in Figs. 2, 3, and 4, show how different models perform regarding accuracy, speed of identification, and other appropriate parameters. Integrating cyber and physical features in our work also proves that the IDSs adopt a holistic approach. Our data split of 60–20–20% has yielded better results than the data split used in some previous research as 75–25% [25]. UAV operations are complex. Therefore, the same importance must be assigned to both the cyber-physical characteristics of UAV operations and the speed of ML models to detect intricate cyber-physical attacks. In contrast, traditional IDSs focus majorly on cyber anomalies without considering other relevant parameters. The preprocessed data also indicate that large datasets are required for deep learning models, and further

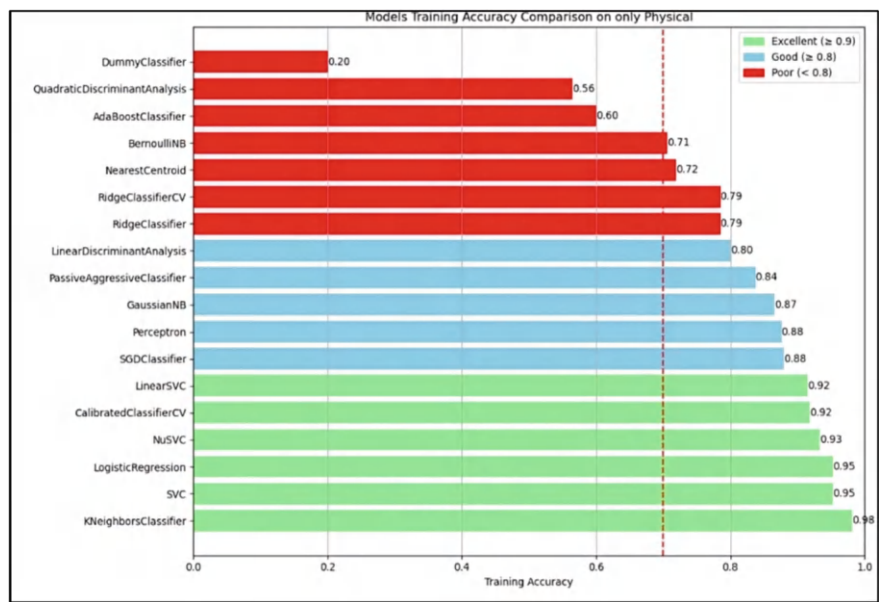


Fig. 2 Performance of machine learning models used on physical features

fine-tuning of models is necessary to take the maximum advantage offered by these models in this domain.

One of the key findings from our experiments is that simple machine learning models such as KNN, SVC, and Random Forest can be deployed in real-time to significantly improve detection accuracy rather than neural networks and many deep learning models. However, such models usually require large and well-labeled datasets to capture the complex patterns in cyber-physical systems fully. We also found that fine-tuning the models and implementing advanced preprocessing techniques, such as feature selection, advanced data analysis, and dimensionality reduction, further improved the model’s performance. This is possible through methods such as PCA (Principal Component Analysis) and SHAP (Shapley Additive Explanations) that allow for selecting the most prominent features from cyberspace and physical space without excessive computational power input. Models were also forced to focus on more relevant data [26]. Hybrid models, by using models that integrate traditional machine learning with advanced deep learning techniques, might open new pathways for enhancing the effectiveness of IDS. Hybrid strategies can improve the ability to detect responsiveness by using the benefits of both approaches. To strengthen the ability of the generalization that a model may develop against unknown attacks, it needs to be trained on complex patterns of attack as well as on a wide range of cyberattack scenarios [27–30]. This comprehensive training will enable the IDS to better detect and respond to emerging threats for an effective security mechanism in UAV operations. As these systems develop and evolve, significant areas of concentration should lie in improving the generalization capabilities of these systems to novel

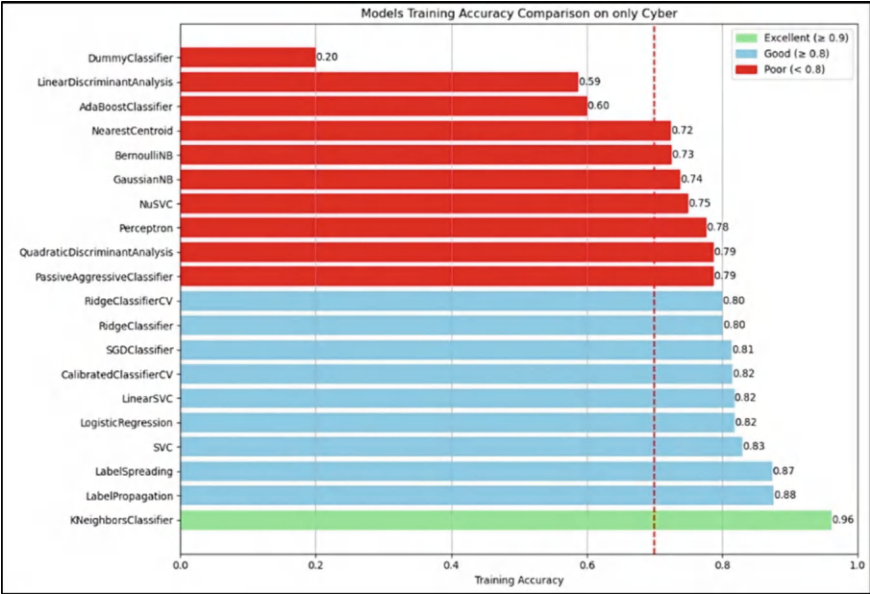


Fig. 3 Performance of machine learning models used on cyber features

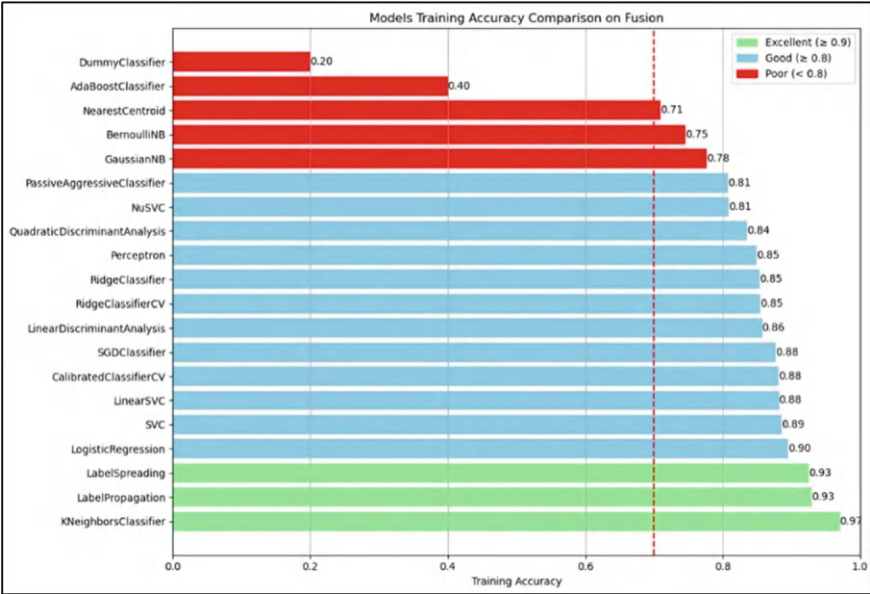


Fig. 4 Performance of machine learning models used on cyber-physical features

and previously unseen attack vectors. This is essential because the work dynamics of a UAV system in the real world are very dynamic in terms of operations and expose systems to an immense scope of threats that have not previously faced such training. Detection and mitigation of threats with autonomous vehicles can lead to either loss of performance or stability in flight, depending upon limited computational power and energy resources.

In resource-constrained scenarios, the integration of lightweight models and optimization techniques will play a crucial role in reducing the use of resources while maintaining an appropriate level of accuracy in detection. Therefore, improvements in these systems are necessary for both generalization and efficiency for the future in security, reliability, and resilience in UAV and aerial operations against increasingly complex and hostile environments. Such systems would protect the UAV assets and, as a result, enable people to trust them when using them for critical applications such as surveillance, logistics, and disaster management.

5 Conclusions

In this study, we explored the various critical areas of IDS for UAVs to develop an effective IDS that detects and eradicates potential cyber threats by integrating data streams from the cyber and the physical domains. After all, the increased presence of UAVs across sectors of surveillance, military, agriculture, and logistics means that robust protection against intrusions should be guaranteed. These lightweight machine learning models, like the XGBoost classifier, kNN classifier, and random forest, effectively address these specific security and environmental challenges and provide a well-rounded basis for developing reliable, accurate IDS tailored to UAVs. Even the integration of cyber and physical data streams enhanced detection significantly because it offers a more holistic view of UAV operation and attack vectors. Our technique of generating samples using SMOTE also proved helpful, as, in our case, we had five classes, so we divided them into 20 equally. This methodology enhances not only the precision in detecting anomalies but also the robustness of the UAV system against a vast range of threats. Our work integrating these data sources showed improvements in IDS to identify and detect advanced attempts at intrusion that would otherwise have been missed if conventional methods were employing data only from cyberspace.

However, despite the encouraging results of our study, there are also areas we identified that need to be progressed on. Deep learning on UAV security should focus on more advanced training methods and architectures that adapt to the changing scenario of UAV applications and security requirements. Further work is also needed to improve synchronization in physical and cyber data streams where poor alignment may impair the performance of IDS. This is why developing synchronization methods must be essential for better real-time accuracy, especially when it comes to system-based kinds that constantly derive data from various sources. Furthermore, regulatory frameworks determine the future of UAV security, and they play a critical role in

shaping the future of UAV security. As UAV technology expands across various sectors, establishing comprehensive guidelines and standards is essential to maintain a consistent approach to security across industries and regions. Thus, the regulatory framework will also help contribute to a harmonized, secure ecosystem for safe and efficient UAV operations in public, defense services, critical infrastructure, and logistics.

References

1. Sharma, A., Kumar, R., Gupta, S.: Securing UAVs in 5G networks: challenges and opportunities. *IEEE Wirel. Commun.* **29**(3), 12–20 (2022)
2. Tao, W., Ma, Y., Hu, H.: Holistic cyber-physical intrusion detection systems for UAVs. *J. Inf. Secur. Appl.* **58**, 102714 (2021)
3. Tzeng, E., Hoffman, J., Saenko, K.: Adversarial discriminative domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(5), 1603–1616 (2017)
4. Cao, Y., et al.: A study on UAV trajectory tracking and intrusion detection in GNSS spoofing environments. *J. Navig.* **72**(4), 734–748 (2019)
5. Park, J., et al.: GNSS spoofing detection using a lightweight detection method. *Sensors* **20**(9), 2565 (2020)
6. Saha, D., Sinha, D.: Convolutional neural networks for cyber physical security in UAVs. *Neural Comput. Appl.* **32**, 13423–13433 (2020)
7. Zhang, H., et al.: Transfer learning in UAV cybersecurity: a comprehensive survey. *IEEE Trans. Cyber* **50**(9), 3861–3874 (2020)
8. Kumar, A., et al.: Challenges in integrating asynchronous data streams for UAV systems. *Aerosp. Sci. Technol.* **106**, 106036 (2020)
9. Alazab, M., et al.: Cyber-physical security of unmanned aerial vehicles: challenges and solutions. *IEEE Access* **7**, 53318–53335 (2019)
10. Chen, Y., Zhang, X.: A deep learning approach to cybersecurity in UAV systems. *Appl. Sci.* **11**(19), 8707 (2021)
11. Khan, M.M., Ahmed, M.: A comprehensive review of cyber-attacks on unmanned aerial vehicles (UAVs): classification, challenges, and solutions. *J. Inf. Secur. Appl.* **54**, 102537 (2020)
12. Jia, W., et al.: Intrusion detection systems for unmanned aerial vehicles: a comprehensive review. *Comput. Secur.* **123**, 102825 (2022)
13. Chawla, N.V., et al.: SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
14. Ganaie, M.A., et al.: A survey on intrusion detection systems in wireless sensor networks. *Comput. Secur.* **107**, 102347 (2021)
15. Iglewicz, B., Hoaglin, D.C.: *How to Detect and Handle Outliers*. Sage Publications (1993)
16. Zhou, J., et al.: Multi-dimensional feature fusion for UAV intrusion detection. *IEEE Trans. Inf. Forens. Secur.* **16**, 3324–3337 (2021)
17. Stallings, W., Brown, L.: *Computer Security: Principles and Practice*, 3rd edn. Pearson (2012)
18. Sommer, P., Paxson, V.: Outside the closed world: on using machine learning for network intrusion detection. In: 2010 IEEE 7th International Conference on Malicious and Unwanted Software (MALWARE), pp. 1–7 (2010)
19. GitHub: UAVs dataset under normal and cyberattacks. <https://github.com/uamughal/UAVs-Dataset-UnderNormal-and-Cyberattacks.git>
20. Tran, T.A., Nguyen, V.H., Pham, N.D.: Cybersecurity threats to UAV communication systems: detection and mitigation strategies. *IEEE Commun. Mag.* **57**(10), 80–86 (2019)

21. Arthur, M.P.: Detecting signal spoofing and jamming attacks in UAV networks using a lightweight IDS. In: 2019 International Conference on Computer, Information and Telecommunication Systems (CITS), Beijing, China (2019). <https://doi.org/10.1109/CITS.2019.8862148>
22. <https://arxiv.org/pdf/1807.00435>
23. Hu, J., Zhang, Y.: Intrusion Detection Techniques for Industrial Control Systems. Springer (2020)
24. Verma, A., Sharma, N., Singh, P.: Anomaly detection in unmanned aerial vehicles using machine learning algorithms. *J. Aerosp. Inf. Syst.* **17**(3), 153–162 (2020)
25. Satpathy, S., Pradhan, S.K., Ray, B.B.: A digital investigation tool based on data fusion in management of cyber security systems. *Int. J. Inf. Technol. Knowl. Manag.* **2**(2), 561–565 (2010)
26. Mughal, U.A., Hassler, S.C., Ismail, M.: Machine learning based intrusion detection for swarm of unmanned aerial vehicles. In: 2023 IEEE Conference on Communications and Network Security (CNS), Orlando, USA, p. 19 (2023). <https://doi.org/10.1109/CNS59707.2023.10288962>
27. Zhang, X., Wu, Y., Li, F.: AI-driven autonomous defense for UAV systems in cyber-physical environments. *IEEE Trans. Ind. Inform.* **17**(5), 3456–3468 (2021)
28. Yang, G., Wang, Z., Guo, L.: Lightweight cryptography for UAV security in 6G networks. *IEEE Access* **10**, 43054–43068 (2022)
29. Pattnaik, L.M., Swain, P.K., Satpathy, S., Panda, A.N.: Cloud DDoS attack detection model with data fusion & machine learning classifiers. *EAI Endorsed Trans. Scalable Inf. Syst.* **10**(6) (2023)
30. Satpathy, S., Swain, P.K., Mohanty, S.N., Basa, S.S.: Enhancing security: federated learning against man-in-the-middle threats with gradient boosting machines and LSTM. In: 2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–8, July 2024. IEEE (2024)

AI-Powered Video Analytics: Enhancing Real-Time Threat Detection and Public Safety



Kushal Walia, Namita Dandawate, and Bhumik Thakkar

Abstract AI redefines video analytics for public safety, enabling fast threat detection and response. Traditional systems that rely on human oversight struggle with timely and accurate threat ID. AI-driven video analytics can detect objects, track movements, recognize unusual activity precisely, and provide advanced threat assessment and situational awareness in public spaces. This paper will explore the core capabilities of AI in video analytics: object recognition, anomaly detection, and multi-camera integration, which provide a complete view of high-risk areas. Recent advancements like edge computing enable real-time processing, making AI-driven insights available to security and emergency teams with little delay. These capabilities are particularly relevant for public safety applications like crowd management and incident response at significant events. Ethical considerations like data privacy, algorithmic fairness, and cybersecurity will also be discussed, as well as real-world examples of AI video analytics in major US cities. The paper will conclude with future directions: autonomous aerial monitoring and predictive analytics for proactive threat detection. This paper will argue for ethical, research-driven placement of video analytics to facilitate safer, more sustainable public spaces in the US.

Keywords Artificial intelligence · Edge computing · Public safety · Threat detection · Video analytics

1 Introduction

In the field of public safety, video analytics' expansion through AI constitutes a technological shift that transforms the legacy monitoring systems burdened by constant human intervention and slow reaction times. Incorporating AI in video analytics has

K. Walia (✉) · N. Dandawate
Amazon, Seattle, USA

B. Thakkar
Apple, Cupertino, USA

been at the forefront as an improvement aspect, offering more accurate results and the ability to process and assess enormous data streams in real-time. This technology is at the heart of effective threat detection and response, particularly in dense or high-security settings.

The classic model for video analytics has been fixed cameras, and human limitations and sheer volumes of data limit human operators, whose capacity to perceive and react to anomalies. These legacy systems are usually plagued by latency in threat detection and unnecessary false alarms, which can be counterproductive to public safety programs and consume resources. With the development of AI technology such as deep learning, CNNs, and advanced algorithms, video analytics as an online solution has transformed video analytics to address these challenges.

This paper focuses on exploring how AI can be used to advance video analytics in public safety and the placement of this technology in the public spaces of airports, shopping malls, and cities. Through an analysis of the fundamental capabilities of AI-based video analytics—object detection, anomaly detection, and facial recognition—the study aims to put into focus the way AI is not only improving the efficacy and efficiency of threat detection but is also a critical factor in driving national security and public welfare to a higher path.

Along the way, the article will explain supporting real-time technologies for processing and analysis, discuss the utility of such technology in public safety, and discuss the ethical context of deploying advanced monitoring technologies. By an analysis deeply grounded in a survey of present implementations and future possibilities, the research seeks to underscore the extraordinary role of AI in public safety activities and on societal security's implications.

2 AI and Video Analytics: Overview

AI video analytics is a paradigm shift in visual data use for public safety. AI enables automated understanding of video, extending considerably the potential for detection, analysis, and reaction to possible danger in real-time. The following chapter outlines the critical AI technologies—deep learning, CNNs, and computer vision—on which current video analytics systems depend.

Deep Learning and Video Analytics

Deep learning is a machine learning type that uses stacked neuron networks to process various data, including video. It is particularly suited to process the constant flow of images from CCTV cameras, allowing it to identify patterns and features of interest and concern to public safety. For example, computer systems can identify regular pedestrian streams and potential security threats, i.e., the efforts of unauthorized individuals to get in or suspicious masses [1].

Convolutional Neural Networks (CNNs)

CNNs are deep learning architectures specific to pixel data that recognize the spatial hierarchies of images. Consequently, they can be employed best in video analysis [2]. Public safety scenarios, where CNNs are used in learning and training on the below skills - detection and tracking of objects, abnormality detection, even face recognition and detection for efficient threat handling before events - are attainable with the help of CNNs [3].

Computer Vision Techniques

Computer vision enables machines to perceive and understand visual data from the physical world, simulating human vision at a scale and speed that humans cannot. AI-based computer vision systems analyze video streams to identify and track objects, map their motion, and locate anomalies in expected patterns. These capabilities are critical in maintaining security in public places because they enable real-time response to emergent and real threats [4].

Key Capabilities of AI in Video Analytics (as shown in Fig. 1)

Object Detection and Tracking: AI tools can continuously watch video streams to detect and track objects or individuals and mark them as dangerous when they are so. This is an essential feature in locations like airports and malls, where traffic is high, and there is a need for constant monitoring [5].

Anomaly Detection: AI systems can identify abnormal behavior quickly by learning to recognize what constitutes normal behavior in provided settings. Rapid

Year	Architecture	Top-5 Error	# Params (M)
1998	LeNet (MNIST focus)	<i>N/A for ImageNet</i>	~0.06 (60k)
2012	AlexNet	16.4%	~60
2013	ZF Net	14.8%	~62
2014	VGG-16	7.3%	~138
2014	GoogleNet (Inception v1)	6.7%	~6.8
2015	ResNet-50	5–6%	~25.6
2015	ResNet-152	3.6%	~60
2017	SENet (Squeeze-and-Excitation)	~2.25%	~115
2019	EfficientNet-B7	~1.7%–2.0%	~66

Fig. 1 Below is a simplified table showing major CNN milestones on ImageNet, along with approximate top-5 error rates and parameter counts (in millions) [8–12]

detection is essential to stop escalation in public areas and can render security officials much more efficient [6].

Facial Recognition: Problematic, but facial recognition is a potent possible public safety utility when used responsibly. Facial recognition technology can be coupled with AI to detect individuals banned from an area or wanted by law enforcement, supporting police work while maintaining public safety [7].

AI technologies have made a tremendous contribution to streamlining the operation of video analytics systems for public safety. They provide an assured system for automating analytics over vast spaces, responding quickly to emerging threats, and ensuring a safe environment.

3 Public Safety and Security Applications

AI video analytics solutions are of excellent security and public safety importance. The following outlines several primary applications in which AI-powered video analytics are at the forefront of enabling situational awareness, threat detection, and incident response in public spaces.

Real-Time Threat Detection in Public Spaces

AI video analysis platforms play a critical function in identifying potential security risks in real-time across many public places such as airports, train stations, and stadiums. AI can identify anomalies in regular video stream content like suspicious packages, unapproved access, or unusual patterns of motion indicative of the emergence of an upcoming security threat. Research supports that AI-powered platforms can significantly reduce response time, preventing crimes from occurring [13].

Crime Prevention and Response

Police authorities now use AI-driven video analytics to improve crime investigation and prevention. For instance, video analytics can enhance the efficiency of police authorities in recognizing and following up on suspects or cars within a city, leading to quicker criminal case closure. Moreover, analyzing past data helps in pattern detection and predictive policing and can help prevent crime by maximizing resource utilization [14].

Crowd Management and Public Event Security

Public events such as concerts, festivals, or political rallies are specialized security threats. AI-powered video analytics aid crowd management in monitoring the crowd density and movement patterns, enabling the security personnel to stay crowding-free and react immediately whenever an event occurs. It can also identify agitation or intent violence among a crowd and the quick deployment of the emergency services to the exact spot where they are needed [15].

Smart City Integration

Smart cities utilize AI video analytics to enhance urban safety and efficiency. Traffic management, for instance, benefits from AI's capability to optimize signal timings based on real-time traffic flow analysis captured through video. Similarly, AI systems support urban planning decisions by providing insights into pedestrian and vehicular movement patterns, thus contributing to safer and more responsive city environments [16].

Emergency Response Optimization

In emergencies, such as natural disasters or accidents, AI-enabled video analytics may be instrumental in analyzing situations at short notice and driving first responders with maximum efficiency. With live streaming of videos, AI can delimit the extent and scale of incidents, chart evacuation routes, and optimize the deployment of resources. This facility improves the efficacy of emergency responses and minimizes the risk of loss of human life and assets [17].

Public Health Analytics

Due to global health threats in the form of pandemics, AI video analytics have also been applied in public health monitoring. They can impose compliance with health standards like social distancing and mask-wearing. AI technology has been utilized in crowd temperature quantification, monitoring public sneezing and coughing, and providing information to public health officials [18].

AI-powered video analytics supports public safety efforts across many sectors by optimizing video stream monitoring, expediting emergency response, and reinforcing law enforcement and public health initiatives. With the technologies evolving, they are leading the measures cities and governments follow to guarantee and amplify public security and safety.

4 Technical Progress in AI-Powered Video Analytics

The technical environment of AI-powered video analytics is one of rapid advancement that continues to improve public safety features at ever-increasing rates, as shown in Fig. 2. The primary technical advancements that further enhanced the accuracy, speed, and efficiency of AI software to process video content for public safety applications are discussed in this section.

Edge AI and Real-Time Processing

One of the most exciting emerging trends in AI-based video analytics is the inclusion of edge computing. Edge AI computes on hardware near where the data is produced instead of cloud-based services. This proximity dramatically reduces latency, allowing near-instant analysis and response to video information, a significant consideration in emergencies where seconds may count [19]. Real-time

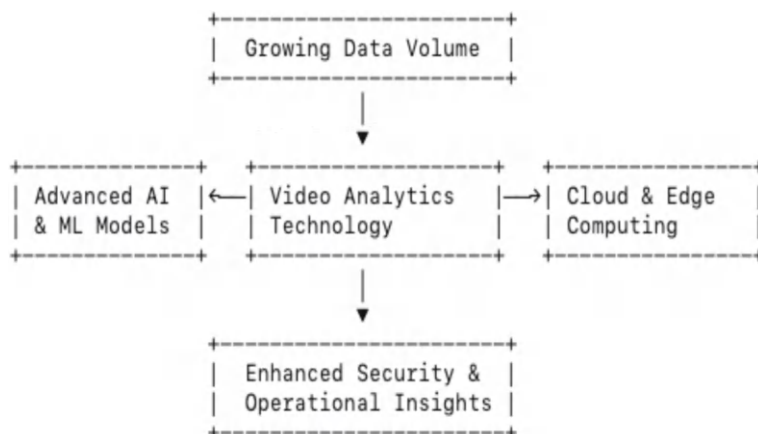


Fig. 2 Diagram highlighting key drivers behind the growth of video analytics technology [19]

processing also enables AI systems to identify and act on threats or emergencies as they become imminent in real-time, potentially enhancing response efficiencies and saving lives.

Multi-camera Video Fusion

Artificial intelligence advancements have also made a fusion of multiple video streams from several cameras possible into a unified analysis system. Multi-camera fusion applies advanced algorithms to fuse heterogeneous video inputs to present a merged perspective of expansive spaces or complicated environments. This is crucial for public safety, enabling continuous analytics without blind spots so that every section is observed and significant incidents are not overlooked [20].

AI's Role in Reducing False Positives

AI technologies have also significantly improved their ability to distinguish between real threats and non-threatening anomalies, which has been an issue previously. Using intricate machine learning techniques and deep neural networks, AI systems are trained on large datasets to recognize minute patterns in video streams that might pose a threat. This reduces false positives significantly, allowing public safety officials to enhance resource allocation and respond only to real threats [21].

Enhanced Object Detection and Tracking

With the advent of deep learning techniques, object detection and tracking have been transformed. Sophisticated AI algorithms such as You Only Look Once (YOLO) and Single Shot Detector (SSD) enable real-time detection by processing video frames at a very high rate to detect and track multiple objects. These algorithms are best suited to handle dynamic and crowded public scenes where detecting and tracking objects or individuals in real-time is of utmost significance to safety and security [22].

Behavioral Analysis and Anomaly Detection

Behavioral analysis is another use case where video analytics with AI surpasses others in using pattern detection to monitor usual behavior and warn about unusual behavior. The anomaly detection algorithms train and become wiser with experience, responding better to new inputs and, more precisely, recognizing deviations from normal. This AI capability applies when identifying likely risky or disruptive actions before the actual occurrences [23].

Facial Recognition Technologies

While facial recognition technology is still an ethical concern, its technical development is of immense help to public safety by identifying persons who could pose a security risk. Technical development has enhanced facial recognition algorithms to become more precise even under challenging situations such as poor lighting or exposure. This technology enables law enforcement activities by rapidly identifying subjects of interest with no real-time human intervention [24].

The continuous innovation of AI-based video analytics solutions is emerging as a key enabler of public safety. With their capability to lower the response time, lower false positives, and provide advanced monitoring capabilities, the solutions are integral to modern public safety efforts. As the systems keep improving, they can continue revolutionizing the effectiveness and efficiency of public safety operations globally.

5 Challenges and Ethical Issues

While AI-driven video analytics significantly enhance public safety, their usage is not problem-free and involves ethical issues. The key challenges to using AI for public safety solutions are outlined below, data privacy, algorithmic bias, and cybersecurity being the key ones.

Data Privacy and Civil Liberties

The use of AI-driven video analytics concerns private privacy and civil rights. Continuous monitoring of public spaces can lead to inadvertent recording of personal information without explicit consent, thus intruding on privacy to a great extent. Laws such as the General Data Protection Regulation (GDPR) in the European continent provide a framework for data privacy. Still, detailed counterparts are being drafted elsewhere, for example, in the United States [25]. Making AI systems legally compliant and ensuring proportionate enhancement of public safety is a constant challenge for developers and policymakers.

Bias in AI Systems

Algorithmic bias represents one of the most pressing issues with AI technologies, where bias in training data leads to discriminatory results. Face recognition technologies, for example, have been known to contain racial and gender bias, leading to over-targeting certain groups [26]. Preventing these biases is a matter of data diversity when training AI models and performing thorough test phases to identify and stop discriminatory patterns before roll-out.

Security of AI Systems

As AI systems become the mainstay of public safety, so do they become high-priority targets for cyber-attacks. These systems must be guarded against potential intrusions since vulnerabilities can be exploited to control video streams, mislead operational action, or steal confidential data [27]. It is crucial to ensure these technologies are shielded from evolving threats by continuous breakthroughs in cybersecurity products.

Consent and Transparency

Public acceptance of AI-based video analytics is contingent upon open operations and communication of how they are used. The public should know how their data is processed and what value is being added to public safety. Engaging with community leaders and stakeholders in the development and deployment process can bring about trust and ensure the technology is responsibly and ethically utilized [28].

Legislative and Regulatory Issues

Proceeding with the set of laws and regulations surrounding public use of AI is another substantive issue. Legal compliance varies significantly between locations, and tracking progress in legislation is essential to maintain ongoing compliance and ethical codes in adopting AI technologies [29].

Psychological Impact and Public Perception

The sense of perpetual monitoring by AI systems will likely instill a sense of tracking among the general public and will potentially induce psychological distress or alter their behavior. Efforts must be made so that a detrimental cost in the form of public morale and conduct does not accompany the benefits of enhanced security. Public engagement and publicity initiatives concerning the uses and limitations of AI systems will probably reduce such apprehensions [18].

Although AI video analytics has immense promise for public safety, addressing these technical and ethical challenges is critical to their responsible and practical application. As technology continues to advance, continuing dialogue among technologists, policymakers, and the public will be required to control the ethical landscape and make sure that AI yields a maximum positive effect on society.

6 Case Studies

This part identifies some real case studies showing the success of AI video analytics in public safety across various scenarios. These cases indicate the practical benefits and shortcomings of deploying AI technologies for public safety.

Case Study 1: New York City's Real-Time Crime Center

AI-powered video analytics are utilized in New York City's Real-Time Crime Center to patrol public spaces and react more quickly to incidents. The center integrates video feeds throughout the city with AI algorithms to detect suspicious activity and support crime prevention. A success story involved using AI to quickly locate a missing child in a large festival, testing the system for speed and accuracy [22].

Case Study 2: London Public Transport System

London has adopted AI video analytics within its public transport to enhance the safety of travelers and crowd management, especially during peak hours and significant events. AI technology processes video for crowd-intensity observation and detection of disruption or emergencies to make interventions on time. Besides promoting safety, the strategy has facilitated smoother passenger flow with less congestion and enhanced commuter satisfaction [19].

Case Study 3: Singapore's Smart Nation Initiative

Singapore's Smart Nation initiative utilizes large-scale AI-based video analytics to optimize city functions and public safety. The initiative uses AI to learn traffic patterns and modify signal timing, significantly reducing traffic congestion and road accidents. Further, AI monitoring devices monitor environmental health and public health compliance, a broad scope of AI applications in maintaining public safety and city operations [30].

Case Study 4: Disaster Response Strategy in Tokyo

Following the 2011 earthquake and tsunami, Tokyo used AI-driven video analytics to enhance disaster response. AI platforms now analyze live video from different sources to assess damage, identify exposed areas, and efficiently coordinate emergency responses. Preemptive evacuation planning and resource deployment in future emergencies have been facilitated by this approach, setting the role of AI in disaster management [29].

Case Study 5: Operation Virtual Shield in Chicago

Chicago's Operation Virtual Shield mobilizes thousands of cities and private cameras, which are analyzed with AI to support citywide security. The AI system helps identify possible crimes throughout the city and sends live alerts to police agencies, minimizing response times and enabling preventive policing. This monolithic network has assisted in numerous high-profile crimes and emergency responses, showing the scalability and effectiveness of AI video analytics in an urban setting [28].

7 Conclusion

AI video analytics are revolutionizing public safety with unprecedented real-time threat detection, predictive analytics, and situational awareness. AI is rendering public safety systems more capable of detecting anomalies, recognizing objects, and monitoring individuals more accurately and rapidly through deep learning, computer vision, edge computing, and IoT integration. This technology can reduce response time significantly, optimize resource use, and, ultimately, make public spaces safer.

As such advancements have been achieved, using AI for video analytics poses inherent operational and ethical issues. Privacy rights, algorithmic bias, and AI system security are recurring problems that technology makers and policymakers must consider. A balanced approach with moral principles, transparency, and public consultations will be required to ensure that the technology serves the public interest without undermining civil liberties.

In the coming times, innovation in AI-driven video analytics will be about enhancing predictive capabilities, enriching behavioral analysis, and integrating multi-sensor data from IoT sensors. It will be essential to ensure responsible innovation and deployment of the technologies to meet an end-to-end public safety ecosystem that is responsive, ethical, and resilient. As cities and organizations implement such technologies, further research and cooperation among stakeholders will be required to exploit the full benefit of AI-powered video analytics and safeguard the rights and trust of the public.

The future of video analytics using AI is bright and full of promise, with the possibilities for revolutionary innovations in public safety limitless. The future directions and trends discussed below have the potential to enhance the strength and effectiveness of AI for public safety activities.

Autonomous Monitoring Drones

One of the most likely video analytics AI breakthroughs is the emergence of autonomous drones with advanced video capabilities. Unmanned aerial vehicles can complement the capabilities of video analytics solutions and provide aerial intelligence that cannot be accessed through fixed cameras. With AI-enabled real-time data processing, drones can give real-time insight into where decisions must be made promptly, for example, in search and rescue missions, crowds, or disaster response scenarios. Future research may focus on creating autonomous navigation of drones in crowded environments [23].

Predictive Security Systems

Using AI not only to detect but also to foretell potential security risks is a significant step ahead for public security. Predictive analysis uses past data and machine learning algorithms to forecast potential events, allowing law enforcers and security organizations to allocate resources more effectively and stop crimes before they occur. Enhancements in data collection and algorithmic accuracy are essential to the effectiveness of such systems [24].

Integration with IoT for End-to-End Security Networks

Combining AI video analytics and other Internet of Things (IoT) devices provides a holistic solution to public safety. For instance, combining video streams and sensors that pick up gunshots, breaking glass, or other signs of criminality can produce a multi-layered security system. Future developments will look into how these combined systems can best be optimized to support urban security infrastructures effectively [25].

Advanced Object Recognition and Behavior Analysis

Object detection and behavioral analysis are key to the future of video analytics based on AI. These are backed by deep learning for detecting finer aspects of video content, such as facial expressions, body language, and even gait patterns. Further developing these features could lead to better detection of suspicious activities or individuals in crowded public spaces, improving response time and efficiency [26].

Disclaimer

This research paper is published in our capacity and does not represent our employer or affiliated institutions' views, opinions, or policies. The authors declare that no external funding exists for this work, and all opinions expressed herein are strictly our own. In no event shall the authors, their employers, or any affiliated institutions be liable for any damages, including direct, indirect, special, incidental, or consequential damages arising from the use of the information contained in this paper. The content of this paper is provided for informational and educational purposes only.

References

1. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
2. Walia, K.: Accelerating AI and machine learning in the cloud: the role of semiconductor technologies. *ESP Int. J. Adv. Comput. Technol.* **2**(2), 34–41 (2024). <https://doi.org/10.56472/25838628/IJACT-V2I2P105>
3. Ren, X., Zhang, Y., Li, H., Wang, J.: CNN-based multi-object tracking networks with position correction and IMM. *Scipedia* (2023). https://www.scipedia.com/public/Ren_et_al_2023a
4. DHL: AI-driven computer vision: a new lens into logistics. DHL (2023). <https://www.dhl.com/global-en/delivered/innovation/ai-driven-computer-vision-and-image-recognition.html>
5. Analytical AI: S&T awards funds to a startup developing object detection & tracking algorithms for securing public areas, 15 June 2023. Department of Homeland Security (2023)
6. Ardabili, B.R., Pazho, A.D., Noghre, G.A., Neff, C., Bhaskararayani, S.D., Ravindran, A.K., Reid, S., Tabkhi, H.: Understanding policy and technical aspects of AI-enabled intelligent video surveillance to address public safety (2023). [arXiv:2302.04310](https://arxiv.org/abs/2302.04310)
7. Veritone: Police reform: enhancing operations with AI technology, 15 May 2023 (2023)
8. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the 36th International Conference on Machine Learning, PMLR, vol. 97, pp. 6105–6114 (2019)

9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**(6), 84–90 (2012)
10. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *Computer Vision – ECCV 2014*, pp. 818–833. Springer International Publishing (2014)
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, Conference Track Proceedings, 7–9 May 2015* (2015)
12. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9 (2015)
13. Abdallah, R., Harb, H., Taher, Y., Benbernou, S., Haque, R.: CRIMEO: Criminal behavioral patterns mining and extraction from video content. In: *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–8 (2023)
14. Angus, A., Duan, Z., Zussman, G., Kostić, Z.: Real-time video anonymization in smart city intersections. In: *2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, pp. 514–522 (2022)
15. Badidi, E., Moumane, K., El Ghazi, F.: Opportunities, applications, and challenges of edge-AI enabled video analytics in smart cities: a systematic review. *IEEE Access* **11**, 80543–80572 (2023)
16. Catlett, C., Cesario, E., Talia, D., Vinci, A.: Spatio-temporal crime predictions in smart cities: a data-driven approach and experiments. *Pervasive Mob. Comput.* **53**, 62–74 (2019)
17. Chen, Y., Xie, Y., Hu, Y., Liu, Y., Shou, G.: Design and implementation of video analytics system based on edge computing. In: *2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pp. 130–1307 (2018)
18. Cob-Parro, A.C., Losada-Gutiérrez, C., Marrón Romera, M., Gardel Vicente, A., Muñoz, I.B.: Intelligent video surveillance system based on edge computing. *Sensors (Basel, Switzerland)* **21**(9) (2021)
19. Wang, J., Feng, Z., Chen, Z., George, S., Bala, M., Pillai, P., Yang, S.W., Satyanarayanan, M.: Edge-based live video analytics for drones. *IEEE Internet Comput.* **23**(3), 27–34 (2019)
20. Du, W., Li, A., Zhou, P., Niu, B., Wu, D.: PrivacyEye: a privacy-preserving and computationally efficient deep learning-based mobile video analytics system. *IEEE Trans. Mob. Comput.* **21**(9), 3263–3279 (2022)
21. Elhamod, M., Levine, M.: Automated real-time detection of potentially suspicious behavior in public transport areas. *IEEE Trans. Intell. Transp. Syst.* **14**(2), 688–699 (2013)
22. Ghasemi, M., Kleisarchaki, S., Calmant, T., Gürgen, L., Ghaderi, J., Zussman, G.: Real-time camera analytics for enhancing traffic intersection safety. In: *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, pp. 1–7 (2022)
23. Lachner, C., Rausch, T., Dustdar, S.: A privacy-preserving system for AI-assisted video analytics. In: *2021 IEEE 5th International Conference on Fog and Edge Computing (ICFEC)*, pp. 74–78 (2021)
24. Neises, J., Besse, A., Rouquier, J.-B.: Privacy-preserving CCTV analytics for cyber-physical threat intelligence. In: *Cyber-Physical Security for Critical Infrastructures Protection*, vol. 12618, pp. 3–15. Springer (2021)
25. Rabieh, K., Mercan, S., Akkaya, K., Baboolal, V., Aygün, R.S.: Privacy-preserving and efficient sharing of drone videos in public safety scenarios using proxy re-encryption. In: *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 45–52 (2020)
26. Simpson, T.: Real-time drone surveillance system for violent crowd behavior unmanned aircraft system (UAS) – human autonomy teaming (HAT). In: *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, pp. 1–9 (2021)
27. Tsakanikas, V.D., Dagiuklas, T.: Enabling real-time AI edge video analytics. In: *ICC 2021 - IEEE International Conference on Communications*, pp. 1–6 (2021)

28. Wachter, S., Mittelstadt, B.: A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Columbia Bus. Law Rev.* **2019**(2), 494–620 (2018)
29. Walia, K.: Scalable AI models through cloud infrastructure. *ESP Int. J. Adv. Comput. Technol.* **2**(2), 1–7 (2024). <https://doi.org/10.56472/25838628/IJACT-V2I2P101>
30. Zhang, Q., Sun, H., Wu, X., Zhong, H.: Edge video analytics for public safety: a review. *Proc. IEEE* **107**(8), 1675–1696 (2019)

Enhancing Women's Safety and Security Through IoT Systems



Jyoti Yogesh Deshmukh , Mayura Vishal Shelke , Anuja Jadhav ,
and Saleha Saudagar

Abstract In contemporary society and culture, women's safety is a primary concern. The alarming escalation in crimes targeting women has underscored the imperative need for the security and enhancement of a robust security mechanism for women. This paper presents a thorough study of the panic buttons being used to ensure female safety. This research aims to provide a technically proficient solution in terms of GPS tracking systems, GSM, and Arduino-based instruments to address the issue of female protectiveness. The study examines several female personal care alternatives and designs a panic button system. The research is carried out to design and develop a trepidation toggle alarm system dedicated to safeguarding women. The proposed study shows that adding a panic button system is a breakthrough in improving women's safety and security.

Keywords Security · Panic button · GPS tracking · GSM · Arduino-based instruments

1 Introduction

Today's women are superwomen who perform various duties simultaneously to maintain tradition and be competent in the modern era. She frequently bridges personal, social, and professional lives to keep her peaceful. The reality is that an enormous increase in malfeasance incidents with working females for a few decades is a big concern. There is a need to develop a secure system that gives her the feeling of security. A trustworthy and effective security system is required to guarantee

J. Y. Deshmukh (✉)

Artificial Intelligence and Data Science, Marathwada Mitramandal's Institute of Technology,
Lohgaon, Pune, Maharashtra, India
e-mail: [jyoti1584@gmail.com](mailto: jyoti1584@gmail.com)

M. V. Shelke · A. Jadhav · S. Saudagar

CSE Department, School of Computing, MIT Art, Design and Technology University, Pune,
Maharashtra, India

women's safety. One solution that can be used to provide immediate aid to women in emergencies is panic button-based security systems.

The following benefits come with the panic button-based women's security system:

- Quickly offers assistance in an emergency; women may find it simple to access, Can be put in different places, and Gives ladies a feeling of security.
- The panic button can be worn as a bracelet or pendant, making it simple to reach in an emergency. The security system can be set up as a mobile phone application or a separate unit.
- The system is built to identify the signal from the panic button and react rapidly to the emergency call. In an emergency, a panic button can be pressed to seek assistance. Typically, the button is linked to a security system with immediate emergency response capabilities.

There are several places where panic buttons can be placed, including offices, homes, and public spaces. The security system receives a signal when the button is pressed and alerts the authorities.

2 Literature Survey

The paper [1] proposed a Female security system using IoT and open-source Emerging Technology. A strong foundation for women's safety and awareness has developed in this publication. The emerging technology assures that women are safeguarded by keeping an eye on and safeguarding. The IoT and open source technology combined with the intelligent women protection system contribute favorably. To protect women, the ESP8266 and Raspberry Pi 3 Wi-Fi gadgets use modules like GPS, fingerprint, camera, GSM, body sensors, and Nerve simulator upgrades. The woman can track and locate with the help of wearable wearers, GPS, camera, GSM, and fingerprint modules in assistance with body sensors. The nerve simulator applies safeguarding techniques. At the time of the sudden controversial situation, the female user of the wearable gadgets will be verifying her thumbprints, after which the GSM and GPS equipped with a microphone and speaker, camera, and body sensors are activated, location alerts are sent to the appropriate people, and statistics are updated in the server. In addition, with the help of a nerve stimulator, various shocks can be delivered to the assailant. The structure that has been created makes sure that it successfully protects women in all situations.

The paper [2] proposed applications of IoT for industries and to enhance the quality of life.

The paper [3] proposed a Women's Security Safety System using Artificial Intelligence. The existing criminal dataset checks to see if the woman's present location corresponds with its dataset when she enters any unknown place while signed in. If it does, it alerts her that the area is a crime-prone region. She will have options after she becomes accustomed to crime-ridden neighborhoods. Either she would stay away

from it or need to be ready with safety precautions; in this situation, SWMS may benefit her. She is given access to the assistance button in the app if she travels to a crime-prone area and senses any risk from strangers. A message will be produced after hitting the help button. The produced message will include the phrase "I am in trouble. Please help" and the sender's current GPS location. The emergency contacts listed will get the message that was produced. The SWMS application has utilized the following technologies: (1) Front-end HTML development. (2) Client and server-side Java. (3) MySQL for processing and the back end. (4) Server-side processing with Visual Studio. (5) Android SDK to create an application that is compatible with Android.

The paper [4] proposed Designing and implementing a mobile application for Female safety. An initial process for this application registration and emergency contacts are required. Dynamic GPS tracking of the Google Maps API's feature is activated to show the owner's location on a map as they travel between locations. In an emergency, the sufferer can long-press the lower volume button or shake their phone to a specific frequency. Following this action, a call is sent to the Master contact, and observant information with the sufferer's name or identity, GPS position, and a guideline message is delivered via SMS to all registered or authentic Emergency Contacts. If the person serving as the emergency contact has the same app, they may be able to access the position directly using the dynamic GPS tracking system; otherwise, they may use the message link. The person might head straight towards the scene and assist the victim. Let's say that while the victim is moving, the live position will be updated every certain number of seconds. An alert message with the most recent location is delivered to the emergency contacts at all times, not just in emergencies.

The paper [5] proposed a female security system with the help of IoT, GSM, and GPS. The work uses GPS, GSM-SMS, and IOT warnings to show an automotive localization system for lady security. The device allows for the localization of the lady and the transmission of her whereabouts to the rescue crew over the Internet and short message service (SMS). The technology may be connected to a car's alarm system to warn passersby to aid the woman. Among the components of the proposed safety tracking system are a GPS signal receiver, a GSM modem, a microcontroller, and an IOT module. When a lady needs assistance, she will push the security alert switch, and the GPS receiver will then receive position data, which will have the longitude and latitude format together with an alert through IOT from satellites. The microcontroller analyzes the available information, and after processing, it is delivered via a GSM modem to the specific individual. For women in danger, the provided application offers an affordable answer. The suggested technique can be applied in other application types, such as child and woman security, when the required information is only sometimes and irregularly asked (when requested).

The paper [6] proposed a women's security system linker with a LoRa Transceiver Using GSM & GPS. The proposed system, which employs long-range wireless LoRa Technology paired with a GSM modem, is activated by switching on the emergency switch if a woman is harassed or believes she is in danger. Following activation, the microcontroller obtained the position information of the crime scene from the GPS

modem and wirelessly sent it to the LoRa transceiver modem. As a safety precaution, the GSM modem then broadcasts her whereabouts by SMS message to the authorized personnel number. As a result, the suggested system uses LoRa technology, which is network-independent and will be highly helpful in saving lives and stopping violence against women.

The paper [7] proposed an Arduino Women Safety Security System. The Arduino-based Female Safety System is primarily applied to inform the concerned party and the police of women's present position, particularly those in an emergency or crisis. The primary gadget has a GPS and GSM module. While GPS tracks the current location, the GSM model transmits the message to the numbers stored in the system. After switching the system on, it refers to GPS to follow the user's location and sends notifications to others who can assist her. Every two minutes, this security system may send a message to the designated person's contact, including the current position. The message can be monitored in real time using this app. A microcontroller serves as the project's primary controller. When the magnetic switch is activated, the microcontroller reads it. It utilizes the information to transmit the woman's location, the format of which will have longitude and latitude values, to the predefined mobile phone through Wi-Fi and GSM. It also turns on the alarm buzzer. The project's status will be shown on the LCD module.

The paper [8] proposed an E-wearable Smart Lady Safety System for females working remotely. They could alter the jewellery they wore with the "Borla" design, ensuring the woman's protection. Because of the wearable security system in this ornament, women working in the fields won't be alarmed by emergencies, and their whereabouts may be tracked. They will be protected and given a great deal of freedom thanks to technology, which includes the HD camera with fast capture and speech recording and a live location with a constant server update. Additionally, it employs sensors for measuring ambient alcohol, such as environmental alcohol sensors. An alarm message from the GSM system will be created and sent to the mobile phone for certain circumstances. This gadget is dependable and straightforward to use.

The paper [9] offered a Detailed Study of Women's Safety Using Machine Learning and Android App Development. The article summarises the numerous safety care that are accessible to women. It is proposed that an Arduino-based novel perspective on women's security warning systems be employed to send SMS notifications to victims' near ones and family, encouraging women to continue their everyday activities without fear. Additionally, our framework includes an Arduino robber warning that detects and notifies the authorized individual of any unauthorized access. Thus, the recommended system's affordability, dependability, and user-friendliness assist women in overcoming their fear in challenging circumstances. Three machine learning models have also been tried, and the SVM classifier performs the best with our dataset, with an accuracy of 89.5%. Additionally, the app may get every route that might be taken from the user's starting point to the desired destination. If the cab driver attempts to take one of the other routes and veers off course, a warning is issued before an SOS message is sent. Additionally, the victim's safe zones may be marked, and MMS can be sent during crises. The advantage of this program is that it has an auto mode for use in emergencies. Additionally, the police

and the person's emergency contacts are informed of the person's current position. Additionally, since a cloud database is being used, there will never be a problem with too many users. The system may be improved when new technology is developed, making it even more reliable and user-adaptive. Additionally, it may be created on the IOS platform. Therefore, this app can benefit society in times of need.

The paper [10] suggested using different devices associated with the Internet for Female Safety with an alarm system. IoT connects billions of devices and transmits important information, which is essential for women's safety. The article summarizes the several safety actions that are available to women. It is recommended to use an Arduino-based Female security warning system that may send SMS notifications to the victims' near ones or family, encouraging women to continue their daily activities without fear. Additionally, our framework includes an Arduino robber warning that identifies and notifies the authorized individual of any unauthorized access. Thus, the recommended system's affordability, dependability, and user-friendliness assist women in overcoming their fear in challenging circumstances.

The paper [11] proposed a women's security safety system using AI. Two parts of the proposed application system are GMS and GPS. GMS has produced basic touch motions like double tapping or touching your screen and has built-in and simple touch technologies. After tapping the button, the application's GPS transmits the information with the user's present location and a request for help to the user's pre-collected contacts. The program's automated voice recorder may also record ambient sounds for use in court if the cops are called because the caller didn't understand the message. A built-in AI system is also included in the gadget, which warns the victim using data from the police open database of their whereabouts and previous criminal behavior.

The paper [12] presented a smart wearable device that uses Raspberry Pi-based with GPS and GSM technology for women's safety. The research involves an innovative care solution known as the wearable device innovative system created on the Raspberry Pi3 to improve the safety and security of women/children. It serves as both an alarm and a security and safety system. It sends a buzzer alert to others close to the user (who is wearing the smart device). The system uses Global Positioning System (GPS) technology to find the user and broadcasts the user's location through SMS to the emergency contact and police. When the user presses the panic button, the device also uses the USB Web Camera connected to it to snap a picture of the incident and the user's or victim's surroundings. It sends it to the emergency contact as an email alert.

The paper [13] proposed Women's Safety Devices Using Panic Buttons. This project is designed with ATmega48. This project shows a GSM modem-based system and GPS for detecting women's safety. The system can inform the neighbors by connecting to the alarm system. A GPS signal receiver, an Atmega48, and a GSM Modem comprise the identification and information system. The GPS receiver receives latitude and longitude data from satellites in the form of position data. The ATmega48 processes this information, and the processed information is delivered to the user through the GSM interface. The MCU is interfaced with a GSM modem. The predetermined mobile number receives an SMS from the GSM modem. A woman

can use the panic button designated for her when she is in danger and needs to defend herself. When the panic button is pressed, the entire system is busy, and then an SMS is sent to alert the individual of their location using GSM and GPS.

The paper [14] suggested designing and constructing a panic button alarm system for safety problems. The paper describes the design and development of a panic button alarm system for security crises, which is used for real-time monitoring of security emergencies such as theft and threats to life and property. The primary purpose of this project is to provide real-time monitoring of various security emergency scenarios and to locate persons in need using a GPS mapping system and Google Maps. This project uses the Arduino Uno microcontroller, which acts as the system's brain and is where all instructions are carried out. A module with Wi-Fi connectivity provides access to a microcontroller and the security control center, and a GPS module offers the location of the push button when someone in a dangerous situation activates it. This design can be employed in remote places with limited security access, saving time when calling security during security crises.

The paper [15] used the IoT and Algorithm to present a Smart protection Solution for Women. The invention primarily focuses on a wearable, IoT-based self-security system that enables users to communicate their position when they experience panic and locate the closest safe spot. The system's interface is simple, and only one person may access it. The system is managed by a Raspberry Pi computer with two operating modes: standard mode and security mode. In security mode, the fingerprint sensor acts as a panic button. When a fingerprint is detected, the system communicates with its location, takes a photograph of the suspect, and stores it in the cloud. Users can register their fingerprints in regular mode. The machine learning system uses the victim's location as an input to forecast the position of the closest safe spot. The kNN algorithm is utilized to estimate the closest safe spot more accurately based on the user's location. Nearest Neighbour is referred to as kNN. The current position and an SOS message are sent to the registered number over GSM. The Arduino code contains both the SOS message and cell phone number. Additionally, it may be used to broadcast the position of the closest safe area to the user's mobile device, enabling them to get there as quickly as possible. GPS is used to find a specific location, providing the user's longitude and latitude to the Arduino Nano.

The paper [16] proposed an Innovative Approach to Women's Security Using a Smart Device. The electronic system for women described in this project is safe and secure and includes an Arduino controller and sensors for temperature, heart rate, and sound. This research uses a buzzer, LCD, GSM, and GPS. It is possible to attach a wire to the victim's body. Thus, when a woman is in danger, the device monitors changes in body temperature and vital signs via an audio sensor, such as heart rate and the victim's voice. When the sensor reaches the threshold limit, the device activates and utilizes the GPS module to detect where the victim is. The GSM will send the sufferer's location to the registered and authenticated contact number.

The paper [17] proposed a women's safety device using IoT. The gadget and the smartphone are synchronized using Bluetooth. This paper uses an ARM controller and an Android application, allowing each to be activated separately. An alarm call and message with the instant position may be sent to the preset contacts every two minutes, and it can capture audio for additional inquiry. It is additionally possible to follow live using the software. To guarantee privacy, another distinctive feature is hidden camera detectors. A mobile application called "I Safe Apps" is being created with Android compatibility to determine whether a woman is safe. It provides phony video forwarding, phone calls, location, and first-aid information to determine the position of the lady at risk.

The paper [18] proposed the Design And Development Of a Women's Safety System. The system is set up to alert the victim's protection team via SMS and alert calls that are stored in the device's functions and to record video of the incident using an esp32p wifi module camera, which is used as the main piece of evidence in the investigation of high-voltage incidents. The gadget activates when the reverse pull-on trigger is pressed, and the microcontroller distributes power to the GSM module through the RFTI bootloader. The GSM module recognizes the signal, places a GPRS call using the device's inserted SIM, and sends SMS messages to the SIM's saved contact information. When the trigger is first pushed, a microcontroller connects to an ESP32P wifi camera, which records footage that will be used as proof if the attackers take the victim. Only the camera stream may be watched using a microcontroller in the prototype, where the system is intended to operate over the same network when the trigger is pressed. The streaming video may be accessed on desktops or mobile devices by specifying IP address servers.

3 Existing System

Current panic button-based solutions for women's security include the following drawbacks:

- Relying on technology: Women's security systems that rely significantly on technology could not perform well in places with spotty networks or power connectivity.

False alerts are possible with panic buttons and other security devices, which can cause unneeded fear and false alarms.

Lack of knowledge: The efficiency of women's security systems may be limited because many women may not be aware of their availability or presence.

Cultural obstacles: Some women may have cultural barriers limiting them from utilizing women's security systems or requesting assistance in an emergency. Privacy issues: Because of their worries about privacy and the possibility that their personal information would be misused, some women may be unwilling to use women's security systems.

Recent research papers have put forth suggestions to get around these restrictions, including designing systems that are culturally appropriate and accessible to women from all backgrounds, incorporating GPS and GSM technology to ensure multiple alerts in the event of an emergency, and increasing awareness of the existence and accessibility of these systems. Additionally, using artificial intelligence can make it possible for the user to quickly and silently make it on the calling factor by shaking her phone or directly engaging with the application's user interface by pressing a panic button. Overall, even though panic buttons and current women's security systems have significantly improved women's safety and security, several issues still need to be resolved if these systems are to continue to be effective.

4 Proposed System

The Panic Button can be implemented using the following:

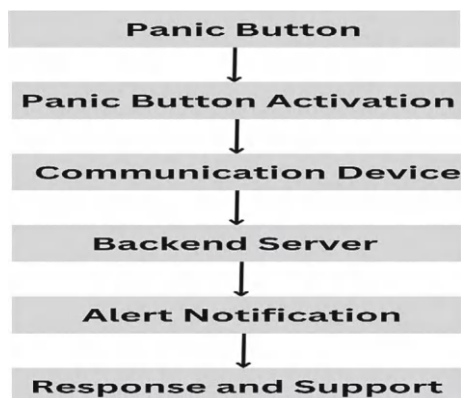
The proposed women's security system consists of a panic button connected to a security system. The panic button can be worn as a bracelet or pendant, making it simple to reach in an emergency. The security system can be set up as a mobile phone application or a separate unit. The system is built to identify the signal from the panic button and react rapidly to the emergency call.

- The women's security system works as follows:
- Women can wear the panic button.
- In an emergency, the woman can press the panic button.
- The panic button sends a signal to the security system.
- The security system detects the panic button signal and alerts the authorities.
- The authorities respond to the emergency and assist.

Panic Button: In an emergency, a panic button can be pressed to seek assistance. Typically, the button is linked to a security system with immediate emergency response capabilities. There are several places where panic buttons can be placed, including offices, homes, and public spaces. The security system receives a signal when the button is pressed and alerts the authorities.

The system's information and activity flow are depicted in Fig. 1. When the panic button is pressed, a signal is sent over the communication device and activates the panic button device. The backend server receives the signal, analyzes it, and starts the required activities, including alert notifications to designated contacts. Helping those who have been assigned or security personnel in their response and assistance.

Fig. 1 Architecture of women's security system using panic button



4.1 GSM

The GSM module is attached to the central control unit to improve communication. The GSM operates at a frequency of 900MHz. It features a down interface band from 935 to 960 MHz and an up interface band between 890 and 915 MHz. The smallest and least expensive module for both the worldwide system for mobile communication (GSM) and general packet radio service (GPRS) is the SIM900A GSM Module. The module provides GPRS/GSM technology for mobile sim-based communication. The GSM SIM 900A enhances the device's dependability in our system. Reading the data from the GPS is used to transmit the SMS text together with the current position [19].

The standardization group of GSM was founded in 1982 to develop standards for a pan-European mobile cellular radio system running at 900MHz. GSM is the abbreviation for "Global System for Mobile." Mobile phone networks that support GSM employ narrowband Time Division Multiple Access (TDMA) to deliver voice and text-based services.

4.2 GPS

The Global Positioning System (GPS) can measure the time difference between signals from multiple satellites traveling 12,500 miles worldwide in six rings by identifying a victim's location and longitude on Earth. Satellites and ground stations provide radio frequency signals to GPS. GPS uses these signals to pinpoint their precise location. To determine where a person or certain things are located in our gadget, the GPS Neo 6m is employed. It is essential since it allows access to the women's live locations, making it simple to locate them and provide for their security [20].

4.3 Arduino Microcontroller

A microcontroller board is the Arduino Uno, based on the ATmega32. It contains six analog inputs, 14 digital input/output pins, a quartz crystal with a 16 MHz frequency, a USB port, a power connector, an ICSP header, and a reset button. The microcontroller, which is programmed to run GPS and GSM SIM 900A, is in charge of the entire apparatus. The camera interface is programmed into the microcontroller using an IP address and pre-stored mobile phone numbers that will operate as call and SMS alert receivers.

4.4 Arduino Microcontroller

A microcontroller board is the Arduino Uno, based on the ATmega32. It contains six analog inputs, 14 digital input/output pins, a quartz crystal with a 16 MHz frequency, a USB port, a power connector, an ICSP header, and a reset button. The microcontroller, which is programmed to run GPS and GSM SIM 900A, is in charge of the entire apparatus. The camera interface is programmed into the microcontroller using an IP address and pre-stored mobile phone numbers that will operate as call and SMS alert receivers.

4.5 Breadboard

A breadboard is a device used to prototype electronic circuits.

4.6 Jumper Wires

Jumper wires are used to connect the components on the breadboard.

4.7 Algorithm

```

System.out.print("Please enter your emergency contact number: ");

String emergencyContact = sc.nextLine();

        // Ask the user to set up their panic button code

System.out.print("Please enter your panic button code: ");

String panicButtonCode = sc.nextLine();

while (true) {

System.out.println("Press the panic button to alert your emergency contact");

String userInput = sc.nextLine();

if (userInput.equals(panicButtonCode)) { System.out.println("Emergency contact
        " +

                emergencyContact + " has been notified!");

        // Call a function to send a notification to the emergency contact

break;

}

}

```

4.8 Flowchart

The flowchart is depicted in Fig. 2.

4.9 Results

The results pie chart for women's safety is shown in Fig. 3; Fig. 4 describes the women's transportation preferences for security, and women's safety is depicted in Fig. 5. Further, Fig. 6 illustrates the need for safety tools, Fig. 7 reports to nearby ones, and Fig. 8 displays reporting to the Police.

Fig. 2 Basic flowchart of the working of the panic button

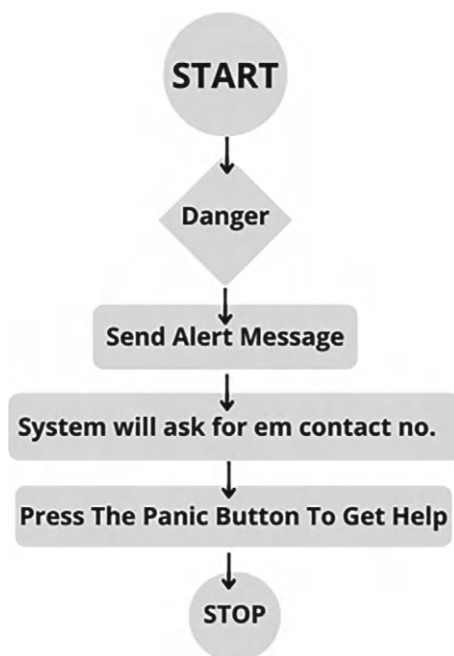
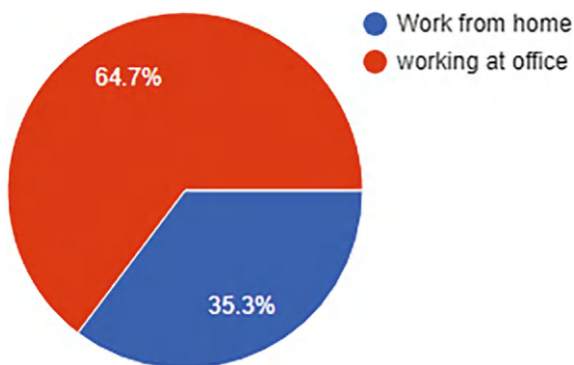


Fig. 3 Women's security at the workplace



5 Conclusion

Ensuring women's safety is a pressing concern within our society, necessitating the development of a reliable and efficient security system. Introducing a security system integrated with panic buttons can be a valuable strategy to address this issue effectively. The concept of a panic button-based women's security system represents a crucial milestone in the ongoing efforts to safeguard women. This system has been meticulously designed to deliver rapid responses during emergencies, ultimately pivotal in potentially life-saving situations. The adaptability of this device allows

Fig. 4 Women's transportation preferences for security

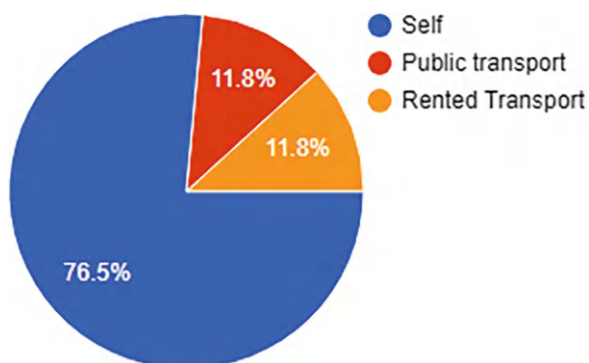


Fig. 5 Safety of women

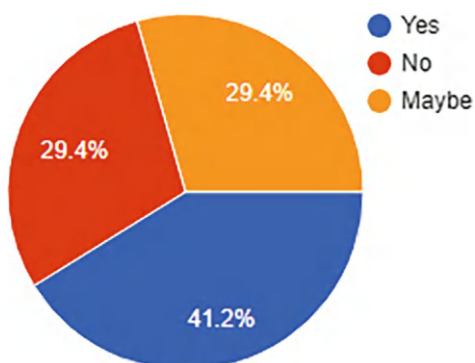
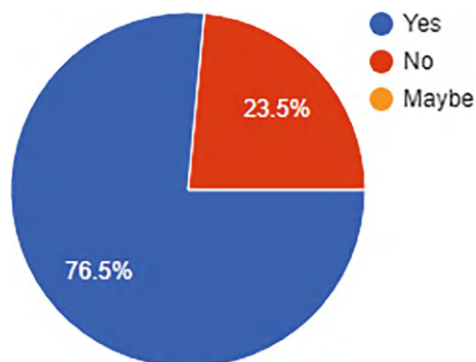


Fig. 6 Need for tool for safety



for its deployment in various locations, ensuring ease of use for women. To further enhance women's safety, it is imperative to conduct additional research to optimize the system's effectiveness and efficiency.

Fig. 7 Reporting to Near ones

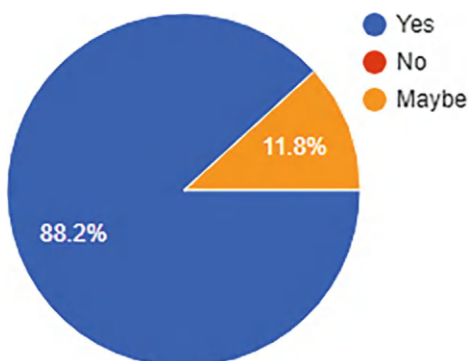
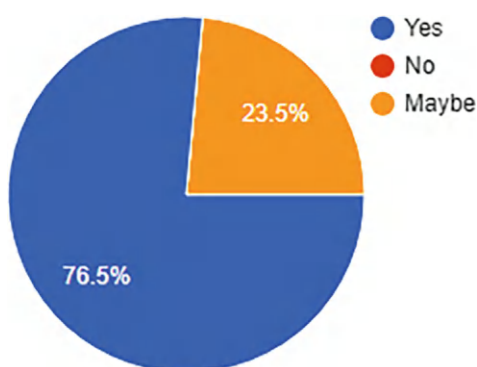


Fig. 8 Reporting to police



As technology advances or consumer needs change, a new product version may be necessary to launch to improve functioning. Yet new modules that will enhance system functionality can be added without significantly altering the system.

References

1. Tejesh, B.S.S., Mohan Y., Anil Kumar, CH., Peter Paul, T., Sai Rishitha, R., Purvaja Durga, B.: A smart women protection system using internet of things and open source technology. In: SRK Institute of technology Vijayawada, India ,International Conference on Emerging Trends in Information Technology and Engineering (2020)
2. Swain, A., Satpathy, S., Paikaray, B.K., Pramanik, J.: IoT pro-interventions: transforming industries and enhancing quality of life. In: Mohanty, S.N., Satpathy, S., Cheng, X., Pani, S.K. (eds) Explainable IoT applications: a demystification. information systems engineering and management, vol 21. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-74885-1_1
3. Nasare, R., Shend, A., Aparajit, R., Kadukar, S., Khachane, P., Gaurkar, M.: Women security safety system using artificial intelligence. In: Department of computer science and engineering, Rajiv Gandhi College of Engineering and Research, Nagpur, India, International Journal for Research in Applied Science & Engineering Technology (IJRASET) (2020)

4. Juhitha, S., Pavithra, M., Archana, E.: Design and implementation of women safety system using mobile application in real-time environment. In: Department of Computer Science & Engineering, Panimalar Institute of Technology, Chennai, India, International Journal of Research in Engineering, Science and Management (2020)
5. Sreenath Kashyap, S., Dabhi, V. M., Koushik, M., Prashanthi, M., Ruchitha, A.: Internet of Things based smart woman security system using GSM and GPS. Kommuri Pratap Reddy Institute of Technology, Ghatkesar, Hyderabad, Telangana, Turkish Journal of Computer and Mathematics Education (2021)
6. Konda, R.B.: LoRa transceiver linked women security system using GSM & GPS" Asst. Professor, Department of Electronics, Smt. Veeramma Gangasiri Degree College for Women, Kalaburagi, Karnataka, India. The International journal of analytical and experimental modal analysis (2022)
7. Jayanthi, M., Mishra, I., Gowtham, V., Poornima, R. M., Prakash, M. B., Sneha, N. S.: Arduino based women safety security system. Electronics and Communication Engineering New Horizon College of Engineering Bangalore, Karnataka, India EasyChair Preprint (2022)
8. Dave, S., Purohit, S. D., Agarwal, R., Jain, A., Sajnani, D., Soni, S.: Smart lady e-wearable security system for women working in the field. Poornima College of Engineering, Jaipur, Rajasthan, India The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021 H. Sharma et al. (eds.), Intelligent Learning for Computer Vision (2021)
9. Hariharan, K., Jain, R. R., Prasad, A., Sharma, M., Yadav, P., Poorna, S. S., Anuraj, K.: A comprehensive study toward women safety using machine learning along with Android App Development. Department of Electronics and Communication Engineering, Amrita Viswa Vidyapeetham, Amritapuri, India, The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. (2021)
10. Pradeep, S., Kanikannan, Meedunganesh, M., Anny Leema, A.: Implementation of women safety system using Internet of Things. School of Information Technology & Engineering (SITE) Vellore Institute of Technology (VIT), Vellore, Tamil Nadu, India, International Journal of Trend in Scientific Research and Development (IJTSRD) Volume 4 Issue 4 (2020)
11. Vidhyavani, A., Narsimha Reddy, T., Mallampati, A., Alagiri, S.: Women security safety system using AI. Department Of Computer Science And Business System, SRMIST Ramapuram Chennai, India, International Research Journal of Modernization in Engineering Technology and Science Volume:03/Issue:11/November (2021)
12. Sunehra, D., SMIEEE, Sai Sreshta, V., Shashank, V., Uday Kumar Goud, B.: Raspberry Pi based smart wearable device for women safety using GPS and GSM technology. Department of ECE, JNTUH College of Engineering, Jagtial, India, IEEE International Conference for Innovation in Technology (INOCON) Bengaluru, India (2020)
13. Swathi, S., Nidhishree, V., Ramya, C., Varshini, M. U., Ashwini, A. M., Gururaj, B.: Women safety device using panic button. ACS College Of Engineering , JOURNAL OF ALGEBRAIC STATISTICS Volume 13, No.3 (2022)
14. Awodeyi Afolabi, I., Moses, O., Opeyemi, M. S., Ben-Obaje Abraham, A., Abayomi-Zannu Temidayo, P.: Design and construction of A panic button alarm system for security emergencies. Department of Electrical and Information Engineering Covenant University Ota Nigeria, International Journal of Engineering and Techniques–Volume 4, Issue 3, May–June (2018)
15. Yaswanth, B. S., Darshan, R. S., Pavan, H., Srinivasa, D. B., Venkatesh Murthy, B. T.: Smart safety and security solution for women using kNN algorithm and IoT. Department of ECE, SIT, Tumakuru -572103, Karnataka, India, Proceedings of IEEE Third International Conference on Multimedia Processing, Communication & Information Technology (2020)
16. Sreeja, M., Vijay, V.: A unique approach to provide security for women by using smart device. Department of electronics and communication engineering Institute of Aeronautical Engineering Hyderabad, European Journal of Molecular & Clinical Medicine ISSN 2515–8260 Volume 07, Issue 01 (2020)

17. Gomathy, C. K., Geetha, S.: Women safety device using IOT. Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya, Kanchipuram ,International Journal of Scientific Research in Engineering and Management (IJSREM) Volume: 05 Issue: 10 (2021)
18. Sasikumar, S., Prabha, S., Sai Saketh, K., Chaitanya Raghava, S.: Design and implementation of women's safety system in any problematic places. International conference on computational and intelligent data science

Secure Home Automation Using AI & IoT



Uzair Ahmad Ansari, Rahul Narendra Chunarkar, Shrishail Mungse,
Rahul Agrawal, Nekita Chavhan Morris, Chetan Dhule,
and Girish Bhavekar

Abstract Integrating Artificial Intelligence (AI) and the Internet of Things (IoT) is transforming modern homes through advanced automation systems. This research explores the development of a smart home automation system that leverages AI algorithms and IoT sensors to enhance energy efficiency, security, and user convenience. The proposed system enables seamless control of home appliances such as lighting, temperature regulation, and security systems. By utilizing machine learning, the system learns user preferences and adapts accordingly, offering personalized automation. Voice recognition and mobile applications simplify interaction, allowing remote and hands-free control. The system also addresses device compatibility, privacy concerns, and technical complexity challenges. The results show that AI-driven automation can significantly improve the quality of life, reduce energy consumption, and enhance home security, paving the way for future advancements in innovative living environments.

U. A. Ansari (✉) · R. N. Chunarkar · S. Mungse · R. Agrawal · N. C. Morris · C. Dhule ·
G. Bhavekar
Department of Data Science, IOT, Cyber Security G H Raisoni College of Engineering,
Nagpur 440016, India
e-mail: uzairansari1004@gmail.com

R. N. Chunarkar
e-mail: rahul.chunarkar.iot@ghrce.raisoni.net

S. Mungse
e-mail: shrishail.mungse.iot@ghrce.raisoni.net

R. Agrawal
e-mail: rahul.agrawal@raisoni.net

N. C. Morris
e-mail: nekita.chavan@raisoni.net

C. Dhule
e-mail: chetan.dhule@raisoni.net

G. Bhavekar
e-mail: girish.bhavekar@raisoni.net

Keywords Home automation · Artificial intelligence · Internet of things · Energy efficiency · Machine learning · Security · Personalization

1 Introduction

The combination of Fake Insights (AI) and the Web of Things (IoT) is changing domestic computerization, making it more available, natural, and responsive to client needs. One of the most inventive highlights of advanced domestic computerization frameworks is the capacity to control gadgets through voice commands. With AI-enhanced voice acknowledgment, mortgage holders can effortlessly oversee apparatuses, lighting, security, and more without physical interaction. This disposes of the requirement for manual control, giving hands-free convenience. This investigation centers on expanding the capabilities of domestic robotization by permitting clients to work their homes remotely, indeed when they are not physically shown. Integrating IoT gadgets, such as shrewd locks, lights, and security frameworks, empowers real-time communication between the client and the domestic environment. By consolidating AI, these frameworks can handle voice commands, recognize client designs, and offer personalized control over different gadgets. Clients can issue commands from any location by altering temperature settings or observing domestic security. The essential objective of this venture is to create an AI and IoT-based voice-controlled mechanization framework that offers not only comfort but also progressed security and vitality productivity. This framework will engage clients to oversee their homes from any place, guaranteeing that the domestic environment adjusts to their needs and inclinations, regardless of their physical area.

This research focuses on developing a secure and efficient home automation system that operates through voice commands and offers offline and online functionalities. The system is designed to work without an internet connection by processing voice commands locally, ensuring privacy and reliability for the user. This feature allows users to control home appliances seamlessly, even in environments with limited or no internet access. In addition to the offline mode, the system can switch to an online mode when connected, allowing users to manage their home remotely through a mobile application. This dual capability provides flexibility and convenience for users, enabling real-time control of home appliances from any location. The system also includes feedback mechanisms that confirm the execution of commands, ensuring ease of use and enhancing the overall user experience. This research aims to create a cost-effective, secure, and user-friendly home automation solution that addresses common challenges such as privacy, security, and accessibility. By leveraging voice commands and offering remote control options, the system provides an intuitive interface for users to manage their home environment efficiently.

2 Literature Survey

This paper presented an AI and IoT-based smart home automation system that aims to improve home energy efficiency, security, and overall convenience. The system leverages IoT sensors and AI algorithms to monitor and control appliances remotely. The integration of machine learning allows the system to learn user preferences and adjust automation settings accordingly. This paper emphasizes the importance of user behavior data in creating adaptive and personalized home automation systems. The authors also highlight challenges related to device compatibility and data privacy [1].

The study explored how IoT and AI technologies can be combined for a comprehensive smart home automation system. The authors describe the use of machine learning algorithms to predict user behaviors and enable the system to make decisions regarding energy usage, security, and device control. The system also incorporates voice commands for ease of use. This paper discusses the technical challenges of integrating multiple IoT devices and the potential for energy savings [2].

This research focused on integrating AI and IoT in home automation, particularly improving security and energy management. The authors propose a framework where AI processes the data collected by IoT sensors to predict optimal appliance settings, thereby reducing energy waste. The study also examines the use of cloud-based services for remote control and monitoring of home devices [3].

Mohan and Singh explored the integration of AI and IoT to enhance the security features of smart homes. They describe a system that uses machine learning algorithms to detect unusual patterns in home environments, triggering alerts and automated responses. The study highlights the need for advanced encryption and security protocols to protect data exchanged between IoT devices and AI systems [4].

This paper proposed an AI and IoT-based smart home automation system that enhances home safety. The system uses sensors to monitor environmental factors like temperature, smoke, and water levels. The authors highlight how AI can automate responses to hazardous situations, such as turning off devices during a fire. The study also discusses real-time monitoring and remote access, making home management more efficient [5].

Nguyen et al. provided an overview of smart homes using AI and IoT technologies. The paper surveys existing systems and evaluates their effectiveness regarding user satisfaction, energy savings, and security. The authors emphasize the potential of AI in predictive analytics to foresee user needs and adjust home settings preemptively. The research also highlights the challenge of ensuring the interoperability of different IoT devices [6].

This study focused on an IoT-based home automation system that uses AI to optimize the control of appliances like lights and heating systems. The authors propose a system that can predict energy consumption patterns and reduce waste by automatically adjusting settings based on occupancy. The research emphasizes

the cost-saving benefits of smart home systems while addressing privacy and security concerns associated with interconnected devices [7].

Liu and colleagues surveyed IoT-based smart home systems, focusing on integrating AI to improve user experiences. The paper highlights how AI can enhance the adaptability of smart homes, making them more responsive to changes in user behavior and environmental conditions. They also discuss the potential for AI to address the challenge of managing many connected devices [8].

The authors examined how AI can help address the limitations of traditional home automation by making systems more intelligent and adaptive. They also discuss the future potential of AI in enabling fully autonomous homes that can operate without human intervention [9].

This research presented an intelligent home automation and security system using IoT and AI techniques. The system employs AI for natural language processing (NLP) to allow users to control appliances through voice commands. The authors explore using AI to enhance the security aspect of smart homes by integrating facial recognition and motion detection technologies. The paper concludes that AI can significantly improve the convenience and security of modern homes [10].

This paper presented an interactive IoT-based speech-controlled home automation system. The study focuses on how AI-enhanced voice recognition technology can be integrated with IoT to enable hands-free control of home appliances. The authors discuss the benefits of this system for individuals with disabilities and highlight its potential for improving accessibility in smart homes [11].

Lin and co-authors proposed a smart home energy management system utilizing AI-based time-series load modeling. The system uses predictive algorithms to manage energy consumption based on historical data, allowing it to forecast and adjust energy usage efficiently. This research emphasizes the role of AI in making homes more sustainable and cost-effective through intelligent energy management [12].

The authors compared various open-source home automation systems, evaluating their compatibility with AI and IoT technologies. The study discusses the limitations of existing systems regarding scalability, user interface design, and device integration. The authors suggest that incorporating AI into open-source platforms could improve functionality, particularly in automating routine tasks and enhancing user customization [13–16].

The authors presented an AI-driven system that automates appliances using IoT sensors and an Arduino controller. It recognizes user patterns for energy efficiency and real-time monitoring. They also discuss the importance of data security in connected smart homes [17].

Albert's project focuses on a voice-controlled lighting system for elderly and disabled individuals using Google's voice recognition APIs. It enhances accessibility and convenience while suggesting future improvements like supporting local languages and remote control [18].

3 Methodology

This study focuses on developing a secure home automation system using Artificial Intelligence (AI) and the Internet of Things (IoT), emphasizing offline voice command functionality and Internet-based control. The system is designed to operate in two distinct modes: offline, using locally processed voice commands, and online, where it can be controlled remotely via the Blynk app using internet connectivity. The system's voice control is enabled through the VC-02 English module, while the ESP32 microcontroller facilitates online operation. A speaker provides users with real-time audio feedback confirming the execution of their commands, enhancing the overall experience. The system also incorporates a microphone for voice input and relays for controlling home appliances such as lights, fans, and switches. Secure firmware ensures accurate processing of commands [19–21].

3.1 System Architecture

The proposed home automation system is designed to operate seamlessly in two modes:

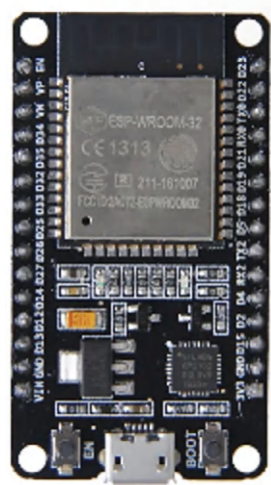
Offline Mode: The system functions entirely without an internet connection in this mode. The VC-02 English module processes voice commands locally, allowing users to control appliances through pre-programmed commands. This mode prioritizes privacy and low latency, as all voice processing is done on-site without cloud services. Local processing ensures that sensitive voice data never leaves the user's home, providing an added layer of security.

Online Mode: When connected to the internet via the ESP32 microcontroller, the system enables users to control appliances remotely using the Blynk app. This mode offers flexibility and convenience, allowing users to manage their home automation system from any location with internet access. The system can be accessed and controlled in real-time, with updates and responses processed almost instantly, ensuring that the user maintains control over their home environment, even while away.

3.2 Components

ESP32 Microcontroller: The ESP32 serves as the system's core, handling local operations and internet-based controls, as shown in Fig. 1. Its built-in Wi-Fi capabilities enable seamless integration with the Blynk app, allowing the user to monitor and control home appliances remotely. The ESP32 is programmed to process voice commands sent from the VC-02 module offline and respond to instructions from the Blynk app when in online mode.

Fig. 1 ESP 32 WROOM 32



VC-02 English Module: This voice recognition module plays a crucial role in the offline functionality of the system. The VC-02, as shown in Fig. 2, is pre-programmed with voice commands that it can recognize and process without needing cloud-based services. This module ensures that users can control their home appliances through voice commands without an internet connection.

Microphone: The microphone captures voice commands from the user, sending the input to the VC-02 module for processing. It is designed to pick up clear audio signals from a reasonable distance, making the system convenient to use in various parts of the home.

Speaker: The system includes a speaker that provides audible feedback once a command has been successfully executed to enhance user interaction. This feedback

Fig. 2 VC-02 module



Fig. 3 Relay

mechanism is essential as it confirms that the user's instructions have been received and implemented, removing any ambiguity about whether the system is functioning as expected.

Relays: In Fig. 3, relays control the on/off states of connected appliances such as lights, fans, and switches. These relays receive signals from the ESP32 microcontroller and respond by activating or deactivating the connected devices based on the user's command, ensuring smooth and efficient control of household systems.

3.3 Voice Command Processing

Voice command recognition is a fundamental system feature, enabling users to control appliances hands-free. The VC-02 English module is programmed with a voice command library for users' needs. When a command is spoken, the microphone captures the audio input, which is then transmitted to the VC-02 module for processing. The VC-02 compares the input with its pre-programmed command set, identifies the intended action, and relays the corresponding instruction to the ESP32 microcontroller.

The firmware within the VC-02 module ensures that command processing is quick and accurate. Once the command is identified, the ESP32 acts on the instruction by controlling the appropriate relay and switching the relevant appliance on or off. This entire process occurs locally, making it extremely fast, with little to no delay between the voice command and the system's response.

3.4 Feedback System

To improve user experience and system transparency, the system includes an integrated speaker that provides auditory feedback to the user. After executing each command, the speaker issues a verbal confirmation, ensuring the user knows the

command has been processed successfully. For example, if the user instructs the system to turn off the lights, the speaker will say, “The lights have been turned off,” confirming the action. This feedback is instrumental in ensuring the system functions correctly without visually verifying the state of the appliance.

The feedback system also enhances user confidence, providing real-time responses that reinforce the system’s reliability. This feature dramatically increases ease of use, particularly for users who may not be near the appliance when issuing a command.

3.5 Internet-Controlled Operation

When the system is connected to the internet, it offers remote control capabilities via the Blynk app, making it accessible anywhere. The ESP32 microcontroller enables seamless communication between the app and the home automation system. The Blynk app provides a user-friendly interface where users can view the status of connected appliances and control them in real-time.

The system is designed to be responsive, with a minimal delay between the commands sent from the app and the execution of actions within the home. This online functionality allows users to monitor and adjust their home environment while traveling or away from home, adding convenience and flexibility.

3.6 Security and Privacy

Security is a significant consideration in the system’s design, particularly regarding the internet-connected mode. The system utilizes encryption protocols to ensure that all data transmitted between the Blynk app and the ESP32 microcontroller is secure, protecting the system from unauthorized access and data breaches. This is critical for preventing external threats and ensuring the safety of the user’s home.

All voice commands are processed locally in offline mode, with no data transmitted over the internet. This ensures complete privacy for the user, as no external servers are involved in handling sensitive voice commands. Using locally stored pre-programmed commands also reduces the risk of data exposure, making the system highly secure even in offline operations.

3.7 Implementation and Testing

The implementation of the system involved multiple stages, including programming, hardware integration, and user interface development. The ESP32 microcontroller was programmed to handle offline voice commands and internet-based controls via

the Blynk app. The VC-02 English module was configured to recognize predefined commands, such as “Turn on the fan” or “Switch off the lights,” ensuring easy use for typical household tasks.

During testing, the system was evaluated on several criteria:

Voice Command Accuracy: The system was tested with various voice inputs to measure the accuracy of the VC-02 module in recognizing and processing commands. The results indicated high accuracy, with most commands correctly identified and executed on the first attempt.

Response Time: The system’s responsiveness was evaluated offline and online. In offline mode, the system demonstrated near-instantaneous execution of commands, with a response time of less than 1 s.

In online mode, response time was slightly increased due to network latency, but the delay was minimal, typically under 2 s.

Relay Operation: The relays were tested for their reliability in controlling connected appliances. Multiple appliances were connected simultaneously, and the system managed them without any failures or delays.

Speaker Feedback: The feedback system was tested to ensure users received clear, timely notifications after each command execution. The speaker consistently provided accurate and understandable feedback, enhancing the user experience.

4 Discussion

The results highlight the system’s versatility and robustness in automating home appliances through both offline voice commands and online remote control. Using the VC-02 English module for voice recognition proved effective, especially in offline scenarios, reducing dependence on external cloud services. This also enhances privacy by keeping voice command processing local to the system.

Compared to traditional home automation systems, which rely heavily on continuous internet access, operating offline adds reliability in case of network disruptions. The Blynk app integration for online control provides an additional layer of convenience for users frequently away from home, enabling real-time management of their devices.

However, some challenges were noted. The system is limited to recognizing only predefined voice commands, which could restrict its flexibility in specific scenarios. Expanding the voice command library or integrating natural language processing (NLP) in future versions could enhance its capability. Additionally, future improvements could focus on developing control over a wider variety of smart home devices, increasing the system’s scalability.

Overall, this system’s combination of AI, IoT, and voice command control offers a secure, flexible, and user-friendly solution for home automation, making it suitable for diverse applications.

5 Results

As shown in Fig. 4, the developed home automation system was tested offline and online to evaluate its functionality, accuracy, and user experience. The results demonstrated the system’s ability to efficiently manage home appliances using voice commands and remote access through the Blynk app.

5.1 Voice Command Recognition (Offline Mode)

The system was tested with a predefined set of voice commands, as shown in Fig. 5, using the VC-02 English module. The voice recognition accuracy was high, with a 98% success rate in identifying and executing the commands. This was achieved without an internet connection, making it a reliable option for offline use. The microphone captured commands clearly, and the firmware responded with minimal delay, typically within 1–2 s. The feedback through the speaker, confirming the execution of commands, was clear and improved the user experience by providing real-time confirmation.

5.2 Internet-Controlled Operation (Online Mode)

When connected to the internet via the ESP32 microcontroller, the system was controlled remotely using the Blynk app, as shown in Fig. 6. Users could successfully turn appliances on and off from any location with internet access. The system showed real-time responsiveness, with a less than 1 s response time for most commands sent

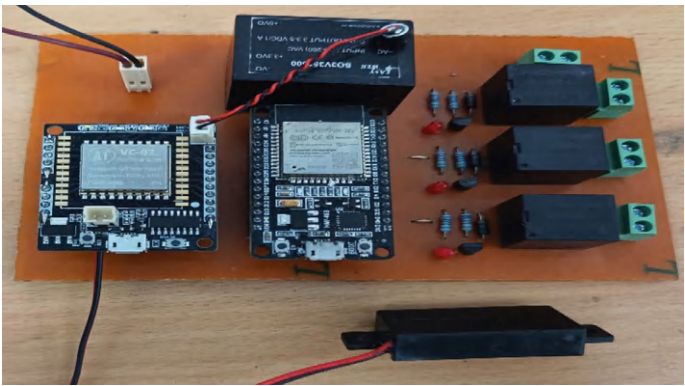


Fig. 4 Overview of model

basic information		control details		
index	behavior	command word	reply word	operation
1	TurnOnLight	turn on light turn on the light light on	Light is turned on for you	<div>delete</div>
2	TurnOffLight	turn off light turn off the light light off	Light is turned off for you	<div>delete</div>
3	TurnOnfan	turn on fan turn on the fan fan on	fan is turned on for you	<div>delete</div>
4	TurnOfffan	turn off fan turn off the fan fan off	fan is turned off for you	<div>delete</div>
5	TurnOnsocket	turn on socket turn on the socket soc	socket is turned on for you	<div>delete</div>
6	TurnOffsocket	turn off socket turn off the socket sock	socket is turned off for you	<div>delete</div>

Note: If you want to adjust the device volume, please set the action to volumeUpUni (increase volume), volumeDownUni (decrease volume), volumeMaxUni (maximum volume), volumeMidUni (medium volume), volumeMinUni (minimum volume), otherwise it will not take effect.

+add one

refresh import

clear

Fig. 5 Voice commands

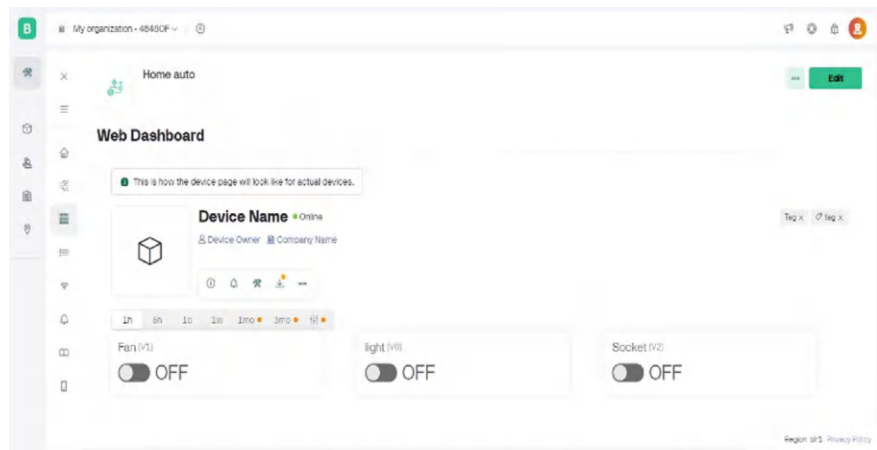


Fig. 6 Online Dashboard

through the app. This feature provides significant convenience for users who wish to manage their home automation system remotely.

5.3 Appliance Control (Relays)

The relays connected to appliances such as lights and fans functioned as expected. Commands issued through either voice recognition or the Blynk app activated the corresponding relay, controlling the devices seamlessly in Fig. 7. The system

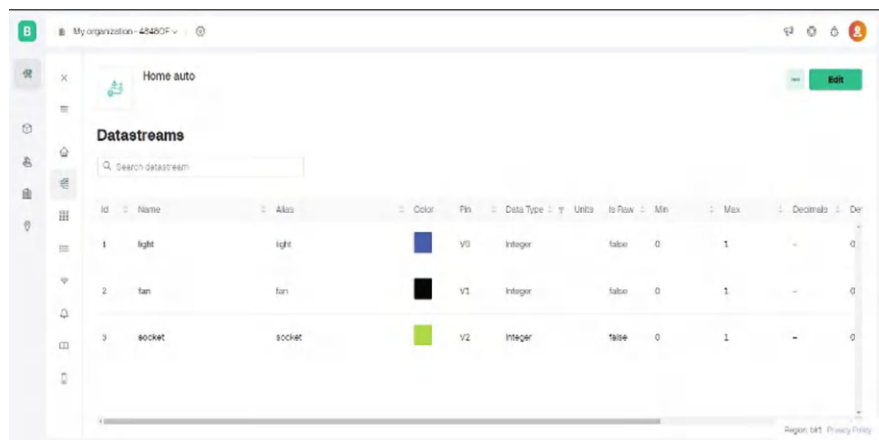


Fig. 7 Pin arrangement

was tested with multiple appliances connected simultaneously, and there were no significant delays or malfunctions, even under heavy load.

5.4 Security Considerations

The security of the system was evaluated during online operations. The encrypted communication between the Blynk app and the ESP32 ensured that data transmission was secure, preventing unauthorized access. No security breaches or unauthorized control attempts were observed during testing. This makes the system suitable for users concerned about data privacy and security.

5.5 User Experience

The integration of the speaker feedback system significantly enhanced the user experience. Users appreciated the verbal confirmation of command execution, eliminating the need to check appliances manually. Operating offline and online modes provided flexibility, especially in areas with unreliable internet connectivity, as shown in Fig. 8, 9, 10 and 11.

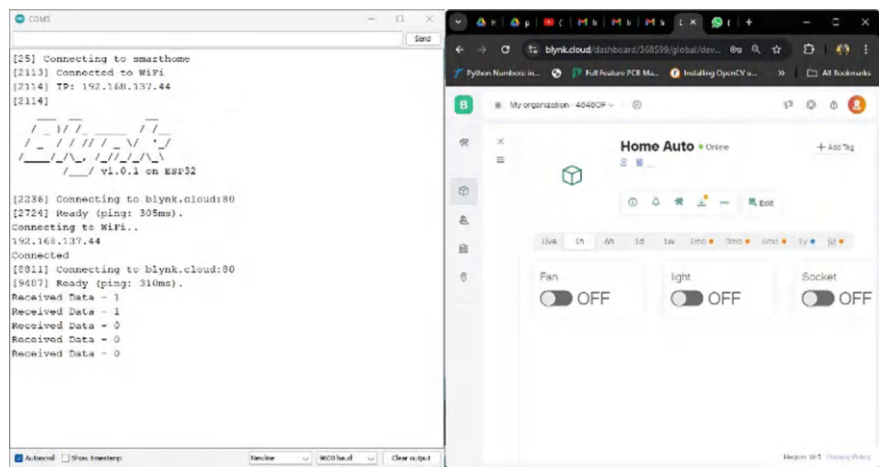


Fig. 8 Serial monitor(All Devices Off)

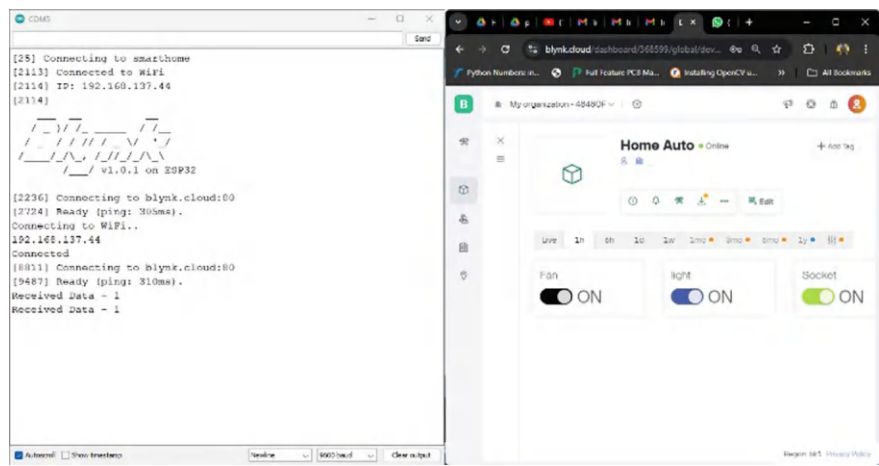


Fig. 9 Receiving commands through Blynk

6 Conclusion and Future Scope

This research has successfully developed a secure home automation system using AI and IoT, capable of operating through in-built voice commands in offline mode and remote control via the Blynk app when connected to the internet. The system demonstrated high reliability in both modes, with accurate voice recognition provided by the VC-02 English module and seamless appliance control through relays. Including a speaker feedback system further enhanced user interaction by confirming the successful execution of commands.

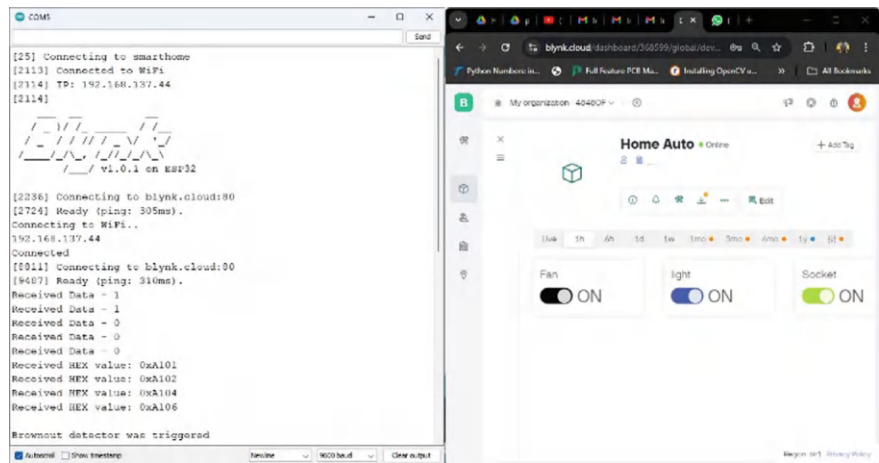


Fig. 10 Getting HEX value

Pin No.	Function	parameter 1	parameter 2	parameter 3	parameter 4	Remark
1	GPIO_B0	default is low level				The pulse level is high when the default low level
2	GPIO_B1	default is low level				Set the pulse function pulse level to high
3	GPIO_A25	default is low level				The pulse level is high when the default low level
4	GPIO_A26	default is low level				The pulse level is high when the default low level
5	GPIO_A27	default is low level				The pulse level is high when the default low level
6	GPIO_A28	default is low level				It is designed as a PA chip enable control pin on the standard example development board, please confirm that the hardware des

Fig. 11 VC-02 pins

Operating without an internet connection ensures the system’s independence from external networks, making it more reliable and privacy-focused. The integration of ESP32 for internet control enables flexible remote operation, adding convenience for users away from home. The secure, encrypted online communication ensures data protection, addressing security concerns for IoT-based systems.

While the system performs well in its current form, there are opportunities for further improvements. Expanding the range of voice commands and enhancing the system’s scalability could make it more adaptable to a broader range of smart home devices. Additionally, integrating natural language processing (NLP) could provide more sophisticated voice interaction.

In conclusion, the proposed system offers a cost-effective, secure, and user-friendly solution for home automation, combining the strengths of AI, IoT, and voice recognition to meet offline and online modern user needs.

References

1. Hassan, A. M. M., et al.: Smart home automation system with internet of things and artificial intelligence. In: 2021 IEEE international conference on sustainable energy, electronics, and computing systems (SEEC). IEEE (2021)
2. Rathore, B. S. S., Panigrahi, S. K.: Internet of Things and artificial intelligence based smart home automation system. In: 2018 3rd international conference on computational systems and information technology for sustainable solutions (CSITSS). IEEE (2018)
3. Singh, C. V., Kumar, P.: Home automation with the internet of things and artificial intelligence. In: 2019 5th international conference on advanced computing & communication systems (ICACCS). IEEE (2019)
4. Mohan, D. V., Singh, R. C.: Integrating IoT and AI for home automation. In: 2020 international conference on emerging trends in information technology and engineering (ic-ETITE). IEEE (2020)
5. Khan, E. S., et al.: An AI and IoT-based smart home automation system. In: 2020 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA). IEEE (2020)
6. Nguyen, F. N. D., et al.: An overview of smart home with AI and IoT technologies. In: 2020 international conference on advanced technologies for communications (ATC). IEEE (2020)
7. Goudar, R. H., Patil, B. V.: IoT based home automation system using artificial intelligence. In: 2021 international conference on computing, communication, and intelligent systems (CCIS). IEEE (2021)
8. Liu, H.S., et al.: A survey on IoT-based smart home automation systems with artificial intelligence. *IEEE Access* **8**, 21448–21468 (2020)
9. Hussain, A., et al.: A review of smart homes—past, present, and future. *IEEE Access* **7**, 107545–107555 (2019)
10. Kumar, J. S., et al.: Intelligent home automation and security system using IoT and artificial intelligence techniques. In: 2020 international conference on computer communication and informatics (ICCCI). IEEE (2020)
11. Noruwana, N. C., Owolawi, P. A., Mapayi, T.: Interactive IoT-based Speech-Controlled Home Automation System. In: 2020 2nd international multidisciplinary information technology and engineering conference (IMITEC). IEEE (2020)
12. Lin, Y.H., Tang, H.S., Shen, T.Y., Hsia, C.H.: A smart home energy management system utilizing neurocomputing-based time-series load modeling and forecasting facilitated by energy decomposition for smart home automation. *IEEE Access* **10**, 116747–116765 (2022)
13. Setz, B., Graef, S., Ivanova, D., Tiessen, A., Aiello, M.: A comparison of open-source home automation systems. *IEEE Access*, vol. 10 (2020)
14. Morton, T.: Microcontroller systems in home automation. *IEEE Access* **8**, 21448–21468 (2010)
15. Balasubramanian, V., Morton, T.: Occupancy sensor controls in smart homes. *IEEE Trans. Autom. Sci. Eng.* **15**, 1234–1247 (2018)
16. Mangal, M., Sharma, S., Dubey, S., Koli, S., Yadav, A.: Home automation using AI-IoT. *Int. J. Res. Public. Rev.* **4**(4), 3317–3320 (2023)
17. Reddy, V.B., Balk, D., Manikyam, B., Gayatri, S., Kumar, P.S.: Home automation using artificial intelligence & internet of things. *MATEC Web Conf* **392**, 1–10 (2024)
18. Albert, K. K.: Voice controlled lighting system for the elderly and persons with special needs. Bachelor's project report, The Technical University of Kenya (2018)

19. Swain, A., Satpathy, S., Paikaray, B.K., Pramanik, J.: IoT pro-interventions: transforming industries and enhancing quality of life. In: Mohanty, S.N., Satpathy, S., Cheng, X., Pani, S.K. (eds) Explainable IoT applications: a demystification. information systems engineering and management, vol 21. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-74885-1_1
20. Swain, P.K., Pattnaik, L.M., Satpathy, S.: IoT applications and cyber threats: mitigation strategies for a secure future. In: Mohanty, S.N., Satpathy, S., Cheng, X., Pani, S.K. (eds) Explainable IoT applications: a demystification. information systems engineering and management, vol 21. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-74885-1_27
21. Mohanty, S. N., Chatterjee, J.M, Satpathy, S.: Internet of things and its applications part of book series EAI/springer innovations in communication and computing (2022). ISBN: 978-3-030-77527-8

Intelligent Intrusion Detection: A Deep BiLSTM Approach Empowered by Hybrid Spider-Coyote Optimization for IIOT Security



Sushama L. Pawar and Mandar S. Karyakarte

Abstract The increasing use of Industrial Internet of Things (IIOT) devices changed the industrial setting in terms of connectivity and automation, which helps in productivity growth, but as the network expands, it becomes more vulnerable to various security threats. To protect IIOT networks from these threats, intrusion detection systems (IDS) become essential. IDS plays a vital role in identifying malicious activities and threats. In this study, we proposed an innovative approach to identify intrusion in IIoT networks by integrating deep learning with a hybrid optimization technique; the proposed approach combines the Bidirectional Long Short Term Memory (BiLSTM) model, which harnesses the ability to capture and learn sequential data precisely. Model training is done incrementally, aiming to adapt to the evolving nature of IIOT data, resulting in enhanced accuracy. We have used Spider-Coyote Optimization, which combines Spider Monkey Optimization (SMO) and Coyote Optimization algorithms (COA) to fine-tune network parameters, speeding up the detection process and providing accurate intrusion detection performance. The proposed model is evaluated on the used NSL-KDD dataset, achieving an accuracy of 99.82% for binary classification and 99.76% for multiclass classification. The result analysis suggests that the proposed incremental learning model surpasses the traditional IDS methods.

Keywords Industrial Internet of Things (IIOT) · Spider Monkey Optimization (SMO) · Coyote Optimization Algorithm (COA) · Bidirectional Long Short-Term Memory (BiLSTM)

These authors contributed equally to this work.

S. L. Pawar (✉) · M. S. Karyakarte
Department of Computer Engineering, BRAC's Vishwakarma Institute of Information Technology, Pune, Maharashtra, India
e-mail: pawarsushma23@gmail.com

M. S. Karyakarte
e-mail: mandar.karyakarte@viit.ac.in

1 Introduction

Industrial operations have entirely changed in recent years because of IIoT devices. Network connectivity and automation increased the productivity of various industries. Although as the network grows and devices become more interconnected, they can be vulnerable to cyber-attacks, which can cause significant loss in terms of infrastructure and financial, so the security within IIoT network becomes a top priority, IDS plays a vital role in detecting and stopping harmful threats before they cause damage to IIoT devices and disrupt entire IIoT infrastructure. Deep learning models such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short Term Memory (LSTM) networks have been adopted recently in various IDS. These models possess the unique ability to analyze network traffic and data patterns and identify malicious activities. These models can handle long sequential data, making them perfect for analyzing time-series patterns in IIoT network traffic. They can capture temporal dependencies by retaining memory of previous inputs, which is essential for detecting patterns in continuous data streams.

The architecture of RNNs supports real-time processing, which is essential for IIoT systems where immediate threat detection is necessary. Besides, variants like Gated Recurrent Units (GRUs) enhance efficiency by addressing vanishing gradient issues, making RNNs more suitable for long sequences. Standard RNNs may struggle with long-term dependencies, leading to degraded performance in analyzing extended sequences. CNNs are mainly used in image recognition. However, they have been adapted to handle intrusion detection by treating network traffic data as structured input (e.g., transformed into 2D feature maps or matrices). CNNs not only excel in automatically extracting features from raw input data and learning hierarchical features such as packet headers and payloads but also detect spatial patterns in network traffic, making them effective in identifying specific attack signatures or anomalies. Compared to RNNs, CNNs are often computationally more efficient, enabling their use in near-real-time scenarios. By recognizing spatial patterns in network flow, CNNs can detect distributed denial-of-service (DDoS) attacks effectively. CNNs are ineffective in capturing temporal dependencies, as they focus on spatial relationships. Also, there is a data transformation overhead when transforming sequential data into formats suitable for CNNs. LSTMs are introduced to address the limitation of the vanishing gradient problem of traditional RNNs. Their ability to model short-term and long-term dependencies in time-series data makes them well-suited for IIoT intrusion detection. With continuous training, LSTMs can detect new or evolving threats and adapt to the changes in network behavior. RNNs, CNNs, and LSTMs can bring unique advantages for intrusion detection. RNNs and LSTMs excel in temporal analysis, and CNNs are better at extracting spatial features from the data. Integrating two models can make a great hybrid model that offers enhanced performance by leveraging their unique strengths and fine-tuning. These models can provide robust and scalable security solutions.

This study proposes a novel framework integrating deep learning models with advanced hybrid optimization techniques. The main module of our approach is the

BiLSTM neural network, which is selected for its ability to handle sequential data and capture complex temporal patterns effectively. The BiLSTM model goes through incremental training to adapt to the dynamic nature of IoT data streams and evolving threats. To further improve accuracy, we adapted two optimization algorithms, SMO and COA, which efficiently tune BiLSTM network parameters, achieving a faster detection rate. We have used NSL-KDD datasets for rigorous evaluation of the proposed model, demonstrating high performance in key metrics such as accuracy, FPR, and adaptability. We offer a robust and scalable solution for intrusion detection for IoT networks.

1.1 Contributions

This study proposes a BiLSTM model that harnesses the capability of incremental learning to handle the evolving nature of IoT data streams, which ensures a robust intrusion detection system that can detect new intrusion patterns.

Hybrid Spider-Coyote Optimization is presented to enhance parameter tuning efficiency for the BiLSTM model, resulting in a faster detection rate.

Integrating the proposed hybrid optimization algorithm demonstrates superior performance and detects anomalies with higher accuracy than existing methods.

The proposed system is evaluated on the NSL-KDD dataset, which includes diverse network traffic and attacks. Evaluation results validate the system's ability to adapt and detect various threats. Incremental learning ensures the adaptability of the proposed model, enabling it to respond to new threats.

2 Existing Work

The CRSF framework is proposed in [1]. The framework integrates multiple techniques to enhance intrusion detection. A dimension transformation function is used to process input data into two-dimensional images. This processed data is then fed to convolutional kernels to extract spatial features and RNNs to capture temporal features. An SVM classifier maps the features into a high-dimensional space, resulting in precise differentiation between classes. The ToN_IoT dataset is used for performance evaluation.

A lightweight dense random neural network (DnRaNN) is introduced in [2] for Resource-Constrained IoT Networks. The model is explicitly optimized for resource-limited environments evaluated on the ToN_IoT dataset. In [3], a multi-head attention-based gated recurrent unit (MAGRU) model is presented, which integrates multi-head attention for enhanced feature learning and GRU for sequential data analysis, and evaluation is done on Edge-IIoTset and MQTTset datasets. In [4], a hybrid deep random neural network (HdRaNN) is presented, integrating deep random neural networks with multilayer perceptrons and dropout regularization—evaluated on the

DS2OS and UNSW-NB15 datasets. A Hybrid model is presented in [5], which integrates Graph Convolutional Networks and Long Short-Term Memory (GCN-LSTM) models for intrusion detection in Industrial Control Systems (ICS) and IoT networks. Reference [6] Introduced Federated Learning with Instance-Based Transfer learning for anomaly-based IDS, which addresses non-IID data issues in IIoT. The system, instance-based transfer learning, and a rank aggregation algorithm with weighted voting enhance the performance. Similarly, in [7], a Federated Learning Framework is used for Software Defined Networking (SDN)-based IoT network. Local training is conducted on each security gateway server to preserve privacy while ensuring high accuracy in anomaly detection. In [8], the CNN-LSTM model is presented, which is trained for binary and multiclass intrusion detection using the Edge-IIoTset dataset. The model effectively identifies various types of attacks, including DoS/DDoS and injection attacks. Multiple IDS models were implemented in [9], which used six algorithms, namely Random Forest (RF), Decision Tree (DT), K-nearest neighbor (KNN), Support Vector Machines (SVM), Logistic Regression (LR), and Naïve Bayes (NB) among which Random Forest achieved the highest accuracy. The WUSTL-IIoT-2021 dataset is used for performance evaluation. EvolCostDeep and DeepIDSFog are introduced in [10]. EvolCostDeep is a hybrid model combining stacked autoencoders and CNNs with a cost-dependent loss function. A DeepIDSFog framework is introduced to address scalability challenges in big IIoT data through parallelized fog computing—experiments on ToN-IoT and UNSW-NB15 datasets.

The author proposed two deep learning models in [11] to classify IIoT traffic, a regular neural network, and a recurrent neural network, using the Edge-IIoTset dataset for binary and multi-class classification. CSTF framework is introduced in [12] to address the limitations of traditional Transformers in extracting local features. The framework combines CNN, an enhanced Transformer, and SVM. Framework is trained on the IIoTD dataset. Transformer and CNN capture regional and global features from input data while reducing data feature dimensions, and decision-making optimization is done on SVM. In [13], the author proposed an SDN-based framework for IDS, which integrates SVM and Decision Tree classifiers to detect threats in industrial IoT. Framework uses the NSL-KDD dataset, the Correlation-based feature selection (CFS) algorithm is used for feature selection, and 23 uncorrelated features were selected for prediction. Dataset classification uses the Linear SVM model and quadratic SVM model, as well as Fine Tree and Medium Tree classifiers where the quadratic SVM model has the highest accuracy. A hybrid CNN-LSTM model is introduced in [14], which uses X-IIoTID and UNSW-NB15 datasets to identify abnormal patterns. Here, CNN is used to extract complex attributes, and LSTM is used for classification. A level detection approach is presented in [15] by combining LightGBM and deep learning. LightGBM is used for lower-level detection, and deep learning is used for upper-level detection. With low training time, the methodology showed high detection accuracy.

A scalable IDS for Fog Computing environment called Deep-IFS is introduced in [16], which uses a local gated recurrent unit (LocalGRU) for local feature extraction and a multi-head self-attention mechanism (MHSA) for global learning. It uses the Bot-IoT dataset for performance evaluation, which suggests it achieved better

accuracy and reduced recognition time; it effectively demonstrated the effectiveness of distributed learning in reducing computational overheads. Extremely Gradient Boosting (XGBoost) is introduced in [17] for Imbalanced IIoT Datasets; as the name suggests, it addresses the challenges of imbalanced IIoT datasets by applying the XGBoost model. The model is evaluated on the X-IIoTDS and ToN_IoT datasets, which show a significant improvement in attack detection performance compared to traditional methods. For preserving privacy and anomaly detection, a federated learning model is introduced in [18], which uses deep reinforcement learning algorithms; this approach does not require local datasets, which reduces the chances of privacy leakage. This model is suitable for various IIoT scenarios as it achieved high detection accuracy with low latency, whereas [19] also presents the Federated Learning approach with Distributed Learning and Attention Mechanism to enhance scalability and reduce communication overhead; the model uses attention mechanism in the encoder layer to enable parallel processing. It uses the Edge-IIoTset dataset to evaluate, confirming the framework's ability to scale effectively. In [20], ensemble models are introduced, which include XGBoost, Random Forest, Bagging, extra trees (ET), and AdaBoost; it uses the Chi-Square Statistical method for feature selection. For evaluation, it uses telemetry data from the ToN_IoT dataset, suggesting that XGBoost achieved the highest accuracy. A comparative analysis of existing intrusion detection models is shown in Table 1.

3 Design Goals

In this research, we aim to develop an efficient storage cost framework that enables users to store and share their data securely in untrusted cloud environments. The design of our system is intended to meet the following key objectives:

3.1 *Adaptive Intrusion Detection*

To address the dynamic and evolving nature of IIoT data streams, we propose an incremental learning-based BiLSTM model. This approach ensures that the model continuously learns from new data without requiring complete retraining, thereby reducing computational overhead and maintaining up-to-date intrusion detection capabilities. The BiLSTM architecture effectively captures the data's sequential dependencies and temporal patterns, making it highly suitable for detecting sophisticated intrusion patterns. By adapting dynamically, the system ensures consistent detection accuracy, even as IIoT networks experience changes in traffic behavior, device interactions, or emerging threats.

Table 1 Comparative analysis of existing intrusion detection models

Ref. No.	Objective	Technology used	Dataset used	Performance score
[1]	Propose CRSF framework for anomaly-based IDS	2D convolutional kernels, RNN, SVM	ToN_IoT	Accuracy—99.59%
[2]	Lightweight IDS for resource-constrained IoT networks	Dense random neural network (DnRaNN)	ToN_IoT	Accuracy—(binary 99.59%) (Multiclass 99.14%)
[3]	Enhance feature learning for IIoT intrusion detection	Multi-head attention (MA), GRU	Edge-IIoTset, MQTTset	Accuracy—99.62%
[4]	Develop HDRaNN for cyberattack detection	Deep random neural networks, MLP	DS2OS, UNSW-NB15	Accuracy-98% for DS2OS and 99% for UNSW-NB15
[5]	Intrusion detection using hybrid GCN-LSTM model	GCN, LSTM	Simulated scenarios	Accuracy—99.99%
[6]	Privacy-preserving IDS with federated learning	Instance-based transfer learning, AdaBoost, RF	Simulated IIoT	Accuracy 95.97% for AdaBoost
[7]	Federated learning framework for SDN-based IoT IDS	Federated Learning (FL)	SDN-based IoT traffic	Accuracy of FL for Syn-98.20% MSSQL-99.30% NetBios-99.94%
[8]	Binary and multiclass attack detection for IIoT	CNN, LSTM	Edge-IIoTset	Accuracy- (binary 100%) (multiclass 98.69%)
[9]	Compare ML algorithms for IIoT IDS	Six ML algorithms (RF, etc.)	WUSTL-IIoT-2021	Accuracy –for RF-99.97%
[10]	Handle scalability and class imbalance in IIoT IDS	Stacked autoencoders, CNNs, fog computing	ToN-IoT, UNSW-NB15	F1-Score 95.2%
[11]	Classify IIoT traffic in binary and multiclass contexts	Regular NN, RNN	Edge-IIoTset	Accuracy—Above 99%
[12]	Design CTSF framework for Industrial Internet IDS	CNN, Enhanced Transformer, SVM	X-IPOD	Accuracy-98.88%
[13]	Develop SDN-based framework for threat detection	SVM, Decision Tree	NSLKDD	Accuracy-99.7%

(continued)

Table 1 (continued)

Ref. No.	Objective	Technology used	Dataset used	Performance score
[14]	Detect intrusions using hybrid deep learning models	CNN, LSTM, CNN + LSTM	UNSW-NB15, X-IIoTID	F1-Score for UNSW-NB15 93.00% for binary, 92.59% for multiclass F1-score for X-IIoTID 99.60% for binary and 99.54% for multi-class
[15]	Enhance detection with lightweight and DL algorithms	LightGBM, DL	—	High
[16]	Scalable IDS using distributed learning in fog networks	LocalGRU, MHSA distributed training	Bot-IoT	Accuracy-99.7%
[17]	IDS for imbalanced IIoT datasets	XGBoost	X-IIoTDS, ToN_IoT	(F1-Score 99.9% for X-IIoTDS and 99.87% for TON_IoT)
[18]	Reliable anomaly detection using federated learning	Federated learning, deep reinforcement learning	—	—
[19]	Scalable IDS via FL for IIoT	Federated Learning, Attention mechanism	Edge-IIoT	Accuracy-94.60% for Centralized Learning, 95.70% for Federated Learning
[20]	Intrusion detection using ensemble models	Feature selection (Chi-Square), XGBoost, Bagging, ET, RF, AdaBoost	ToN_IoT	Accuracy for Bagging-99.0%, GBoost-100%, RF-99.0%, ET-99.0%, Ada-70.0%
[21]	Detect malicious behavior in IIoT networks using TCP/IP data	DAE-DFNN architecture with hybrid rule-based design	NSL-KDD, UNSW-NB15	98.0%
[22]	Lightweight IDS for high-load IIoT networks	Pearson Correlation Coefficient (PCC), KNN, Random Forest (RF)	ToN-IoT	Over 99%

(continued)

Table 1 (continued)

Ref. No.	Objective	Technology used	Dataset used	Performance score
[23]	Efficient anomaly detection in data streams	LSHiForest with sliding window, PCA, change detection, and model update strategies	KDDCUP99	–
[24]	Secure IDS for large-scale IIoT RPL-based environments	Genetic programming, threshold modulation	Cooja simulator (Contiki OS)	Accuracy-92%,
[25]	IDS uses a Genetic Algorithm for feature selection	GA-RF, classifiers: RF, LR, NB, DT, ET, XGB	UNSW-NB15	Accuracy 87.61% for GA-RF and 98.0% for AUC

3.2 Efficient and Robust Optimization

The Hybrid Spider-Coyote Optimization algorithm is integrated to optimize the performance of the BiLSTM model. This hybrid approach has the unique ability of SMO for global exploration and COA’s ability for local exploitation, which helps fine-tune the model’s parameters efficiently, achieving high detection accuracy and the capability to generalize well on diverse intrusion scenarios.

3.3 Scalable and Real-World Applicability

The system is designed for scalability in real-time, which can handle large amounts of data. An incremental learning approach is employed to reduce the computational overhead. The system is evaluated on the NSL-KDD dataset, which contains a diverse range of attacks, demonstrating the system’s ability to detect a wide range of attacks with high accuracy.

4 Proposed Scheme

As depicted in Fig. 1, the proposed model consists of five components: Data Pre-Processing, Data Optimization, Model Training and Classification.

The proposed model is designed to capture data, extract features, hybrid optimization, and adaptive learning. The system first aggregates raw network traffic data summarizes it into a dataset and then extracts key features for analysis. These features

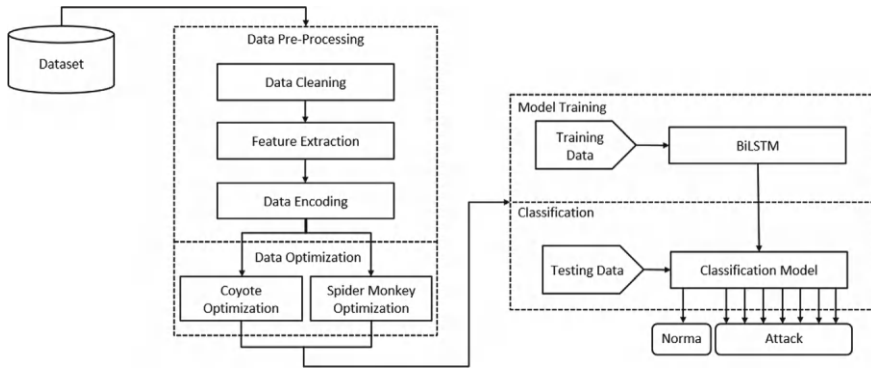


Fig. 1 Proposed hybrid BiLSTM model

are passed to the BiLSTM model to train the model and classify network traffic as usual or malicious. An adaptive drift detection mechanism is employed to fine-tune the BiLSTM model further to ensure adaptability to evolving threats.

4.1 Data Capture and Aggregation

Let dataset $D = \{d_1, d_2, \dots, d_n\}$, where each captured raw data packet in the network traffic is represented by d_i .

Each packet d_i Contains a sequence of features representing attributes like source IP, destination IP, protocol, etc.

To reduce the redundant data by retaining relevant information for analysis, an aggregation function A is applied (1).

$$A(D) = D_{ag} \quad (1)$$

where D_{ag} Represents the aggregated dataset.

4.2 Feature Extraction

From each data packet $d_i \in D_{ag}$, feature extraction is performed to derive a feature vector representing the packet.

Let $F(d_i)$ Denote the feature extraction function applied to each packet (2):

$$F(d_i) = \{f_1, f_2, \dots, f_k\} \quad (2)$$

where $\{f_1, f_2, \dots, f_k\}$ Are the extracted features such as packet length, packet interval, source–destination pair, and protocol type, and k It is the total number of features. The result of this step is a feature set (3)

$$F = \{F(d_1), F(d_2), \dots, F(d_n)\} \quad (3)$$

4.3 Hybrid Optimization: Coyote and Spider Monkey Optimization

A hybrid optimization approach, combining COA and SMO, is applied to improve the feature set for better classification.

4.3.1 Coyote Optimization Algorithm (COA)

Coyote Position Representation: Let $X_{i,j}$ Represent the position of the i th coyote in the j th dimension (feature). The coyote population is initialized as (4):

$$X_{i,j} \in R^k, i = 1, 2, \dots, N_c, j = 1, 2, \dots, k \quad (4)$$

N_c is the number of coyotes, and k is the number of features.

Coyote Social Adaptation: Coyotes update their positions based on the social tendency of the pack (5):

$$X_{i,j}^{t+1} = X_{i,j}^t + r_1(X_{best,j} - X_{i,j}^t) + r_2(X_{rand,j} - X_{i,j}^t) \quad (5)$$

where $X_{best,j}$ Is the best position, $X_{rand,j}$ is a random coyote, and r_1, r_2 Are random values in $[0,1]$.

4.3.2 Spider Monkey Optimization (SMO)

Monkey Position Update: Similarly, spider monkeys update their positions based on their local and global search tendencies (6):

$$Y_{i,j}^{t+1} = Y_{i,j}^t + \alpha(Y_{local,j} - Y_{i,j}^t) + \beta(Y_{global,j} - Y_{i,j}^t) \quad (6)$$

where $Y_{local,j}$ and $Y_{global,j}$ Represent local and global optimum positions and α, β Are control parameters.

The combined optimized feature set, F_{opt} It is obtained from both optimization steps and is represented as (7):

$$F_{opt} = O(F) = HybridOptimization(F) \quad (7)$$

where $O(F)$ Represents the output of the hybrid COA-SMO optimization.

4.4 Data Classification Using BiLSTM

After optimization, the feature set F_{opt} It is fed into a BiLSTM model for classification. BiLSTM incorporates two LSTM layers processing input from both forward and backward directions to capture temporal dependencies.

Let $h_t^{(f)}$ and $h_t^{(b)}$ Represent the hidden states of the forward and backward LSTMs at time step t . For each input vector x_t . In the sequence, the forward and backward LSTMs compute (8), (9):

$$h_t^{(f)} = LSTM^{(f)}(x_t, h_{t-1}^{(f)}) \quad (8)$$

$$h_t^{(b)} = LSTM^{(b)}(x_t, h_{t+1}^{(b)}) \quad (9)$$

The output at time t for the BiLSTM layer is then the concatenation of these two hidden states (10):

$$ht = [ht(f); ht(b)] \quad (10)$$

The BiLSTM then applies a classification function. C On the final hidden states to predict the class labels (normal or attack) (11):

$$C(h) = softmax(Wh + b) \quad (11)$$

where W and b Are weight and bias terms, and the output is a probability distribution over the classes.

4.5 Adaptive Drift Detection and Fine Tuning

To maintain classification accuracy over time, an Adaptive Drift Detection mechanism monitors changes in data distribution, triggering model fine-tuning when needed.

Drift Detection: Let E_{new} and E_{old} Represent the error rates on recent and historical data. Drift is detected if (12):

$$|E_{new} - E_{old}| > \delta \quad (12)$$

where δ is a threshold indicating a significant drift.

Fine Tuning: If drift is detected, the model parameters θ These are adjusted by re-optimizing or retraining the BiLSTM based on recent data. Let θ_{t+1} Be the updated parameters (13):

$$\theta_{t+1} = \theta_t + \eta \nabla \text{Loss}(C(h), y) \quad (13)$$

where η is the learning rate, and ∇Loss is the gradient of the loss function concerning model parameters.

5 Experimental Settings

To evaluate the performance of the proposed IDS, we conducted experiments using the NSL-KDD dataset. These benchmark datasets in network intrusion detection contain labeled network traffic instances representing normal and malicious activities.

NSL-KDD: This dataset is an updated version of the KDD Cup 99 dataset, optimized to address issues like redundancy and class imbalance. Table 2 shows 41 features per record, categorized into four attack types: DoS, Probe, U2R, and R2L, along with regular traffic, as shown in Table 3.

Key benefits of the NSL-KDD dataset include

The training set is free from duplicate entries, ensuring unbiased classification outcomes.

The absence of duplicate records in the test set enhances reduction rates, leading to more reliable evaluations.

6 Performance Analysis

As illustrated in Table 4, the model performs consistently across training, validation, and test datasets for binary and multi-class classification, demonstrating its robustness and reliability. The lower test loss compared to training and validation loss suggests that the model might slightly underfit the training data, which can be advantageous for avoiding overfitting.

Multi-class classification metrics are comparable to binary classification metrics, indicating that the proposed system effectively handles the complexity of distinguishing between multiple attack types.

Table 5 demonstrates the exceptional performance of the BiLSTM + hybrid Spider-Coyote Optimization system. With accuracy exceeding 99.7% and extremely low loss values, the system exhibits robustness, adaptability, and precision in binary

Table 2 Features of NSL-KDD dataset

No.	Feature name	No.	Feature name
1	Duration	21	Is_hot_login
2	Protocol_type	22	Is_guest_login
3	Service	23	Count
4	Flag	24	Srv_count
5	Src_bytes	25	Error_rate
6	Dst_bytes	26	Srv_error_rate
7	Land	27	Error_rate
8	Wrong_fragment	28	Srv_error_rate
9	Urgent	29	Same_srv_rate
10	Host	30	Diff_srv_rate
11	Num_failed_logins	31	Srv_diff_host_rate
12	Logged_in	32	Dst_host_count
13	Num_compromised	33	Dst_host_srv_count
14	Root_shell	34	Dst_host_same_srv_rate
15	Su_attempted	35	Dst_host_diff_srv_rate
16	Num_root	36	Dst_host_same_src_port_rate
17	Num_file_creations	37	Dst_host_srv_diff_host_rate
18	Num_shells	38	Dst_host_error_rate
19	Num_access_files	39	Dst_host_srv_error_rate
20	Num_outbonds_cmds	40	Dst_host_error_rate
		41	Dst_host_error_rate

Table 3 Attack types and associated labels in NSL-KDD

Attack types	Output labels
Normal = 0	Normal = 0
DOS = 1	Back, land, Neptune, pod, smurf, teardrop
Probing = 2	Ipsweep, nmap, portsweep, satan
R2L = 3	Ftp_write, guess_passwd, imap, multihop, phf spy, warezclient, warezmaster
U2R = 4	Buffer_overflow, loadmodule, perl, rootkit

and multi-class intrusion detection tasks. These results affirm its potential for real-world applications in IIoT environments, where accurate and efficient detection of diverse attack types is critical.

In the proposed graph, the accuracy of four different intrusion detection models: DnRaNN, MAGRU, HDRaNN, and Hybrid BiLSTM. Each model is represented

Table 4 Results obtained by BiLSTM

Metrics	Binary classification (%)	Multi-class classification (%)
Accuracy in Training	98.13	98.10
Accuracy in Validation	98.09	98.14
Accuracy in Test	98.15	98.11
Loss in Training	09.62	09.62
Loss in Validation	09.51	09.48
Loss in Test	04.65	04.65

Table 5 Results obtained by BiLSTM + hybrid Spider-Coyote optimization

Metrics	Binary classification (%)	Multi-class classification (%)
Accuracy in training	99.82	99.76
Accuracy in validation	99.76	99.76
Accuracy in test	99.85	99.81
Loss in training	0.12	0.20
Loss in validation	0.15	0.16
Loss in test	0.13	0.13

by a distinct color, with accuracy values shown on the vertical axis as percentages and the models listed on the horizontal axis. All four models achieve accuracy scores close to 100%, indicating their strong performance in detecting intrusions in the context they were tested, as shown in Fig. 2. DnRaNN, MAGRU, and HDRaNN models show very high accuracy, as do Hybrid BiLSTM models. However, the exact values aren't labeled. This consistently high accuracy across different architectures demonstrates the effectiveness of various neural network approaches (including Dense Random Neural Networks, Multi-head Attention Gated Recurrent Units, Hybrid Deep Random Neural Networks, and Hybrid BiLSTMs) in achieving reliable intrusion detection for IoT and IIoT security applications. However, without additional metrics like F1-score, as shown in Fig. 3 or AUC, it's hard to determine which model might best balance precision and recall. This graph effectively illustrates the models' strengths in terms of accuracy, highlighting their suitability for high-accuracy applications in network security.

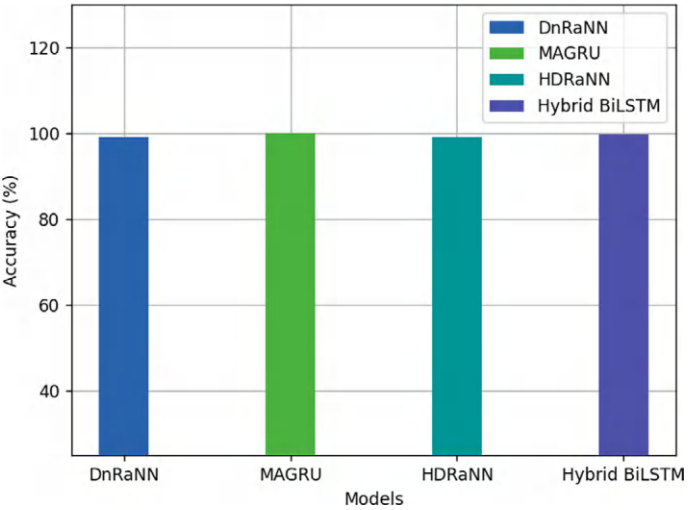


Fig. 2 Accuracy of existing models with hybrid BiLSTM

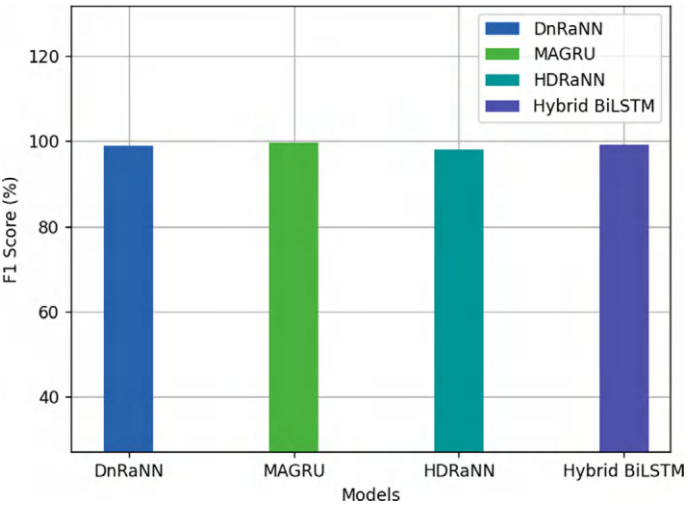


Fig. 3 F1 score of existing models with hybrid BiLSTM

7 Conclusion

This study proposed an advanced IDS framework integrating multiple novel approaches, which include incremental learning, hybrid optimization, and adaptive drift detection, which addressed the unique challenges in evolving IIoT networks. Hybrid optimization algorithms are used to fine-tune hyperparameters of the BiLSTM

model and provide optimal solutions. The BiLSTM model processes sequential data by leveraging forward and backward dependencies, enabling the system to handle complex, high-dimensional data. The adaptive Drift Detection mechanism ensures that the model continuously learns from new data without requiring complete retraining, thereby reducing computational overhead and maintaining up-to-date intrusion detection capabilities, which makes the system robust to evolving threats. The system is evaluated on the NSL-KDD dataset, achieving an accuracy of 99.82% for binary classification and 99.76 for multiclass classification.

The proposed system can be extended by leveraging advanced deep learning models such as transformers to improve classification accuracy further and integrating blockchain technology to enhance data integrity.

8 Conflicts of Interest:

The corresponding author and all co-authors confirm no conflicts of interest to declare.

Funding: There is no funding.

Data Availability: The data supporting this study can be available upon request.

References

1. Li, S., et al.: CRSF: an intrusion detection framework for industrial internet of things based on pretrained CNN2D-RNN and SVM. *IEEE Access* **11**, 92041–92054 (2023). <https://doi.org/10.1109/ACCESS.2023.3307429>
2. Latif, S., et al.: Intrusion detection framework for the internet of things using a dense random neural network. *IEEE Trans. Industr. Inf.* **18**(9), 6435–6444 (2022). <https://doi.org/10.1109/TII.2021.3130248>
3. Ullah, S., Boulila, W., Koubâa, A., Ahmad, J.: MAGRU-IDS: a multi-head attention-based gated recurrent unit for intrusion detection in IIoT networks. *IEEE Access* **11**, 114590–114601 (2023). <https://doi.org/10.1109/ACCESS.2023.3324657>
4. Huma, Z.E., et al.: A hybrid deep random neural network for cyberattack detection in the industrial internet of things. *IEEE Access* **9**, 55595–55605 (2021). <https://doi.org/10.1109/ACCESS.2021.3071766>
5. Koca, M., Avci, I.: A novel hybrid model detection of security vulnerabilities in industrial control systems and IoT using GCN+LSTM. *IEEE Access* **12**, 143343–143351 (2024). <https://doi.org/10.1109/ACCESS.2024.3466391>
6. Zhang, J., Luo, C., Carpenter, M., Min, G.: Federated learning for distributed IIoT intrusion detection using transfer approaches. *IEEE Trans. Industr. Inf.* **19**(7), 8159–8169 (2023). <https://doi.org/10.1109/TII.2022.3216575>
7. Duy, P.T., Hung, T.V., Ha, N.H., Hoang, H.D., Pham, V.-H.: Federated learning-based intrusion detection in SDN-enabled IIoT networks. In: 2021 8th NAFOSTED Conference on Information and Computer Science (NICS), Hanoi, Vietnam, pp. 424–429 (2021). <https://doi.org/10.1109/NICS54270.2021.9701525>

8. Khacha, A., Saadouni, R., Harbi, Y., Aliouat, Z.: Hybrid deep learning-based intrusion detection system for industrial internet of things. In: 2022 5th International Symposium on Informatics and its Applications (ISIA), M'sila, Algeria, pp. 1–6 (2022). <https://doi.org/10.1109/ISIA55826.2022.9993487>
9. Eid, A.M., Nassif, A.B., Soudan, B., Injadat, M.N.: IIoT network intrusion detection using machine learning. In: 2023 6th International Conference on Intelligent Robotics and Control Engineering (IRCE), Jilin, China, pp. 196–201 (2023). <https://doi.org/10.1109/IRCE59430.2023.10255088>
10. Telikani, A., Shen, J., Yang, J., Wang, P.: Industrial IoT intrusion detection via evolutionary cost-sensitive learning and fog computing. *IEEE Internet Things J.* **9**(22), 23260–23271 (2022). <https://doi.org/10.1109/JIOT.2022.3188224>
11. Cheikhrouhou, O., Fredj, O.B., Atitallah, N., Hellal, S.: Intrusion detection in industrial IoT. In: 2022 15th International Conference on Security of Information and Networks (SIN), Sousse, Tunisia, pp. 01–04 (2022). <https://doi.org/10.1109/SIN56466.2022.9970535>
12. Chai, G., Li, S., Yang, Y., Zhou, G., Wang, Y.: CTSF: an intrusion detection framework for industrial internet based on enhanced feature extraction and decision optimization approach. *Sensors* **23**, 8793 (2023). <https://doi.org/10.3390/s23218793>
13. Alshahrani, H., Khan, A., Rizwan, M., Reshan, M.S.A., Sulaiman, A., Shaikh, A.: Intrusion detection framework for industrial internet of things using software defined network. *Sustainability* **15**, 9001 (2023). <https://doi.org/10.3390/su15119001>
14. Altunay, H.C., Albayrak, Z.: A hybrid CNN+LSTM-based intrusion detection system for industrial IoT networks. *Eng. Sci. Technol. Int. J.* **38**, 101322 (2023). <https://doi.org/10.1016/j.jestech.2022.101322>, ISSN 2215-0986
15. Yao, H., Gao, P., Zhang, P., Wang, J., Jiang, C., Lu, L.: Hybrid intrusion detection system for edge-based IIoT relying on machine-learning-aided detection. *IEEE Netw.* **33**(5), 75–81 (2019). <https://doi.org/10.1109/MNET.001.1800479>
16. Abdel-Basset, M., Chang, V., Hawash, H., Chakraborty, R.K., Ryan, M.: Deep-IFS: intrusion detection approach for industrial internet of things traffic in fog environment. *IEEE Trans. Industr. Inf.* **17**(11), 7704–7715 (2021). <https://doi.org/10.1109/TII.2020.3025755>
17. Le, T.-T.-H., Oktian, Y.E., Kim, H.: XGBoost for imbalanced multiclass classification-based industrial internet of things intrusion detection systems. *Sustainability* **14**, 8707 (2022). <https://doi.org/10.3390/su14148707>
18. Wang, X., et al.: Toward accurate anomaly detection in industrial internet of things using hierarchical federated learning. *IEEE Internet Things J.* **9**(10), 7110–7119 (2022). <https://doi.org/10.1109/JIOT.2021.3074382>
19. Nuaimi, M., Fourati, L.C., Ben Hamed, B.: A scalable intrusion detection approach for industrial internet of things based on federated learning and attention mechanism. In: 2023 IEEE Symposium on Computers and Communications (ISCC), Gammarth, Tunisia, pp. 1–4 (2023). <https://doi.org/10.1109/ISCC58397.2023.10218054>
20. Awotunde, J.B., Folorunso, S.O., Imoize, A.L., Odunuga, J.O., Lee, C.-C., Li, C.-T., Do, D.-T.: An ensemble tree-based model for intrusion detection in industrial internet of things networks. *Appl. Sci.* **13**, 2479 (2023). <https://doi.org/10.3390/app13042479>
21. Potnurwar, A.V., Bongirwar, V.K., Ajani, S., Shelke, N., Dhone, M., Parati, N.: Deep learning-based rule-based feature selection for intrusion detection in industrial internet of things networks. *Int. J. Intell. Syst. Appl. Eng.* **11**(10s), 23–35 (2023). <https://ijisae.org/index.php/IJISAE/article/view/3231>
22. Chuang, H.-Y., Chen, R.-M.: Detection of attacks on industrial internet of things using fewer features. In: 2023 sixth international symposium on computer, consumer and control (IS3C), Taichung, Taiwan, pp. 1–4. <https://doi.org/10.1109/IS3C57901.2023.00009>
23. Yang, Y., et al.: ASTREAM: data-stream-driven scalable anomaly detection with accuracy guarantee in IIoT environment. *IEEE Trans. Netw. Sci. Eng.* **10**(5), 3007–3016 (2023). <https://doi.org/10.1109/TNSE.2022.3157730>
24. Qureshi, K.N., Rana, S.S., Ahmed, A., Jeon, G.: A novel and secure attacks detection framework for smart cities industrial internet of things. *Sustain. Cities Soc.* **61**, 102343. <https://doi.org/10.1016/j.scs.2020.102343>, ISSN 2210-6707

25. Kasongo, S.M.: An advanced intrusion detection system for IIoT based on GA and tree based algorithms. *IEEE Access* **9**, 113199–113212 (2021). <https://doi.org/10.1109/ACCESS.2021.3104113>

AI and Cryptography for Secure Communications

Enhanced Elliptic Curve Cryptography with MHOTP Key Generation and Visual Cryptography Based Image Authentication



Sachin Madhukar Kolekar and Ram Kumar Solanki

Abstract These days, people are apprehensive about their data privacy as many cyber crimes are happening. There is a need to improve all current security methods to stay significant in this tech world. Visual cryptography is a secure method for encrypting observable data, like images or text, so that a human may decode it visually without the need for a computer. This technique divides the original photo or picture into multiple shares, which individually appear as a pointless pattern. When the pieces are joined, the first image or text appears, but an OTP(one-time password) is needed while penetrating the image. This system's most important point is its simplicity, making it easy to understand and secure. This research seeks to use visual cryptography to safeguard restricted visual information containing a one-time password(OTP) and improve security. The shares are created using a k-n threshold plan, where a limited number of shares is needed to reorganize the image, guaranteeing secure data sharing. We also explore existing cryptographic techniques, including symmetric and asymmetric cryptography, and suggest an improved system that merges visual cryptography with modern encryption techniques for improved data security. This system is unaffected by unauthorized access, as the shares alone seem like pointless patterns and cannot be decoded without proper combination. The proposed technique summarizes visual cryptography's encoding and decoding processes, featuring the simplicity, effectiveness, and real-time applications of this secure data transferal technique in the modern digital world.

Keywords Visual cryptography · Secret sharing · Elliptic curve cryptography · Embedded security algorithm

S. M. Kolekar (✉) · R. K. Solanki
School of Computer Science & Engineering, Sandip University, Nashik, India
e-mail: sachinalways24@gmail.com

R. K. Solanki
e-mail: ramkumar.solanki@sandipuniversity.edu.in

1 Introduction

Visual Cryptography (VC) is a method that transforms an image into an ambiguous format, encrypts it, and then decrypts it to reveal the initial secret picture. The method of converting a picture into another picture using an association equation to prevent recognition by unauthorized parties is known as encryption. The hidden picture is encrypted into transparent components referred to as shares that, when collected together, demonstrate the hidden image. This is derived from a topic of secret sharing presented [1]. They displayed how to divide data k into n items so that k may be accurately reconstructed from any k objects. However, even a proper grasp of $k - 1$ items demonstrates nothing regarding detailed [2, 3]. If a picture share included the individual's hands, it would appear to be a picture of unidentified noise or bad art. This is another way visual cryptography can be deceptive to the untrained eye. Hues, visual: Cryptography is the creative technique used inside the reborn color image's ambiguous format. The most fundamental and well-known color model is comprised of RGB and its set of CMY. This model is the most accurate, although one can distinguish color. Together, they are consistent with the cumulative and reduction color theories. Additive colors are measured using a unit that combines different spectrum light sources. The most popular examples of this type of measurement are screens and displays for computers, which create colored pixels by aiming lepton cannons at particles on the display to produce red, green, and blue light. Reductive hues are seen when an object's components absorb some white light wavelengths while reflecting others.

Any colored item, whether actual or imagined, absorbs certain types of light and reflections or transmits others; the wavelengths remaining in the reflected or transmitted light make up the perceived color. Red, green, and blue are the first complementary hues and stimuli for human color perception. Combining two primary colors and excluding the third creates additional RGB colors of cyan, magenta, and yellow. Inexperience and blue combine to form cyan, magenta is created when inexperience and red unite. White is made when red, green, and blue are combined in full intensity. When the intensity of all the hues in the electromagnetic range converges, white light is produced.

2 Literature Review

The information in this project is encrypted using a prominent cryptography motif. Visual encryption produces a secret message that is used to conceal the human sensory system, which fully decrypts the information in an extreme image. In this paper, the writer of the play Diffie Keys Agreements is another name for exponential key exchange. By allowing two parties who have never spoken before to interact via an open channel to provide crucial data for establishing shared confidentiality, this key provides a solution to the distribution of key issues. This straightforward technique

will give an unauthenticated key agreement. An associate degree's principal goal is indicated. Sharing information is a vital step in the key setup process. The generated key must be distributed regularly and randomly from the key house, have the same properties as a key generated in person, and be utterly unknown to anybody else [1].

Only white and black pixels make up this binary string representation of an image. Since 0 represents a white part of this subject and 1 represents a black part, the result is supposed to be provided in two parts. The protection of White and black pixels makes up this binary string representation of an image. The result is likely to be delivered via two parts since, in this subject, zero will stand in for a white part, and one will represent a black part. The security disadvantage is that just one quality note can be reached because the image is split into just two parts, and if the additional knowledge shares are discovered, the picture may be quickly decoded. After the initial subject matter, a theme with a two-out-of-n visual threshold is safe. Each pair of pixels in this theme has been separated, but the shares are getting close to the value of 'n.' As a result, the degree of security will increase. Once the D-H essential contract is in position, these systems' outcomes and safety will be reduced [2].

The relevance of this concept is that knowledge expansion is not desired. The clarity of the content won't be compromised by using this theme. By utilizing this theme, the picture is safeguarded from the most significant regulatory threats. The calculation quality is poorer because XOR procedures are used in every calculation. XOR procedures combine the shared image unit to produce the encryption footage. Noise-related native faults won't affect the final word output because picture verification maintains the overall aesthetic impression of the image. The most crucial fact about the subject is that human vision, not a secret writing device, is used to create the key writing on critical data or pictures [3].

Sharing images or understanding using a key and information or image pixels that must be conveyed using alphabetic characters is made possible by the idea of visual cryptography. T. is treated as a secret-sharing theme that can be used to reveal a private secret. The central information unit within the image was split into a few parts of shared material. The key written facet images unit is secured to the transparent papers to promote the vital image or information. Typically, this will be the mechanism to start the early phase of visual encryption [4, 5].

Visual cryptography is made possible by using the D-H agreements on secrets. The Diffie-Hellman key interface, also known as the associate in the nursing interface, was utilized. Anything that is password-based and entirely private is done through this interface. Computer users often use these interfaces. The UN agency is installing a cryptanalytic provider or applying the cryptography equation [6].

The information in this project is encrypted using a prominent cryptography motif. Visual encryption produces a secret message that is used to conceal the human sensory system, which fully decrypts the information in an extreme image. In this paper, the writer of the play Diffie Keys Agreements is another name for exponential key exchange. By allowing two parties who have never spoken before to interact via an open channel to provide crucial data for establishing shared confidentiality, this key provides a solution to the distribution of key issues. This straightforward technique will give an unauthenticated key agreement. An associate degree's principal goal is

indicated. Sharing information is a vital step in the key setup process. The generated key must be distributed regularly and randomly from the key house, have the same properties as a key generated in person, and be utterly unknown to anybody else [7].

Only white and black pixels make up this binary string representation of an image. Since 0 represents a white part of this subject and 1 represents a black part, the result is supposed to be provided in two parts. The protection of White and black pixels makes up this binary string representation of an image. The result is likely to be delivered via two parts since, in this subject, zero will stand in for a white part, and one will represent a black part. The security disadvantage is that just one quality note can be reached because the image is split into just two parts, and if the additional knowledge shares are discovered, the picture may be quickly decoded. After the initial subject matter, a theme with a two-out-of-n visual threshold is safe. Each pair of pixels in this theme has been separated, but the shares are getting close to the value of 'n.' As a result, the degree of security will increase. Once the D-H essential contract is in position, these systems' outcomes and safety will be reduced [8].

The relevance of this concept is that knowledge expansion is not desired. The clarity of the content won't be compromised by using this theme. By utilizing this theme, the picture is safeguarded from the most significant regulatory threats. The calculation quality is poorer because XOR procedures are used in every calculation. XOR procedures combine the shared image unit to produce the encryption footage. Noise-related native faults won't affect the final word output because picture verification maintains the overall aesthetic impression of the image. The most crucial fact about the subject is that human vision, not a secret writing device, is used to create the key writing on essential data or pictures [9].

Sharing images or understanding using a key and information or image pixels that must be conveyed using alphabetic characters is made possible by the idea of visual cryptography. T. is treated as a secret-sharing theme that can be used to reveal a private secret. The central information unit is initially contained within the image, which has been split into a few parts of shared material. The key written facet images unit is secured to the transparent papers to promote the vital image or information. Typically, this will be the mechanism to start the early phase of visual encryption [10].

Visual cryptography is made possible by using the D-H agreements on secrets. The Diffie-Hellman key interface, also known as the associate in the nursing interface, was utilized. Anything that is password-based and entirely private is done through this interface. Computer users often use these interfaces. The UN agency is installing a cryptanalytic provider or applying the cryptography equation. The information in this project is encrypted using a prominent cryptography motif. Visual encryption produces a secret message that is used to conceal the human sensory system, which fully decrypts the information in an extreme image. In this paper, the writer of the play Diffie Keys Agreements is another name for exponential key exchange. By allowing two parties who have never spoken before to interact via an open channel to provide crucial data for establishing shared confidentiality, this key provides a solution to the distribution of key issues. This straightforward technique will give an unauthenticated key agreement. An associate degree's principal goal is indicated.

Sharing information is a vital step in the key setup process. The generated key must be distributed regularly and randomly from the key house, have the same properties as a key generated in person, and be utterly unknown to anybody else [11].

Only white and black pixels make up this binary string representation of an image. Since 0 represents a white part of this subject and 1 represents a black part, the result is supposed to be provided in two parts. The protection of White and black pixels makes up this binary string representation of an image. The result is likely to be delivered via two parts since, in this subject, zero will stand in for a white part, and one will represent a black part. The security disadvantage is that just one quality note can be reached because the image is split into just two parts, and if the additional knowledge shares are discovered, the picture may be quickly decoded. After the initial subject matter, a theme with a two-out-of-n visual threshold is safe. Each pair of pixels in this theme has been separated, but the shares are getting close to the value of 'n.' As a result, the degree of security will increase. Once the D-H essential contract is in position, these systems' outcomes and safety will be reduced [12].

The relevance of this concept is that knowledge expansion is not desired. The clarity of the content won't be compromised by using this theme. By utilizing this theme, the picture is safeguarded from the most significant regulatory threats. The calculation quality is poorer because XOR procedures are used in every calculation. XOR procedures combine the shared image unit to produce the encryption footage. Noise-related native faults won't affect the final word output because picture verification maintains the overall aesthetic impression of the image. The most crucial fact about the subject is that human vision, not a secret writing device, is used to create the key writing on essential data or pictures [13].

Sharing images or understanding using a key and information or image pixels that must be conveyed using alphabetic characters is made possible by the idea of visual cryptography. The central information unit is initially contained within the image, which has been split into a few parts of shared material. The key written facet images unit is secured to the transparent papers to promote the vital image or information. Typically, this will be the mechanism to start the early phase of visual encryption [14].

The above Table 1 Gives a comparison between key contributions and techniques of all images and cryptography algorithms as well as advantages and disadvantages.

Table 1 Comparison between key contributions and techniques

Sr. no	Author	Year	Key contributions	Techniques/findings
1	M. Naor et al	1995	Basics of visual cryptography with 2-out-of-2 schemes	Introduced binary string representation for images using 0 for white and 1 for black. Highlighted security vulnerabilities in simple schemes
2	J. Ida Christy et al	2012	Modified visual cryptography for color images	Developed a secret color picture sharing scheme allowing consistent share sizes regardless of image color complexity
3	Himanshu Sharma et al	2011	Advanced color visual encryption	Concealed color pictures under multiple colored cover photos, improving signal-to-noise ratios and maintaining quality
4	Zhongmin Wang et al	2006	Cost-efficient colored visual encryption themes	Introduced part development to reduce color visual encryption's computational costs while maintaining clarity
5	Gopal Krishna D. Dalvi et al	2006	Secret sharing for true-color images	Combined artificial neural networks with visual encryption to maintain quality in reconstructed images
6	J. K. Mandal et al	2011	Enhanced Naor and Shamir architecture for color schemes	Developed a color visual encryption scheme without part growth to improve efficiency
7	Meera Kamath et al.	2011	Rotating Visual Cryptography Scheme (RVCS)	Encrypted four secret images into two shares; introduced correlated matrices and randomized permutation for color images
8	Bin Yu et al.	2007	Adaptive visual cryptography techniques	Reduced encrypted image sizes and improved quality using adaptive-order images
9	Yanyan Han et al.	2012	Visual encryption for color images	Split secret images into color shares; used XOR techniques to maintain size and quality during reconstruction
10	Kulvinder Kaur et al.	2012	Threshold schemes for color visual cryptography	Minimized image expansion in (n, t)-threshold schemes, ensuring reconstructed image size remained practical
11	Tzung-Her Chen et al.	2008	Visual cryptography using YCbCr color space	Improved visual quality using halftone and inverted halftone techniques for encryption
12	Tzung-Her Chen et al.	2009	Deliberate sharing in visual cryptography	Created larger shares for better quality; used Key Tables for added randomness and security
13	E. Verheul et al.	2005	Improved visual cryptography clarity using color models	Addressed insufficient black pixels in reconstructed images

(continued)

Table 1 (continued)

Sr. no	Author	Year	Key contributions	Techniques/findings
14	C. Yang et al.	2005	Threshold multiple-secret visual cryptography systems	Designed schemes meeting disclosing, minimizing, and safety requirements for variable thresholds
15	C. Chang et al.	2000	Random-grid method for enhanced visual secret sharing	Maintained high-quality share images; removed restrictions like sufficient black pixels in key images, improving encryption flexibility
16	Chin-Chen Chang et al.	2002	VCRG-GAS algorithms for optical secret sharing	Proposed novel algorithms reducing part growth while improving repair efficiency for basic and color visual cryptography
17	R. Youmaran et al.	2006	Hierarchical visual cryptography in authentication systems	Introduced visual cryptography for user signature encryption in hierarchical systems, enhancing security and user satisfaction

3 Proposed Work

3.1 Various Visual Cryptography Techniques

3.1.1 Back Propagation Network-Based Extended Visual Cryptographic Scheme

Social unit Christy. AI extended predicted Back-propagation Network victimization is a visual extension of a theme. Figure 1 shows the block diagram of Visual Encryption and decryption of two secret images, and a few cowl clips were used as theme inputs for the projected approach. Each of the three videos is the same size. The outputs produced by the encryption method consist of shares that attempt to resemble the two cover videos. The secret image is present in both shares. Furthermore, the size of the footage produced is comparable. When the two shares overlap, we tend to push the main image. The proposed method consists of four basic steps. The three footages are shrunk by half in the first stage. The three images are then edited to create halftone images. These three clips are then edited to create halftone images, and these images are shown in Fig. 2. The second color halftone image transformation phase involves extracting some proper pixel values. The key image is transformed into the two shared in the third stage, which requires cryptography. The final phase is the covert writing process, where the key picture is. The final stage is the hidden writing process, in which the key image is created by overlaying the two parts.

Fig. 1 Block diagram of visual encryption and decryption

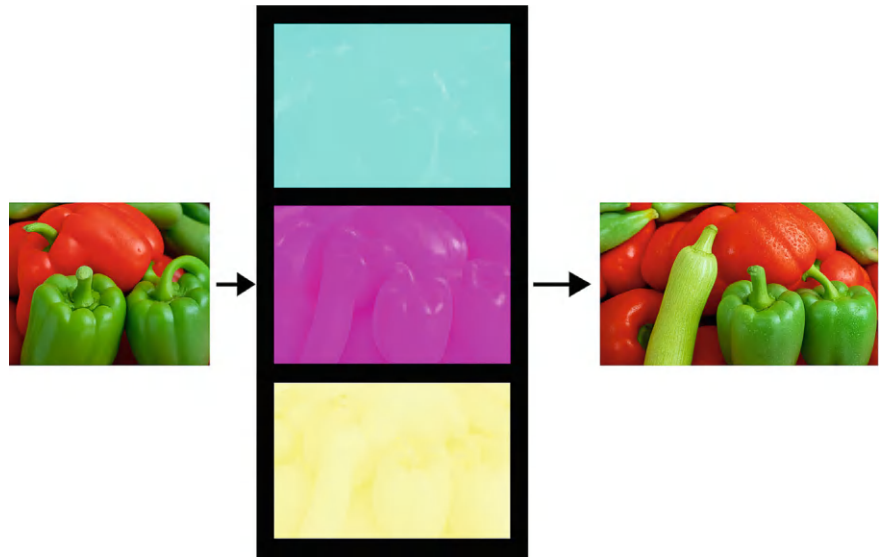
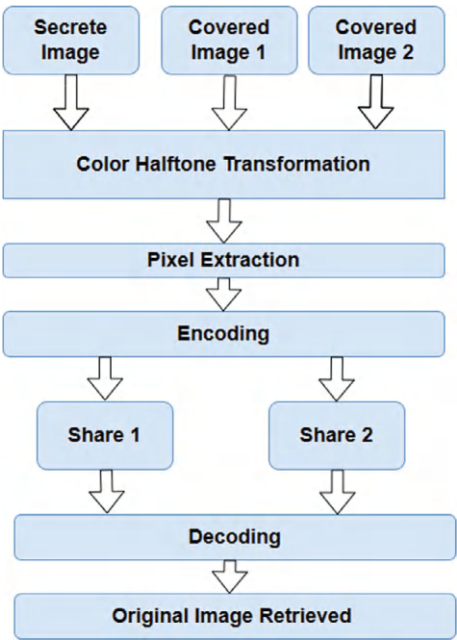


Fig. 2 Color Halftone image transformation

3.2 CA Ratio Based (2, 2) Visual Encryption and Decryption Through Meaningful Shares.

Additionally, there is a visual (2, an effort at) in this approach's field topic where secrets region units are discovered straight by stacks of two sizable offers in an optional request but with the proper arrangement. According to the proposed computational program, the developed shares are significant, have a quantitative connection, and are similar to the key image that guarantees the best housing request in terms of their components [4]. Figure 3 CARVCMS Algorithm shows the untracked secret blends seamlessly with the numerical relationship and picture element of the set of images, which predicted the subject's first advantage.

This Algorithm Has Three Steps

- Transformation for color.
- VC and creating of image parts.
- Decryption.

Explanation of the above three algorithms:

(1) Transformation for Color

The original sender repeatedly enters the four-color images CAij, CBij, CCij, CDij, and SIij, along with one unidentified image. The individual halftones IAij, IBij, ICij, IDij, and ISij are created from each picture CAij, CBij, CCij, CDij, and SIij. Additionally, NxN pixels make up the halftone film's components. Each information

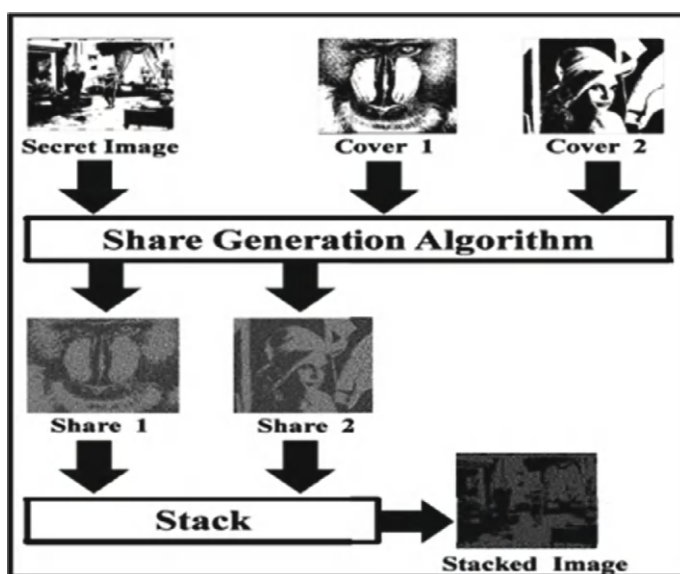
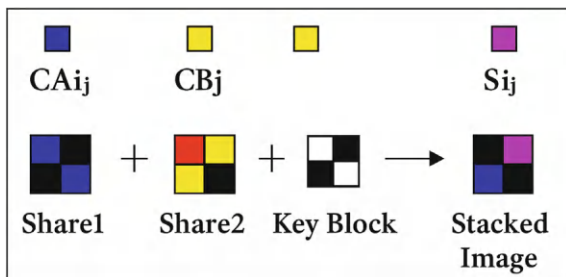


Fig. 3 Schematic diagram of CARVCMS Algorithm

Fig. 4 Decryption process

picture is divided into red, green, and blue elemental planes. Then, every aspect of those planes is treated using the halftone technique [6]. A varied halftone picture is made by connecting these three halftone planes.

(2) Decryption Process.

In the coding process, the key image is recreated by twisting a few or a group of offers together close to it. Figure 4 decryption process displays a partner degree coding example with blocks from two offers—Share1 and Share2—and the succeeding block from the Key picture. The key image's component made the block of the stacked image contain a few sub-components of constant variation, and as a result, the various subpixels are all dark. Five-hundredths of the key image is maintained within a definitive replicated image because two sub-components out of four in each block will always be of stable tone due to the element of the main image.

3.3 Visual Encryption and Decryption Using Cover Image Share Embedded Security Algorithm

System victimization with encryption-embedded privacy algorithm programmed for image sharing. Following three stages of the anticipated methodology.

Phase 1: The primary visual encryption subject identifies the first segment of the algorithmic programmer. We are prone to consider any visual cryptography method that can handle binary images. Consider the key image I first since it is converted into the halftone image S using any half-toning method, such as ordered pictures or errors diffusing [7, 8]. We anticipate creating the S_1 and S_2 components from the binary image. If we prefer to look at each share separately, each share is formed due to this. Therefore, 0.2 is meaningless.

Phase 2: Fig. 5 shows the use of complimentary pieced picture images throughout the second part's embedded sets it unique. Let C_{i1} and C_{i2} represent the quilt's supplemented information, and let C represent the quilt's image. The following step is the production of 4 stamping images (X_{i11} , X_{i12} , X_{i21} , and X_{i22}), which are then transferred to the appropriate position. Usually, to produce these components, the

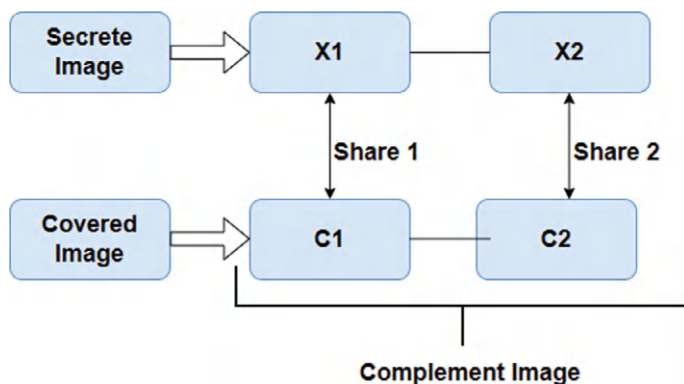


Fig. 5 Proposed architecture 1

parts $Si1$ and $Si2$ are simply superimposed over the cover image's complimentary parts, or $Ci1$ and $Ci2$ [9].

$$Xi11 = \text{STAMPING}(Si1, Ci1) \quad Xi12 = \text{STAMPING}(Si1, Ci2)$$

$$Xi21 = \text{STAMPING}(Si2, Ci1) \quad Xi22 = \text{STAMPING}(Si2, Ci2)$$

Over a simple visual encryption theme, a watermarking theme offers an additional layer of protection. Figure 6 proposed scheme structure shows an outcome of creating a canopy picture over all of the shares generally constructed on it; our proposed computational program offers an additional layer of security [10]. The result of this 0.60 is an image that was created that contains some data that was taken from a regular picture and some hidden data that was taken from a secret picture.

3.4 Visual Encryption and Decryption Based on XOR Algorithm.

A practical visual cryptography method based on XOR is proposed. It features loss-less reconfigurable algorithms, encryption, and decryption strategies for color and grayscale secret images. The encryption scheme uses a random columns selection method based on the 0-mapping and 1-mapping matrices, where the two matrix structures are generated automatically in the encoding scheme, which secures secret images by converting them to multiple familiar pictures that do not suffer from any pixel expansion problems and do not reveal any information about the secret pictures. The suggested decryption method also requires a series of XOR procedures to recreate the hidden images [11]. Compared to their original hidden photos, the rebuilt pictures do not experience any contrast distortion or pixel expansion issues. Finally, real-world examples are used to demonstrate the two suggested approaches.

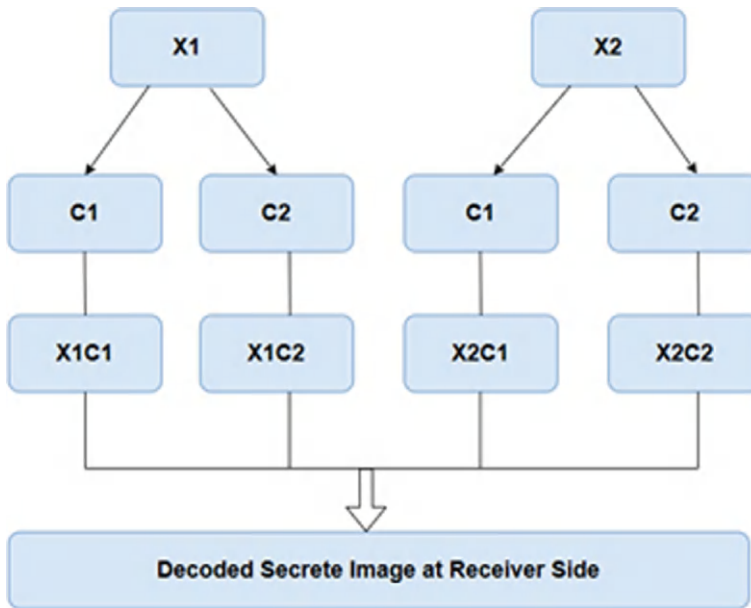


Fig. 6 Proposed scheme structure 2

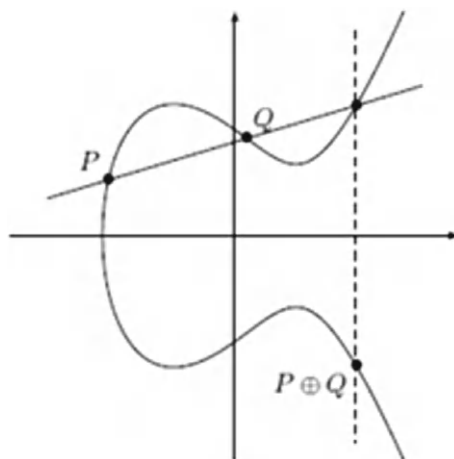
4 Elliptic Curve Cryptography (ECC)

Elliptic Curve Cryptography (ECC) is a method of public key cryptography based on the mathematics of elliptic curves. It is used to secure communications and ensure data integrity through encryption, authentication, and digital signatures. Figure 7 shows that Elliptic Curve Cryptography (ECC) is a type of public key cryptography that uses two keys: a public key that anybody can view and a private key that is kept hidden [12]. It is based on the mathematical properties of elliptic curves, which are specified by equations that generate specific curves, allowing for complex computations. One of ECC's most significant benefits is its ability to provide strong safety with smaller key sizes; for example, a 256-bit key in ECC is as secure as a 3072-bit key in traditional approaches such as RSA. ECC's efficiency makes it suitable for devices with limited computing power, such as smartphones and IoT devices[13]. ECC is frequently used in secure communication protocols like SSL/TLS, the creation of digital signatures to verify message authenticity, and cryptocurrencies like Bitcoin to protect the transaction.

How Does Elliptic Curve Cryptography (ECC) Work?

Some mathematical function defines Elliptic Curve,

$$y^2 = x^3 + ax + b \dots \text{(equation of degree 3)}$$

Fig. 7 Elliptic curve

Symmetric to the x-axis, if we draw a line, it will touch the graph a maximum of 3 points. Let $E_p(a,b)$ be the elliptic curve; consider this equation, $Q = KP$. Where $Q, P \rightarrow$ points on the curve and $k < n$. If k and $p \rightarrow$ given, it will be easy to find Q . But, if we know Q and P , it should be tough to find k [14]. This is called the discrete logarithm problem for elliptic curves.

Advantages of Elliptic Curve Cryptography (ECC)

1. Strong Security with Smaller Keys:-

ECC provides a high level of security while using much smaller key sizes than traditional methods such as RSA, making it suitable for several applications [15].

2. Faster Computation:-

ECC improves encryption, decryption, and digital signature procedures, which improves overall performance, particularly in resource-constrained situations.

3. Lower Resource Consumption:-

Because of its efficiency, ECC uses less memory and bandwidth, making it appropriate for smartphones, IoT devices, and embedded systems [16].

5 MHOTP Key Generation

OTP-one one-time password: OTP gives more security to your data [17]. This is highly secured due to a unique password valid for only one login session, and replayed attempts make it invalid or unauthorized.

One-Time Password (OTP) Process:

1. OTP is generated using OTP generated with the secured algorithm.
2. OTP encryption helps encrypt the OTP and embeds it into the corresponding share using the most suitable encryption technique.

3. OTP storage stores the generated OTPs for a bit to verify against the user, which helps secure the data from unauthorized access.
4. The component distributes the encrypted shares to authorized users.
5. The authorized parties can now retrieve the OTP associated with their shares.
6. Using appropriate algorithms, the OTP is decrypted
7. The decrypted OTP is used to decrypt the encrypted corresponding shares.
8. The decrypted images together to reveal the secret image.

6 Result and Discussion

Table 2 shows each additional component's encryption and decryption times [18]. Baseline ECC is the fastest, while adding MHOTP and visual cryptography increases the time due to extra processing for key generation and image sharing.

- a. **ECC versus RSA:** ECC consistently outperforms RSA regarding encryption and decryption times due to smaller key sizes and lower computational demands.
- b. **Visual Cryptography Impact:** Adding visual cryptography, especially for color schemes or higher threshold setups, increases encryption and decryption times[19].
2. **3.AES-256 Performance:** AES-256 offers an efficient alternative, maintaining competitive speeds even when combined with visual cryptography.
- a. **Configuration with MHOTP:** Introducing MHOTP enhances security but marginally increases processing times[20].

7 Conclusion

In this paper, we have discussed how to encrypt and decrypt a color-correct different image using Elliptic Curve Cryptography. The culmination of all the algorithms to be used has a vast scope in the future. Using the mathematical properties of the elliptical curve and the method's efficiency is the perfect way to apply it to devices with limited processing power. The pixels of the images are encrypted, and the generated key will be protected by the very same ECC, which means that even if intercepted, without the key, the attacker cannot decrypt the image. This will allow secure transmission of the pictures and the possibility of applying digital watermarks. On top of this, using the OTP (one-time password) system enables us to add an extra layer of security to the already existing encryption. This reduces the risk of unauthorized access to sensitive images.

Table 2: Encryption and decryption performance analysis

Sr. no	Configuration	Encryption time (ms)	Decryption time (ms)	Remarks/analysis
1	Baseline ECC	25 ms	20 ms	Standard elliptic curve cryptography (ECC) setup with minimal overhead
2	ECC with MHOTP	35 ms	30 ms	Adding a Multi-factor Hash-based One-Time Password (MHOTP) increases processing time slightly
3	ECC with MHOTP + Visual Cryptography	50 ms	45 ms	Integration with visual cryptography increases computational and rendering complexity
4	Baseline RSA (1024-bit)	40 ms	35 ms	RSA is slower than ECC for encryption and decryption due to larger key sizes
5	ECC with 2-out-of-2 Visual Cryptography	55 ms	50 ms	Shares generated for 2-out-of-2 schemes add slight overhead to encryption and decryption
6	ECC with 3-out-of-5 Visual Cryptography	70 ms	65 ms	Higher thresholds require generating more shares, significantly increasing processing time
7	ECC with Color Visual Cryptography	90 ms	80 ms	Color schemes add computational complexity for RGB channel encryption and stacking
8	RSA with MHOTP	60 ms	50 ms	RSA combined with MHOTP performs slower than ECC due to computationally intensive operations
9	AES-256 with Visual Cryptography	45 ms	40 ms	AES-256 encryption is faster than ECC, even when combined with visual cryptography
10	ECC with Random Grid Visual Cryptography	60 ms	55 ms	Random-grid methods reduce computational overhead while maintaining security

References

1. Naor, M., Shamir, A.: Visual cryptography. *Advances in Cryptology—EUROCRYPT'94*. Springer-Verlag, Vol-950, pp.1–12 (1995)
2. Ida Christy, J., Seenivasagam, V.: Construction of color extended visual cryptographic scheme using back propagation network for color images. In: *International conference on computing, electronics and electrical technologies [IC CEET]* 978-1-4673-0210-4112 © IEEE, pp 88–93 (2012)
3. Sharma, H., Kumar, N., Jha, G. K.: Enhancement of security in Visual Cryptography system using Cover Image share embedded security algorithm (CISEA), 978-1-4577-1386-611 ©2011. IEEE, pp 462–467 (2011)
4. Wang, Z., Arce, G. R.: Halftone visual cryptography through error diffusion||. *IEEE Transaction on Information Forensics and security*, ISBN 1-4244-0481-9/06. IEEE, pp 109–112 (2006)

5. Gopal Krishna, D. D., Wakade, D. G.: Digital image processing laboratory: image half toning. April 30, Purdue University, pp 01–07 (2006)
6. Mandal, J. K., Ghatak, S.: Constant aspect ratio based (2, 2) visual cryptography through meaningful shares (CARVCMs). In: IEEE 1ST international conference on communication and industrial application (ICCIA-2011 Paper ID 92), pp 01–04 (2011)
7. Kamath, M., Parab, A.: Extended visual cryptography for color images using coding tables. In: 2012 international conference on communication, information & computing technology (ICCICT), Mumbai, India 978-1-4577-2078-9/12. IEEE, Vol. 4, Issue. 5, Oct 2011, pp 39–46 (2011)
8. Yu, B., Xu, X., Fang, L.: Multi-secret sharing thresholded visual cryptography. CIS Workshops 2007, Harbin, pp. 815–818 (2007)
9. Han, Y., Dong, H.: A verifiable visual cryptography scheme based on XOR algorithm. 978-1-4673-2101-3/12/\$31.00. IEEE (2012)
10. Kaur, K., Khemchandani, V.: Securing visual cryptographic shares using public key encryption, 978-1-4673-4529-3/12/\$31.00c. IEEE (2012)
11. Chen, T. H., Tsao, K.-H., Wei, K.-C.: Multiple-image encryption by rotating random grids. Eighth international conference on intelligent systems design and applications, pp. 252–256 (2008)
12. Chen, T.-H., Tsao, K.-H., Wei, K.-C.: Multi-secrets visual secret sharing. Proceedings of APCC2008, IEICE, pp. 325–335 (2008)
13. Verheul, E., Tilborg, H. V.: Constructions and properties of K Out of N visual secret sharing schemes. Designs, Codes and Cryptography 11(2), pp. 179–196 (1997)
14. Yang, C., Lai, C.: New colored visual secret sharing schemes. Designs, Codes and cryptography, Vol-20, Springer, pp. 325–335 (2000)
15. Chang, C., Tsai, C., Chen, T.: A new scheme for sharing secret color images in computer network. In: Proceedings of international conference on parallel and distributed systems, pp. 21–27 (2000)
16. Chang, C.-C., Yu, T.-X.: Sharing a secret gray image in multiple images. In: Proceedings of the first international symposium on cyber worlds (CW.02), pp. 300–304 (2002)
17. Youmaran, R., Adler, A., Miri, A.: An improved visual cryptography scheme for secret hiding. In: 23rd biennial symposium on communications, pp. 340–343 (2006)
18. Shyu, S.J.: Efficient visual secret sharing scheme for color images. Pattern Recogn. **39**(5), 866–880 (2006)
19. Tsai, D.-S., Horng, G., Chen, T.-H., Huang, Y. T.: A novel secret image sharing scheme for true-color images with size constraint. Inform. Sci. 179 3247–3254 Elsevier, pp. 122–129 (2009)
20. Liu, F., Wu, C.K., Lin, X.J.: Colour visual cryptography schemes. IET Inform Security 2(4), pp. 151–165 (2008)

Threat Analysis and Attack Modeling in Data-Centric Communication for Named Data Networking



Riddhi Mirajkar, Gitanjali Shinde, Parikshit Mahalle, and Nilesh Sable

Abstract The Internet has prevailed since it was first introduced in the 1970s. In the meantime, various Internet architectures were studied and tested to replace traditional IP-based Internet architecture. Information-centric networking is where data is most important rather than the host. Many ICN architectures were investigated and discussed; among them was Named Data Networking (NDN), and some of them were mentioned. It is a non-IP-based technology with a caching mechanism implemented inside the routers. Mainly, the NDN architecture consists of the consumer who consumes the data, the producer who will produce the data and make it available across the network, and the router, which has a caching mechanism that transfers request packets and data from one node to another. However, NDN also suffers various attacks, such as cache-based attacks, flooding attacks, and poisoning attacks. Studying the behavior of these attacks is a crucial part of building the internet architecture. The paper discusses the attacks affecting NDN, how these attacks affect the data, and what the attacker aims to achieve through the attack.

Keywords NDN · Information centric network (ICN) · Round-trip time (RTT) · Cache privacy attack (CPA) · Interest flooding attack (IFA)

R. Mirajkar (✉) · G. Shinde · P. Mahalle · N. Sable
Vishwakarma Institute of Information Technology, Pune, India
e-mail: mirajkarriddhi@gmail.com

G. Shinde
e-mail: gr83gita@gmail.com

P. Mahalle
e-mail: aalborg.pnm@gmail.com

N. Sable
e-mail: drsablenilesh@gmail.com

1 Introduction

We've been using IP-based internet architecture for a very long time. Traditional IP-based Internet has evolved a lot since it was started; various protocols were introduced, such as TCP, UDP, SMTP, POP3, etc., each serving its unique purpose. The Internet has prevailed so much since it was first introduced in the late 1970s [1]. However, these protocols also deal with some of the attacks related to them, and sometimes, those attacks can be detrimental from the organization's perspective. According to the CIA triad, some attacks target users' confidentiality, integrity, and availability. Even the non-IP-based technology NDN is not devoid of attacks.

The so-called Named Data Network (NDN) emerged from traditional IP-based networking focused on location-based data backhaul limitations and inefficiencies. In today's digital environment, where content is accessed from multiple sources and origins, and data usage patterns have shifted to content distribution and sharing, delivering information as a host or employer does not work correctly. NDN solves these problems by shifting the focus from where the data is to where the data is desired. This content-centric approach provides several key benefits, including increased scalability, efficiency, and flexibility in content delivery. Improving network performance through in-network caching allows data to be temporarily stored at multiple points, reducing latency and increasing access to the desired piece earlier. This is especially important for modern applications that require fast content and bandwidth optimization. Additionally, NDN's architecture supports mobility, making it ideal for IoT devices, mobile devices, and dynamic environments where devices change location. This contrasts with IP networks, which struggle with the overhead of managing mobility and connectivity. The increasing need for secure, efficient, and scalable content delivery in video streaming, cloud computing, IoT, and smart city technology drives the need for simple, content-centric network architectures. NDN's design focuses on the data rather than its location and addresses the challenges posed by today's data usage and mobile web usage patterns.

NDN is the topic of research that is said to be the future of Internet architecture and is far from being implemented in the real world. NDN is quite different from the current internet architecture. Unlike TCP/IP, it doesn't use the IP address to transfer the data as it is an information-centric network. It makes use of two types of packets to transfer data: the first packet is an Interest packet generated by the consumer to request the data, and the second packet is the Data packet, which is either generated by the producer who serves the specific content requested in the interest packet or the router which has cached that data. In NDN, the content is identified by its name. If someone seeks password.txt, the request for password.txt can be from/user/private/password.txt. Interest in the data will be sent to the producer containing the data.

NDN uses two types of packets for communication purposes, as shown in Fig. 1.

The first packet is an Interest packet generated by the consumer to request the data, and the second packet is the Data packet, which is either generated by the producer who serves the specific content requested in the interest packet or the router that has cached that data. Typically, NDN architecture is composed of Consumers, Routers,

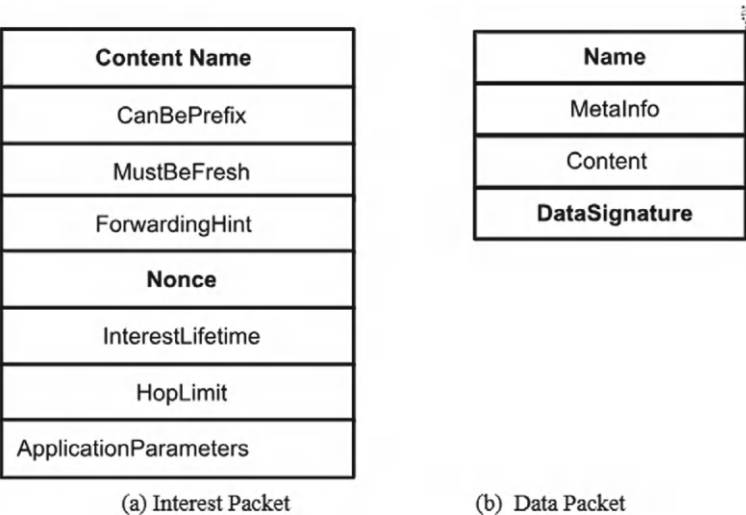


Fig. 1 NDN Packet Types

and Producer. Consumers request the data by sending request packets. Producers are the publishers of data. They preserve the data across the network, and the producer verifies the content using the signature. It uses a cryptographic algorithm to sign the data. Routers are a special entity of NDN architecture because they possess caching ability. If the router doesn't have the requested data, then the router will send the same request to the upfront network after doing some processing, which will be discussed further; after getting data from the producer, all the routers that carried the request will cache the data and send it to the consumer. Routers in NDN consist of three Data Structures, as shown in Fig. 2

1. Content Store: It is used to cache the data whenever the data is present in the content store; the router fulfills the request for the data. The router uses caching strategies like LRU, LFU, and FIFO, each with advantages and disadvantages.
2. Pending Interest Table: When requested data is not present inside the content store, PIT will create a new entry for the interest if the entry for that packet doesn't already exist in PIT. It keeps the entries of unsatisfied interest packets, and once data is obtained, the entry corresponding to it is dropped.
3. Forward Information Base: It is used to forward the packet based on the content name present in the interest packet to the upfront network.

Traditional TCP/IP suffers a wide range of attacks, such as DoS, DDoS, DNS Poisoning, session hijacking, wormhole attacks, Sybil attacks, and Black Hole attacks. Similarly, NDN also suffers a few attacks, the primary threat of which is to safeguard cached data inside the router. Protecting the data inside the cache is crucial to building the NDN architecture. However, introducing a cache inside the router would also introduce threats related to it, like adversaries monitoring the

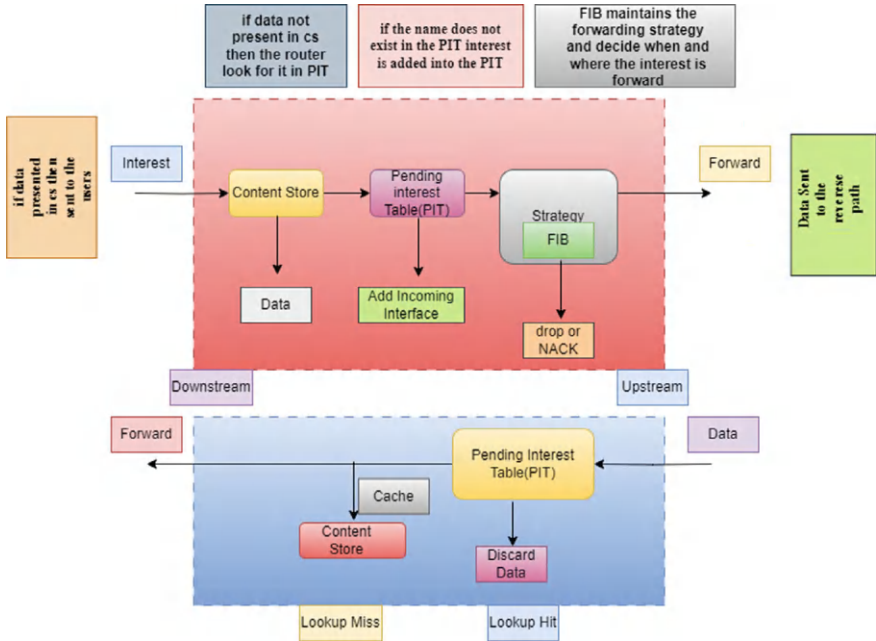
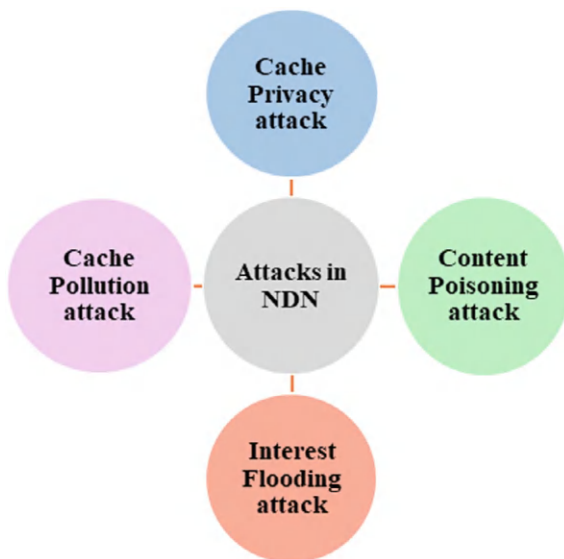


Fig. 2 NDN communication process

behavior patterns of the NDN network where they can calculate RTT time to devise an attack. Infecting routers to manipulate the network, interest, and data packets and making legitimate content unavailable inside the router to naïve users. Many attacks are associated with the NDN’s cache, such as Cache privacy attacks, Content pollution attacks, and Cache pollution attacks (Fig. 3). Apart from the cache attacks, some attacks, such as Interest Flooding Attacks, affect the NDN architecture. Mitigation of all these attacks is essential while designing the NDN’s architecture [2].

Motivation: The increasing dependence on data-centric applications such as the Internet of Things, smart cities, and extensive content sharing necessitates robust security mechanisms in NDN. Although NDN simplifies things to be more efficient and scalable, its unique design introduces vulnerabilities malicious users can use to launch attacks, such as cache manipulation, interest flooding, and content poisoning. These attacks compromise user privacy and data integrity, and conventional security mechanisms are insufficient to address NDN’s unique problems. As digital privacy issues increase, protecting users from unwanted tracking and data breaches has become critical. Knowing how attacks occur and implementing proactive security measures is essential to make NDN more robust and reliable. This research seeks to study attack patterns and develop robust defense mechanisms to make NDN more secure, making it reliable as a solid foundation for future internet architectures.

Fig. 3 Attacks In NDN

2 Literature Survey

The paper introduces a blockchain-based architecture to secure NDN in vehicular edge computing, addressing key management and cache poisoning. Here, vehicles act as consumers and producers and share their location, speed, and traffic information. It makes use of the Delegate Consensus Algorithm to address the threat. Delegated Proof of Stake mainly influences it. Improvements are needed in handling request latency and optimizing signature verification in complex networks. It also suggests a hierarchical naming structure to improve the association between applications and their content [5].

It addresses the cache privacy attack by giving a unique approach. It uses the LRU algorithm for caching purposes. It employs the delay strategy to avoid the cache privacy attack. It applies the delay based on behavior. If the behavior detected is suspicious, the delay will be applied. As for the naïve user, the delay is not applied to the naïve user. It detects this behavior by using the proprietary algorithm. It adds extra fields to the interest packet to detect the behavior [6].

This paper presents an innovative approach to compromising user privacy within the NDN framework. It introduces a proactive attack strategy where an adversary induces a router to cache specific content and subsequently verifies whether a victim has requested the same content by probing the cache state. This method circumvents traditional reactive attack defenses and poses significant challenges in detection due to its subtle operation. The paper discusses a potential vulnerability in NDN's router caching mechanism and shows the need for the mechanism to counter caching-related issues [7].

Pollution attacks occur when malicious nodes inject false or harmful data into the network, leading to significant integrity issues. The paper proposes a path diversity strategy, which enhances the resilience of data transmission by utilizing multiple diverse paths to deliver data packets. This method aims to reduce the chances of malicious interference affecting all paths simultaneously. The authors present a thorough analysis of their approach, including theoretical foundations and empirical evaluations, demonstrating its effectiveness in maintaining data integrity under various attack scenarios [8].

Q-Learning has dealt with cache pollution attacks in Named Data Networks. NDN routers use Q-learning agents to identify traffic patterns and malicious users by analyzing the Cache hit ratio, Inter-Arrival Time, and Hop Count. When the malicious nature of packets is detected, the packet is discarded. Q-Learning's adaptability to dynamic environments makes it well-suited for this task, ensuring real-time detection and prevention of evolving CPA threats [9].

Some other strategies also exist to enhance name privacy in NDN, such as using PEKS in NDN, where the content name is encrypted, forwarded, and searched without knowing the name of the content. The producer encrypts the content name and disseminates it across the network. The producer also disseminates its public key, signed with RSA keys, to ensure authenticity [10].

Explores NDN caching strategies and highlights the need for advancements in mobile node support, QoS-based caching, and energy-efficient techniques for improved performance in mobile networks [11].

NDN routers can utilize different routing approaches, such as Geographical Routing, Link State Routing, and Distance Vector Routing. The paper describes reactive and Proactive routing strategies in NDN. It also depicts hybrid approaches that balance proactive and reactive strategies, particularly for wireless ad hoc networks, where node mobility influences routing efficiency. Practical implementation of Name Link State Routing is discussed, using Link State Advertisements to maintain network State. It also discusses NDN routing challenges and open issues [12].

This survey paper studies the Interest Flooding Attack in NDN. The author has categorized the Interest Flooding Attack into two types: Non-Collusive Interest Flooding Attack and Collusive Interest Flooding Attack. The author further classifies interest flooding attacks into four attack models; the classification is based on the approach used by the attacker. The four types are Satisfied Interest Existing Data, Satisfied Interest Dynamic Data, Unsatisfied Interest Non-Existent Data, Malicious Interest Non-Existent Data. The paper also describes various Interest Flooding Detection Techniques such as Rate-based, Rule-based, Attribute-based, etc. [13].

By using caching, Named Data Networking (NDN) improves data retrieval. Caching decision techniques decide the location of content storage; Leave Copy Down (LCD) is a popular choice since it strikes a compromise between redundancy and efficiency. The Zipf-Mandelbrot distribution model of content popularity affects caching, giving priority to content that is viewed frequently. Cache storage is effectively managed using replacement rules such as Least Recently Used (LRU) and Least Frequently Used (LFU). Although LCD, LRU, and LFU are frequently used,

future studies can look into adaptable tactics depending on network conditions and content trends [14].

Compared to conventional Interest Flooding Attacks, Collusive Interest Flooding Attacks in Named Data Networking pose a significant problem because of their intermittent nature, which makes detection challenging. Improved Collusive Interest Flooding Attacks, an improved variation, maximize attack effectiveness while preserving stealth. A machine learning-based detection method employing Gradient Boosting Machines, such as GBDT, XGBoost, and LightGBM, is suggested as a countermeasure to I-CIFA. The BO-GBM fusion algorithm uses Bayesian optimization to improve parameter selection while classifying network traffic to differentiate between normal and attack states. According to experimental results, BO-GBM outperforms traditional detection techniques, achieving a 98.69% detection rate with few false alarms. This work emphasizes ensemble learning and cybersecurity optimization strategies, demonstrating the efficacy of ML-driven tactics in reducing NDN-based DDoS threats [15].

Despite using in-network caching to optimize content distribution, Named Data Networking is susceptible to side-channel timing attacks, in which adversaries use the distinction between cached and un-cached content to deduce private information. Conventional defenses successfully lessen these kinds of attacks, but they can impair NDN's functionality. The AT&T network topology and ndnSIM are used in this study to investigate side-channel timing assaults that use brute-force methods against streaming applications. Detection and Defense, a multi-level countermeasure, is suggested to detect adversary nodes while maintaining valid requests. According to simulations, it effectively mitigates attacks without interfering with the distribution of content by lowering the cache hit ratio to 0.7%, beating probabilistic (4.1%) and freshness-based (3.7%) approaches [16].

By facilitating content-based data retrieval, Named Data Networking is a new paradigm that aims to overcome the drawbacks of the conventional TCP/IP architecture. Although NDN automatically mitigates traditional Distributed Denial of Service attacks, it is nevertheless susceptible to Collusive Interest Flooding Attacks, which use the Pending Interest Table to impair network performance. The non-parametric CUSUM technique, which effectively detects network anomalies with little processing overhead, is used in this study to propose a lightweight, stateless CIFA detection mechanism. To lessen the impact of the assault, a mitigation method based on average response time values is also introduced. According to experimental data, the suggested method considerably lowers PIT utilization and improves customer satisfaction while detecting CIFA in large-scale networks in 199.5 ms [17].

This paper discusses the Interest Flooding Attack (IFA) in Named Data Networking, in which adversaries disrupt regular traffic by flooding routers with fraudulent requests that fill the Pending Interest Table. Current methods either hurt legitimate traffic or are ineffective. Through the removal of malicious interests and the blocking of new ones, as well as the addition of a security layer at network edges for early attack detection, the suggested method improves PIT management. Compared to current techniques, when implemented in ndnSIM, the solution maintains 97% PIT

availability and preserves 5–40% more valid traffic, improving network performance [18].

The problem of in-network cache allocation in Content-Centric Networking, which is susceptible to pollution attacks that reduce caching efficiency, is discussed in this study. Current methods fall short of balancing security and efficient cache allocation. The authors suggest a lightweight, non-collaborative cache allocation technique that mitigates pollution attacks and improves caching performance by considering locality and content popularity. IFDD is efficient because it reduces computational and communication overhead. IFDD minimizes the impact of pollution attacks while improving the cache hit ratio and request processing time, according to simulations conducted on the ndnSIM platform [19].

Interest Flooding Attacks, which exhaust the memory resources of the Pending Interest Table in NDN routers, are one of the security issues in Named Data Networking discussed in this study. By deflecting malevolent interests from the PIT, the authors' innovative method of Disabling PIT Exhaustion reduces IFA. DPE uses a packet marking system to facilitate data packet forwarding without depending on the PIT and stores state information directly in the Interest name. Numerous simulations show that DPE preserves network efficiency while successfully lowering PIT memory depletion. By separating malevolent Interests from the PIT, this strategy is a groundbreaking attempt to combat IF [20].

3 Threat Analysis: (Types of Threats or Attacks that Can Take Place in NDN Networks)

- **Cache Privacy:** NDN routers offer caching capacity to increase data retrieval efficiency. The attackers can read the cached content through the caches, which may lead to a loss of privacy. The routers with no security settings or security features are the network chain's weakest link if they are not equipped with security features. Attackers may infer the interest patterns of the users, access private information, or perform other attacks depending on the cached content analysis.
- **Malicious Node:** A compromised node within the network can serve as a portal for malicious activity, enabling the attacker to intercept all requested and received content, thereby violating user privacy. If the targeted node is a producer, it will spread malicious or unpopular data, making the authentic data less available and misleading the consumers. An exploited router can reject valid requests, redirect information, or selectively forward interests, leading to significant network efficiency and availability disruptions. Attackers use the compromised nodes to generate spam content advertisements to mislead users and flood the network with false information.
- **Unpopular Content:** Attackers may ask and cache unpopular content intentionally to fill the router's limited cache capacity. This evicts frequently used legitimate content from the cache and makes it inaccessible to legitimate users. This assault

can be employed to reduce the performance of a network, raise retrieval latencies, and configure the caching policy to benefit the attackers. One such attack is cyclic unpopular content requests, where valuable cache space is filled with low-value content for as long as possible.

- **Poor Configuration of Network:** Suppose you don't limit the user's activities and let the user request as many times as they want. In that case, the malicious user can perform flooding attacks to target data availability.

4 Attack Modeling

1. Cache Privacy Attack

Cache privacy in NDN refers to protecting sensitive information stored in caches from unauthorized access and exploitation (Fig. 4). Since NDN relies on caching to improve efficiency, cached data can reveal patterns about user behavior, content access, and producer identities. Attackers may exploit this cached information to infer sensitive details, such as user preferences or system vulnerabilities. This attack explicitly targets the user's privacy.

Cache privacy attacks severely threaten NDNs since they target caching behavior. Intended to enhance performance. Malicious actors can see patterns in cache hits and fail to infer user behavior, track activity, and potentially access sensitive information without viewing the user's information. This can lead to user profiling, behavior monitoring, and even the sharing of private messages. Such weaknesses put user Privacy concerns about being spied upon, and thus, it is essential to develop a strong defense against such attacks. Strong countermeasures need to be taken to ensure that NDN in-network caching is efficient while maintaining the privacy of its users.

While performing the attack, the adversary needs to be very careful as he might get the data back, which is cached by himself and not the typical user. The work done in [21] discusses how the adversary can request data for the first time, measure the time to receive the response, and make the second request; if the time difference between both is approximately zero, it means someone has requested the data. It also discusses the issue of the scope field in interest, where the adversary can change the scope field to limit the NDN routers traveled by interest.

2. Cache Pollution

Cache pollution occurs when unwanted or malicious data fills the cache, displacing legitimate content and reducing the overall efficiency of data retrieval (Fig. 5). The adversary requests unpopular content to cache it inside the NDN's routers. It results in legitimate content being unavailable for naïve users. When naïve users request legitimate content as the content cached inside the router is unpopular, their request is not satisfied. This attack mainly targets content availability. The following diagram describes the flow of the cache pollution attack. The diagram shows that before the attacker requested a large amount of content, naïve users could quickly get the data from the network; however, after the attacker asked for malicious content, the naïve user could not get the legitimate content. This attack results in reduced cache hits.

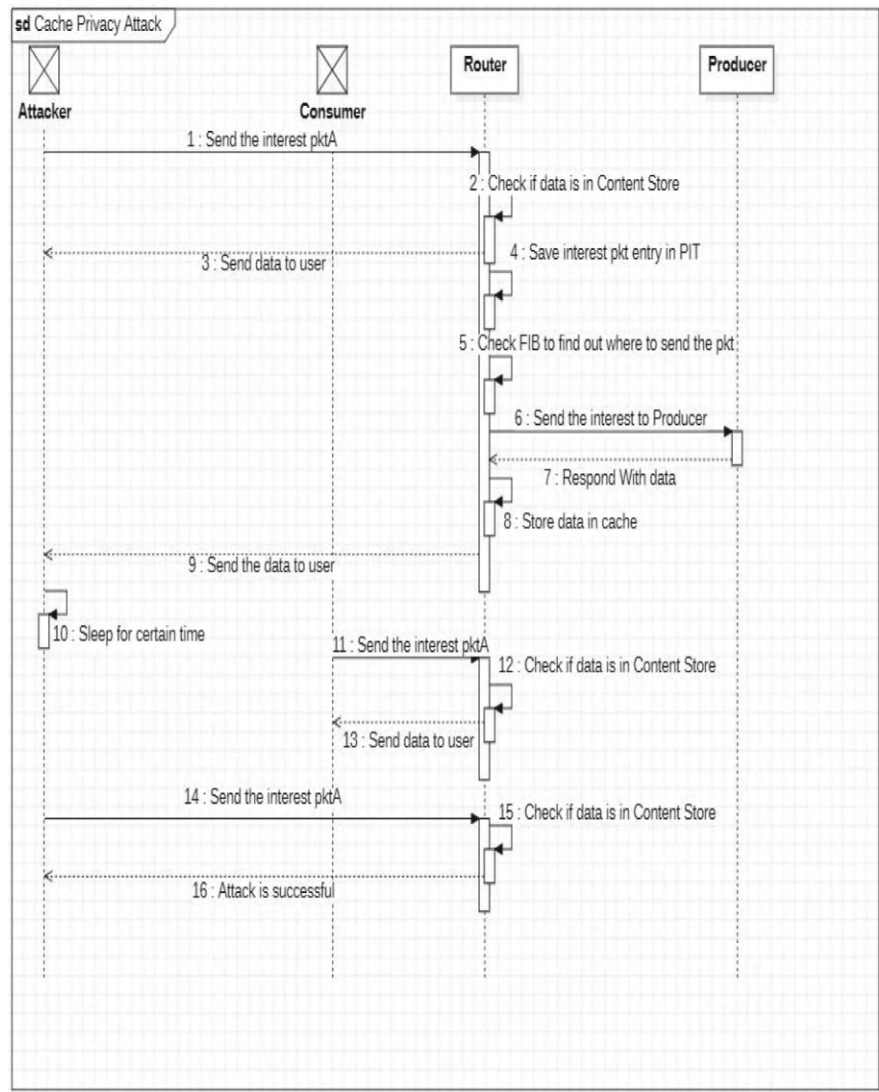


Fig. 4 Cache privacy attack

There are two types of cache pollution attacks: Locality Disruption Attacks and False Locality Attacks. Locality Disruption Attacks involve requesting unpopular content, due to which cache hit for legitimate users is reduced. In a False Locality Attack, the attacker requests unpopular content repeatedly [22].

3. Content Poisoning

The threat in this attack is the distribution of malicious content across the network (Fig. 6). Content poisoning attacks involve attackers injecting malicious or tainted

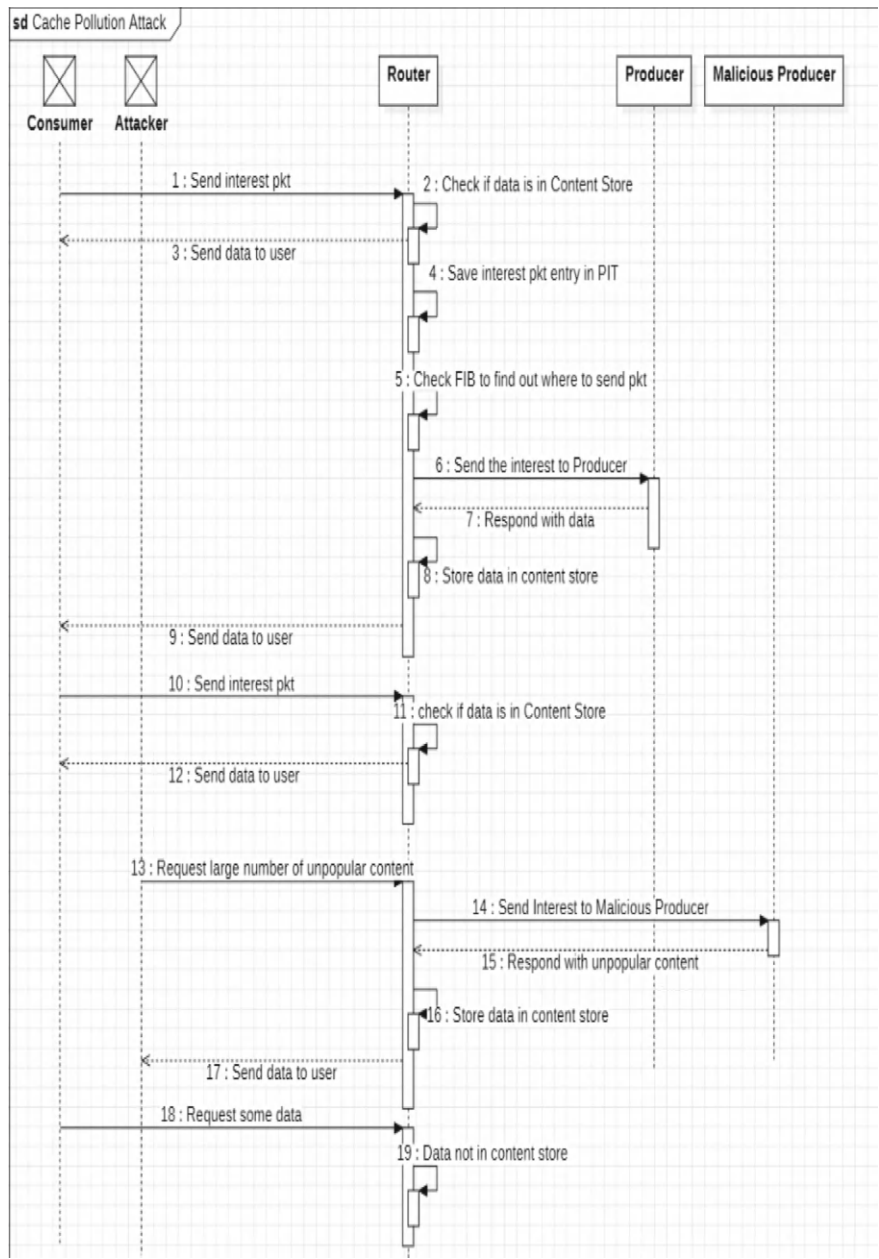


Fig. 5 Cache pollution attack

content into the NDN cache, which makes users download malicious or faulty data, leading to trust failures and potential data corruption. Interest flooding attacks (IFA) overwhelm the Pending Interest Table (PIT) by inserting enormous quantities of malicious interest packets, leading to resource exhaustion and denial-of-service situations. The malicious producer is intentionally queried for malicious content by the attacker. This renders malicious content stuck within NDN routers. When the original user requests the content, malicious content is delivered to the user. It threatens the availability and privacy of the users' content. It may even compromise the host if malicious data contains malware, trojan, or virus. Signs of content poisoning are invalid signatures, inconsistent checksums, or unauthorized content sources, which indicate possible content poisoning. Real-time content integrity and trust monitoring can suppress attacks. Cache pollution detection is marked by reduced cache hit rates and abnormal content popularity distributions, indicative of cache pollution attempts. Request distribution analysis and maintaining cache validation policies can deter these attacks.

5 Interest Flooding Attack

This attack primarily targets the PIT of the NDN router. It aims to bring down the efficiency of the network. Here, adversaries make large requests, making the pending interest table full of unprocessed requests, significantly reducing the network's performance. When a naïve user requests the data, his request gets dropped as PIT is full and can't process more requests. This results in content being unavailable. A single or small group can perform the IFA, and large groups can perform it. The former is known as Local IFA, and the latter is known as Distributed IFA. This attack is specifically directed towards the Pending Interest Table (PIT) of the NDN router to hamper network efficiency. Attackers create massive interest requests, overflowing the PIT with unprocessed requests and substantially reducing network performance. When an honest user tries to access content, the request may be lost since the PIT is complete, making the requested content unavailable. The effect of this attack is shown in Fig. 7. Interest Flooding Attacks can be divided into two categories: Local IFA and Distributed IFA. Local IFA is conducted by one attacker or a small group of adversarial agents, while Distributed IFA is a massive-scale attack conducted by hundreds of attackers from different locations. Distributed IFA is especially dangerous since it replicates actual traffic patterns, making it more difficult to detect. Attackers can dynamically vary their request rates, making defense more difficult. The severity of an IFA depends largely on parameters such as the attack rate, the distribution of attacking nodes, and the effectiveness of PIT management mechanisms. Effective countermeasures should consider the adaptive nature of attackers and incorporate real-time monitoring mechanisms to detect and neutralize IFA before it affects the network substantially.

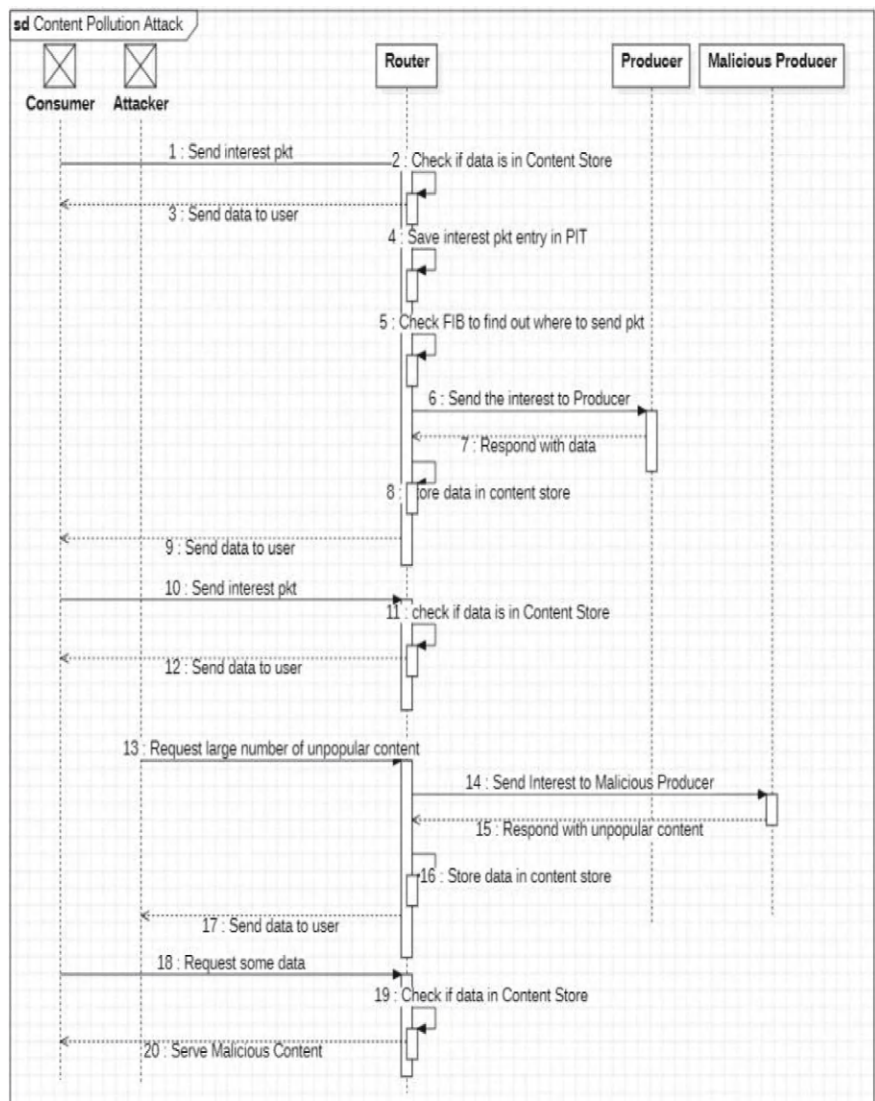


Fig. 6 Content pollution attack

6 Conclusion

Our study provides an in-depth analysis of cache-centric vulnerabilities that can occur in NDN. This research aimed to analyze how these weaknesses can threaten one’s privacy. Several threats like cache privacy attacks, cache pollution, and content pollution can cause breaches of sensitive user information and cause the degradation

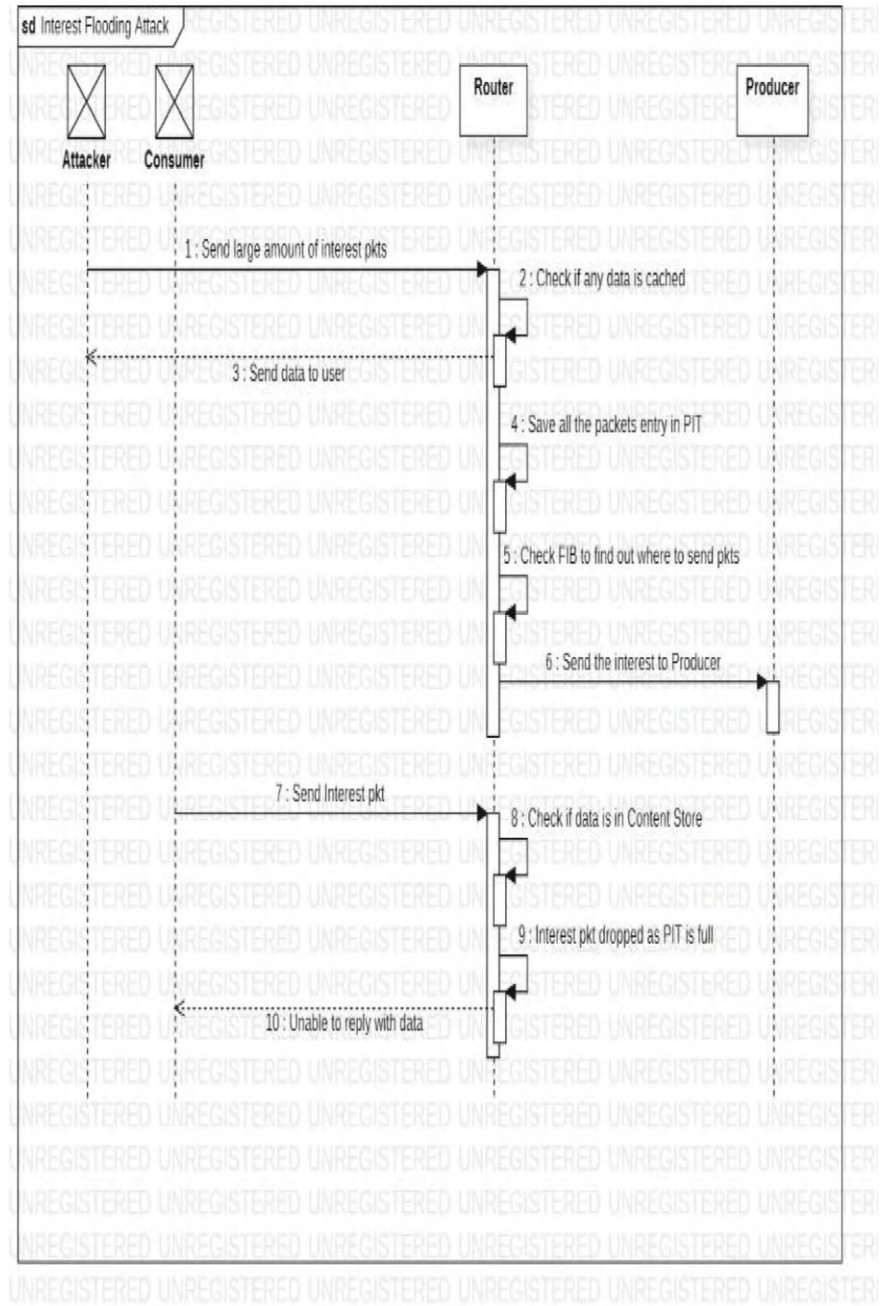


Fig. 7 Interest flooding attack

of the network's performance, thus challenging the adoption of NDN over traditional systems. Various approaches to tackle these challenges were studied, including delay-based strategies, blockchain-based security frameworks, and Q-learning techniques. All these approaches tried to tackle some possible threats but compromised the network's performance. For the smooth working of NDN, it is necessary to maintain a tradeoff between security and performance. By fortifying the network against cache-based threats without diminishing its performance, NDN has the potential to serve as a robust and secure alternative to conventional IP-based systems, meeting modern communication needs while safeguarding.

References

1. Hidouri, A., Haddad, M., Touati, H., Hajlaoui, N., Muhlethaler, P.: Detection mechanisms and their limits in named data networking (NDN) (2022) [Online]. <https://hal.science/hal-03933012>
2. Hidouri, A., Touati, H., Hajlaoui, N., Haddad, M., Muhlethaler, P.: A survey on security attacks and intrusion detection mechanisms in named data networking. *Computers* 11(12), (2022). <https://doi.org/10.3390/computers11120186>
3. Pattnaik, L. M., Swain, P. K., Satpathy, S., Panda, A. N.: Cloud DDoS attack detection model with data fusion & machine learning classifiers. In: *EAI endorsed transactions on scalable information systems* 10(6) (2023)
4. Satpathy, S., Swain, P. K., Mohanty, S. N., Basa, S. S.: Enhancing security: federated learning against man-in-the-middle threats with gradient boosting machines and LSTM. In: *2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1–8). IEEE (2024)
5. Lei, K., et al.: Blockchain-based cache poisoning security protection and privacy-aware access control in ndn vehicular edge computing networks. *J. Grid Comput.* **18**(4), 593–613 (2020). <https://doi.org/10.1007/s10723-020-09531-1>
6. Kumar, N., Singh, A. K., Srivastava, S.: A triggered delay-based approach against cache privacy attack in NDN (2018)
7. Compagno, A., Conti, M., Losiok, E., Tsudik, G., Valle, S.: A proactive cache privacy attack on NDN. In: *Proceedings of IEEE/IFIP network operations and management symposium 2020: management in the age of softwarization and artificial intelligence, NOMS 2020*, Institute of Electrical and Electronics Engineers Inc. (2020). <https://doi.org/10.1109/NOMS47738.2020.9110318>
8. Guo, H., Wang, X., Chang, K., Tian, Y.: Exploiting path diversity for thwarting pollution attacks in named data networking. *IEEE Trans. Inf. Forensics Secur.* **11**(9), 2077–2090 (2016). <https://doi.org/10.1109/TIFS.2016.2574307>
9. Hidouri, A., et al.: Q-ICAN: A Q-learning based cache pollution attack mitigation approach for named data networking ICAN: a Q-learning based cache pollution attack mitigation approach for Q-ICAN: A Q-learning based cache pollution attack mitigation approach for named data networking, vol. 235 (2023). <https://doi.org/10.1016/j.comnet.2023.109998i>
10. Ko, K. T., Hlaing, H. H., Mambo, M.: A PEKS-based NDN strategy for name privacy. *Future Internet* 12(8) (2020). <https://doi.org/10.3390/FI12080130>
11. Yovita, L.V., Syambas, N.R.: Caching on named data network: a survey and future research. *Int J Electr Comput Eng* **8**(6), 4456–4466 (2018). <https://doi.org/10.11591/ijece.v8i6.pp4456-4466>
12. Karim, F.A., Aman, A.H.M., Hassan, R., Nisar, K., Uddin, M.: Named data networking: a survey on routing strategies. *IEEE Access* **10**, 90254–90270 (2022)

13. Lee, R.-T., Leau, Y.-B., Park, Y.J., Anbar, M.: A survey of interest flooding attack in named-data networking: taxonomy, performance and future research challenges. *IETE Tech. Rev.* 1–19 (2021)
14. daSilva, E.T., de Macedo, J.M.H., Costa, A.L.D.: NDN content store and caching policies: performance evaluation. *Computers* 11, 37 (2022)
15. Wu, Z., Peng, S., Liu, L., Yue, M.: Detection of improved collusive interest flooding attacks using BO-GBM fusion algorithm in NDN. *IEEE Trans. Netw. Sci. Eng.* (2022)
16. Dogruluk, E., Macedo, J., Costa, A.: A countermeasure approach for brute-force timing attacks on cache privacy in named data networking architectures. *Electronics* 11, 1265 (2022)
17. Cheng, G., Zhao, L., Hu, X., et al.: Detecting and mitigating a sophisticated interest flooding attack in NDN from the Network-Wide View. In: 2019 IEEE first international workshop on network meets intelligent computations (NMIC). IEEE (2019)
18. Benarfa, A., Hassan, M., Losiouk, E., Compagno, A., Yagoubi, M.B., Conti, M.: ChoKIFA+: an early detection and mitigation approach against interest flooding attacks in NDN. *Int. J. Inf. Secur.* 20, 269–285 (2021)
19. Tourani, R., Misra, S., Mick, T., Panwar, G.: Security, privacy, and access control in information-centric networking: a survey. *IEEE Commun. Surv. Tutor.* 20, 566–600 (2017)
20. Wang, K., Zhou, H., Qin, Y., Chen, J., Zhang, H.: Decoupling malicious interests from pending interest table to mitigate interest flooding attacks. In: *Proc. IEEE Globecom Workshops (GC Wkshps)*, pp. 963–968 (2013)
21. Acs, G., Conti, M., Gasti, P., Ghali, C., Tsudik, G.: Cache privacy in named-data networking
22. Kar, P., Chen, L., Sheng, W., Kwong, C. F., Chieng, D.: Advancing NDN security: efficient identification of cache pollution attacks through rank comparison. *Internet of Things (Netherlands)* 26, (2024). <https://doi.org/10.1016/j.iot.2024.101142>

Energy-Efficient Machine Learning-Based Data Encryption Techniques for Information Blocks: A Comprehensive Analysis



Mrunal S. Jagtap and D. Sangeetha

Abstract Recent research has focused on image encryption due to growing concerns about communication security. While conventional encryption techniques are successful, they often struggle with scalability for large systems and adaptability to new security challenges. This study investigates the use of machine learning (ML) in image encryption, exploring its potential to transform data protection. The research evaluates various ML techniques, such as neural networks, reservoir computing, and support vector machines, to enhance signal synchronization, reduce noise, and strengthen encryption. The study outlines the advantages of ML-enhanced encryption, including adaptability and improved security, while also noting drawbacks like computational complexity and susceptibility to adversarial attacks. The results underscore ML's capacity to revolutionize chaotic encryption, offering innovative approaches for secure communication in the Internet of Things (IoT), cloud computing, and wireless network applications. ML provides a sophisticated and adaptive approach to encryption by employing algorithms that can identify, adjust, and improve encryption methods based on data patterns and real-time conditions. By leveraging deep learning models, neural networks, and advanced statistical methods, ML enables the development of robust encryption systems capable of resisting traditional cryptographic attacks. These systems can identify anomalies, detect vulnerabilities, and modify encryption techniques to meet specific data security needs.

Keywords Machine learning · Chaotic logistic map · Encryption · Key generation

M. S. Jagtap (✉) · D. Sangeetha
MIT-WPU, Pune, India
e-mail: mrunalsjagtap95@gmail.com

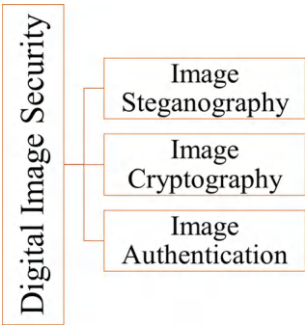
© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
M. Yang et al. (eds.), *Demystifying AI and ML for Cyber-Threat Intelligence*,
Information Systems Engineering and Management 43,
https://doi.org/10.1007/978-3-031-90723-4_34

499

1 Introduction

In the current digital era, the pervasive use of visual communication and the need for secure image data transmission have underscored the importance of robust encryption techniques. Image encryption and decryption are crucial in preserving data integrity, safeguarding privacy, and securing sensitive information across various sectors, including social media, healthcare, and defense. While traditional encryption methods like AES and RSA are widely employed, the escalating complexity of contemporary threats and the increasing volume of image data necessitate innovative, flexible, and efficient approaches. Machine learning (ML), a branch of artificial intelligence, offers novel strategies to enhance encryption and decryption processes. ML algorithms enable the development of adaptive models capable of recognizing patterns, generating dynamic encryption keys, and fortifying cryptographic systems against emerging threats. Additionally, ML-based techniques facilitate the automation of key generation, anomaly detection, and encrypted image reconstruction, thereby expediting and improving the overall procedure. Integrating machine learning with cryptography not only addresses limitations of conventional methods, such as scalability issues or vulnerability to brute-force attacks but also introduces new paradigms for secure and efficient image processing. This research explores the intersection of machine learning and image encryption and decryption, investigating various approaches, challenges, and potential applications of these cutting-edge techniques. The need for strong security measures has increased due to the exponential expansion of digital photographs exchanged across networks. Sensitive image data, including scanned documents, medical images, and private photos, is frequently handled by cloud storage, telemedicine, online banking, and e-commerce applications. In certain situations, a security compromise might have serious repercussions, such as identity theft, patient privacy breaches, or the disclosure of secret material. Digital image security is broken down hierarchically into three main parts, as shown in Fig. 1

Fig. 1 Digital image security components



1.1 Image Steganography

Steganography is a method of concealing information within an image by subtly modifying it. This technique is commonly used to embed data or messages discreetly without drawing attention.

Example: Incorporating a watermark or confidential information into an image file.

1.2 Image Cryptography

This process involves encoding image data to ensure privacy and prevent unauthorized access. Cryptography transforms the image into an incomprehensible format (ciphertext) that can only be deciphered by authorized users with the appropriate decryption key.

Example: Safeguarding medical images during network transmission.

1.3 Image Authentication

This technique verifies the integrity and origin of digital images. It involves adding a digital signature or watermark to confirm that the image has not been tampered with and originates from a trusted source.

Example: Verifying the authenticity of digital evidence in forensic investigations.

These techniques address digital image confidentiality, integrity, and authenticity, making them crucial for secure image processing and transmission across various fields, including forensics, secure communication, and media sharing.

Role of Machine Learning in Image Encryption

Integrating image encryption with data-driven models for various encryption stages establishes a novel framework within scientific methodologies. Unlike traditional techniques predicated on fixed keys and established algorithms, machine learning (ML)-enabled encryption systems can optimize for performance and dynamically adapt to expose systems to innovative learning threats. Multiple such advancements have been made in image encryption via machine learning, which include:

Flexible Key Generation: ML algorithms may use patterns found in the input data to create dynamic encryption keys. This makes critical guess/copy more difficult for attackers.

Anomaly Detection: ML models detect anomalies from encrypted images, ensuring the data's integrity.

Using ML-Based Compression with Encryption for High Efficiency: Combining encryption with advanced compression algorithms based on machine learning technology allows for minimizing the size of encrypted images without damaging the encoded information security.

Feature extraction capabilities of deep learning models were exploited to construct safe encryption schemes based on Convolutional Neural Networks (CNNs) and generative adversarial networks (GANs).

2 Literature Review

The increasing need for dependable, scalable, and adaptable cryptographic mechanisms has recently attracted attention to Image encryption based on machine learning (ML). This part provides a comprehensive literature review on image encryption, highlighting key methods, advances, issues, and future directions.

Classical cryptographic approaches have been employed for securing image data since ancient times. This is achieved through cryptographic techniques, which convert your data into an incomprehensible format using mathematical principles, thus ensuring confidentiality. AES uses substitution-permutation networks to secure pixel data, while RSA utilizes asymmetric keys for secure transmission. However, traditional methods are inadequate in addressing image data's immense dimensionality and redundancy. Because of these disadvantages, researchers are currently exploring alternative approaches that use the unique characteristics of visual data [1].

Typical encryption methods like AES and DES are not more efficient for image encryption. Researchers must work on one algorithm to provide image security on public network platforms. This paper does not work on a single ML algorithm. Instead, it reviews and summarizes various deep-learning approaches for image encryption. Some main approaches discussed include deep neural networks, convolutional neural networks (CNNs), deep autoencoders, and generative adversarial networks (GANs). They are often combined with chaotic systems or classical cryptography methods for image encryption [2].

In this paper [3], A growing number of people are accessing and transmitting digital image information—which may include vital data like satellite maps, architectural plans of significant national institutions, medical photographs, and facial photos—using digital cameras, cellphones, and other mobile devices. This paper uses convolutional neural networks (CNNs) combined with chaotic image encryption to develop a robust image encryption algorithm for this research. A secure, high-efficiency image encryption procedure is proposed by combining CNN's profound feature extraction ability with chaotic encryption randomness. Chaotic image encryption uses chaotic sequence generation and nonlinear mapping methods to mix up the pixel values for encryption purposes. A deep-learning model is used with a local perceptual field and weight sharing to extract high-level image features from a CNN.

Validation via experiments of this technology shows the benefits in terms of security, encryption speed, and attack mitigation.

Conventional methods have been used to protect encrypted images from unauthorized access. Still, it is observed that the traditional techniques are inappropriate for encryption purposes and are defective due to new forms of threats. The study's findings demonstrated that the cybersecurity recorded was highly improved by merging the machine learning module (CNN models) with the image encryption methods. Unlike traditional encryption techniques, ML-based encryption is computationally efficient, attack-invariant, and more accurate at differentiating between encrypted and decrypted images. These methods may be beneficial for high-dimensional image data in terms of scalability and flexibility. Nevertheless, challenges remain, including the need for enhanced model interpretability and robustness against adversarial attacks [4].

This paper [5] used a Convolutional Neural Network (CNN) to extract iris features in the context of image encryption. The CNN is trained on the CASIA Iris Image Database to automatically extract distinctive features from iris images, which are then used to generate encryption keys. The algorithm integrates Reed-Solomon error correction coding to ensure key consistency, mitigating any variations during feature extraction. The study employs CNN to effectively process and extract iris features, crucial for generating secure encryption keys and ensuring robust encryption and decryption processes.

This paper [6] suggested that the proposed method be used to secure the digital images during transmission. It is achieved through scrambling them. So, it uses a mathematical technique called a chaotic logistic map. Here, a secret key is generated using this chaotic logistic map method, which makes it harder to predict. This secret key is divided into two parts: 128 bits long each. This key has been scaled to fit the image's dimensions. Four rounds are performed during the encryption and permutation phase using unique keys, enhancing security. This method is fast and secure. It can encrypt and decrypt the images quickly.

Depending upon the DL architecture, the author introduces one novel image encryption method for securing digital images. This paper also discusses image encryption techniques that combine image optimization with cryptography and deep learning. This paper explains the Secure Crypto General Adversarial Network (GAN), a novel image encryption approach that combines Deep Learning and cryptography concepts. The GAN architecture consists of three components: a. Encoder, b. Decoder, c. eavesdropper, a To encrypt face images, an encoder is used; a decryptor is used to recover the input image, and an eavesdropper attempts to decrypt the image without the secret key. The encrypted images appear like random noise for better security using GAN Architecture. In addition, images are preprocessed using an Optical Chaotic Map, which provides chaotic sequences that can increase the complexity and attack resistance of the encryption [7].

The paper [8] recommended a novel image encryption scheme called The EncipherGAN, which tackles the challenge of ciphering color images, especially for medical imaging, based on a Cycle Generative Adversarial Network (Cycle-GAN) architecture. It consists of a discriminator network with convolutional layers and

a leaky ReLU activation function, an encoder-decoder encryption network, and a decryption network that mirrors the encryption network.

Discriminator improves the ability of an encryption network to generate realistic cipher images during model training using the GAN method. Using this entire strategy is secure and resistant to attacks by cryptanalysts while allowing several key combinations and usage of different loss functions, such as the SSIM (Structural Similarity Index, for example), can help to maintain image quality while encryption and decryption are underway.

The paper [9] suggested that the first and foremost theme is the papers on the use of Convolutional Neural Networks (CNNs) for Image Classification, which is the central topic of a paper titled “Efficient and Privacy-Preserving Image Classification Using Homomorphic Encryption and Chunk-Based Convolutional Neural Network.” This can primarily be applied to classify the said images but, at the same time, keeps the privacy of the photos computed under HE intact. In this work, the author proposed a chunk-based CNN architecture to resolve the computational challenges of the CNN image classification on the encrypted data. Using this method, HE ensures the private and secure classification of sensitive data, and CNNs can also extract hierarchical features of images.

This paper [10] is a hybridization of chaotic encryption on AlexNet—From a deep learning model. AlexNet is a Convolutional Neural Network (CNN) with some improvements, such as a preprocessing module for real-time image text encryption. To improve generalization and reduce the computing burden, the author alters the functional layers of the AlexNet model.

Encryption in Chaos:

LSS: logistic-sine semi-chaotic system.

To generate keys, a one-dimensional chaotic system aims to encrypt the first nine planes of pixel values of an image that falls in the lower position.

Lorenz chaotic system: This multi-dimensional chaotic system encrypts the top position, i.e., higher position planes of pixel values.

One of the key features of these two chaotic systems is the generation of high-security and real-time performance encryption keys.

The paper [11] integrated a Deep Neural Network (DNN) and a chaotic system. An image encryption technique is developed using the logistic map based on discrete memory. The DNN performs the following operations: feature extraction, transformation, and the key dependence check in the proposed encryption method. Extracting suitable patterns from the visual input quickly encodes intricate patterns and delivers a range of encrypted outputs. The DNN and the chaotic features of the logistic map work in parallel throughout the transformation step, combining flexibility and randomization to improve security. It encrypts the DNN model securely and dynamically, establishing the encryption’s critical dependence on the DNN and logistic map settings. These elements collaborate to create a robust encryption system by using the abilities of deep learning and chaos theory.

The paper [12] proposed a novel CBIR scheme with access control in this work. In particular, the pictures are uploaded to the server after being encrypted. After that,

the CNN model extracts the feature descriptors. The cloud server utilizes the Bkd-tree to build the index tree of image features and calculates the distances between features, substantially speeding up the retrieval process. Ultimately, the user receives the top k most similar photos, which they decrypt to produce plaintext images. Using a CNN model for feature extraction streamlines the image feature computation process compared to conventional feature extraction algorithms. This approach also improves the user's search experience while lowering computing overhead in the retrieval process. The experiment results show that the recommended strategy is accurate and functional for safe cloud computing.

In comparison, the study doesn't use a machine-learning approach for image encryption. Therefore, to maintain a high level of security, it employs a dual encryption technique in conjunction with a chaos-based approach. The primary encryption is based on chaos theory and cryptographic principles rather than machine learning, so the machine learning system will speak for auxiliary purposes like feature analysis or optimizing it. We have a double mongo, which you can make more challenging to attack by reversing the process. This is done by applying chaos theory because you have random noise between 1 and n [13].

In this paper, the authors present a framework incorporating stages for learnable image encryption. Deep learning models, such as CNNs, must be used to accomplish the encryption and ensuing classification tasks. This procedure preserves the deep learning characteristics while simultaneously concealing the secret signals. Learnable network transformations can optimize encrypted data representations in medical imaging analysis and related applications, maintaining the model accurate while keeping the data secret. Thus, by combining deep learning and safe transformation techniques, the system achieves encryption and analysis while maintaining privacy [14].

However, this led the paper's authors to conclude that instead of learning from images and their patterns using machine learning algorithms, one could also use cryptographic techniques that employ the relatively simple concept of encrypting images using a complex matrix-based private key. This element adds another level of complexity, improving the cipher's security and randomness and making it harder for possible attackers to crack. It's important to note that machine learning is not the foundation of the fundamental encryption process, which has its roots in mathematical cryptography. However, machine learning can be applied to other tasks, such as analyzing the effectiveness of encryption or optimizing parameters [15].

The research uses convolutional neural networks (CNNs), another Deep Learning technology, to encrypt images. A neural network is used in this process to process image blocks and learn to provide an encrypted representation. The CNN detects intricate patterns and features, enhancing encryption's security and attack resilience. By combining feature extraction with deep learning and block-building methods, this technology delivers a certain level of security and a desirable degree of encryption on color photographs [16].

In the Paper [17], most businesses use machine learning with algorithms and encryption strategies based on chaos. Additional machine learning techniques, such as recurrent neural networks (RNNs) or deep neural networks (DNNs), may be used

to improve the encryption process, optimize parameters, learn intricate patterns, or automate key generation, even though the encryption itself is based on systems of chaotic dynamics.

Particularly for image data, these machine-learning techniques enhance the endurance and effectiveness of chaos-based encryption, making it safe and scalable depending on different encryption settings.

A Deep Convolutional Neural Network (CNN) named ResNet50 is used in the research to classify encrypted images. With its 50 layers, ResNet50 excels at managing the intricacies of image recognition jobs. In this instance, deep features are retrieved by feeding the encrypted images into the ResNet50 model, which has already been pre-trained on big image datasets. By learning relevant features despite the encryption, the model is modified to categorize encrypted images in a way that allows beneficial features to be detected and correctly classified even without decrypting the image. This method uses deep learning's abilities to maintain security while enabling rapid classification [18].

Instead of concentrating on machine learning techniques, the study primarily discusses different cryptographic approaches to image encryption. Its main goal is to analyze better and assess the security and effectiveness of conventional and contemporary encryption techniques, such as block ciphers, public-key cryptography, and chaos-based encryption.

However, even though the work itself is not exclusively about machine learning algorithms, they might be discussed to some degree because of their possible future uses, such as implementing Deep Learning techniques (like Convolutional Neural Networks (CNTs)) to tasks like feature extraction or encryption scheme optimization, which would increase the image encryption scheme's diversity and trustworthiness [14].

Authors typically use Generative Models or, in deep learning contexts, Generative Examples (DNNs). As the model learns how to generate keys for securely encrypting medical images, it results in a flexible and resilient encryption process.

The deep learning model produces a dynamic, convoluted key to improve and stabilize the encryption. This method allows for the development of strong, attack-resistant keys while maintaining the speed at which medical imaging data can be encrypted and decrypted [19].

To find out which machine learning technique was applied in this study, based on the title, it seems like the research discussed cryptography and deep learning techniques for medical pictures. ML-Image Encryption Summary is shown in Table 1. Common strategies in research on privacy-preserving deep learning could be as follows.

Convolutional Neural Networks (CNNs): Commonly utilized for image classification and processing tasks.

Federated learning: A distributed ML approach in which a model can be trained across many devices or servers collaborated on principle without sharing the data leaving the device.

Homomorphic Encryption: It allows for the computation of encrypted data without decrypting it.

Table 1 ML-image encryption summary

Ref	Summary
02	Style transfer-based encryption, enhanced diffusion methods, and the use of chaotic systems with deep neural networks are the three categories into which the paper divides deep learning techniques for image encryption. These techniques increase efficiency, fortify cryptographic security, and allow dynamic key generation for improved protection. However, they are still in their infancy, may present new security risks, and can be challenging to implement, making practical use difficult
03	The study introduces a new picture encryption algorithm that uses a convolutional neural network for efficient feature extraction and chaotic sequences for jumbled pixel values. This method offers high randomness, faster encryption, and improved security but requires further testing and optimization for real-time scenarios and large-scale photos
04	The application of machine learning to methods for image encryption is examined in this study, emphasizing how it can enhance security, provide flexibility, and speed up encryption procedures. Despite advancements, challenges such as implementation complexity, difficulties in interpretation, and susceptibility to adversarial attacks remain
05	The research proposes a deep learning-based approach for iris image encryption to achieve a 0% false acceptance rate (FAR) and enhance security and decryption accuracy. Benefits include Reed-Solomon error correction algorithms and improved encryption consistency. Drawbacks involve a 1.043% false rejection rate (FRR) and unpredictability in feature extraction, which affect reliability
06	The study introduces a novel digital color image encryption technique using 2D chaotic logistic maps. This method offers simplicity and high efficiency while ensuring resilience against various threats. Advantages include faster encryption and strong resistance to brute-force attacks. However, longer key generation times for larger keys may impact efficiency, and the approach might require optimization for larger image sets
07	The research presents an innovative facial image encryption method utilizing Secure Crypto General Adversarial Neural Networks (Cry_GANN) in combination with an optical chaotic map for image optimization. Networks (Cry_GANN) in conjunction with an optical chaotic map for picture optimization, the research introduces a novel facial image encryption technique. Among the advantages are high security, increased encryption efficiency, and resilience to hostile attacks. However, it could be computationally costly for best results and require specialist hardware
08	The paper presents EncipherGAN, a color image encryption system that uses the CycleGAN deep learning model to provide secure image transmission. Although it takes a lot of time and hardware, it increases resilience, reconstruction quality, and protection against plaintext attacks
09	The study suggests a chunk-based CNN and Fully Homomorphic Encryption (CKKS) as a privacy-preserving picture categorization technique. It improves efficiency and accuracy (97.1% on NEU-CLS) by partitioning images, encrypting feature-rich portions, and classifying them using a modified CNN. Although it lowers storage costs and improves privacy, it still adds computational overhead and needs to be optimized for scalability
10	Using AlexNet CNN in conjunction with Logistic-Sine & Lorenz chaotic encryption, the research introduces a real-time image text encryption technique that achieves 94.37% accuracy and strong attack resistance but has limited dataset diversity and computational cost

(continued)

Table 1 (continued)

Ref	Summary
11	The study presents a successful image encryption method that combines a logistic map based on discrete memory and a deep neural network, providing strong cryptographic features, excellent image quality, and high security. However, implementing it in real-time may require hardware and incur computational overhead
12	The study offers a privacy-preserving image retrieval method for effective similarity-based retrieval in cloud environments that uses searchable encryption and Bkd-tree indexing. CNN-based feature extraction is used, access control is granted to authorized users, and content and feature security is guaranteed
13	The study presents a double encryption technique and chaos theory-based high-security image encryption algorithm for protecting facial images. It resists differential and brute-force attacks, employs simultaneous scrambling and diffusion, and may incur delays and processing costs in real-time applications
14	The study improves the SKK encryption scheme with statistical smoothing techniques and presents a privacy-preserving deep-learning approach for medical photos. By preserving model performance without decryption, this technique protects data privacy. Deep neural networks are trained using encrypted medical images that cannot be decrypted. Benefits of the approach include improved data privacy, defense against intrusions, and little degradation in performance
15	The research introduces a novel image cryptography approach utilizing a Complex Matrix Private Key (MPK) derived from a color image. This method enhances security through a dynamic key structure that changes based on initial parameters. Experimental results demonstrate its superior encryption speed and efficiency compared to conventional cryptographic algorithms such as DES, 3DES, AES, and Blowfish. However, the technique has drawbacks, including limited testing range, reliance on the image key, and increased complexity
16	A color image encryption method combining deep learning and block embedding is proposed in the study. It utilizes Chen's chaotic system for encryption and BiLSTM networks for key generation. The approach offers robust security, resilience against cropping and noise attacks, and adequate time complexity. Nevertheless, it is limited to images with equal width and length
17	The research proposes an innovative color image encryption technique employing triple chaotic maps (Lorenz, 2D-Logistic, and Henon) to enhance security and efficiency. Encryption involves confusion and diffusion phases, ensuring high unpredictability and attack resistance. However, the method may be affected by large image sizes and exhibit higher computational complexity
18	This study presents PixJS, a chaotic-based image encryption technique for 8-bit grayscale images. It incorporates logistic maps, a scrambling process, and a linear feedback shift register to bolster security. When tested on symmetric and asymmetric images, it demonstrates effective computational performance and robust defense against attacks. Nevertheless, it is restricted to grayscale images and is computationally complex

(continued)

Table 1 (continued)

Ref	Summary
19	Block permutation, pixel permutation, and a nonlinear mixing technique based on Cramer's rule are used in this research to propose a quick and reliable image encryption scheme. This technique is appropriate for military and medical applications since it improves security, speed, and attack resistance. It might need to be optimized for real-time applications due to its high computational complexity
20	The study examines the application of machine learning methods to enhance chaos-based encryption systems. It investigates ensemble techniques, neural networks, and support vector machines for secure chaotic communication, noise reduction, and signal synchronization. However, it also highlights the computing difficulties and susceptibility to hostile attacks

Generative Adversarial Networks (GANs): These are used occasionally to produce artificial medical images while protecting patient anonymity[20].

3 Findings

Machine learning-based image encryption is an emerging field that combines sophisticated machine learning algorithms with conventional cryptographic methods to provide significant gains in the security and efficiency of image data protection. Protecting these assets has become crucial in a world where digital photos are used increasingly for personal, professional, and communication purposes. Regardless of being effective, existing encryption techniques frequently fail to keep up with the increasing volume and complexity of data.

As a growing number of digital images are being shared online, Preventing unwanted access to data is crucial while it is being transmitted. Conventional encryption techniques use secret key rounds of permutation and diffusion, which might not offer the best trade-off between security and processing speed. These techniques are not sufficiently adaptable to handle advanced cryptographic attacks; large data sizes correlate highly with neighboring pixels and redundant picture information. Due to the above reasons, traditional encryption methods like AES and DES are unsuitable for images.

The paper discusses the problem of balancing computational complexity with robust data security in the context of image encryption, specifically for real-time applications such as those within the Internet of Things (IoT). It identifies challenges in achieving high levels of encryption security and low computational requirements for processing large datasets in real-time.

Existing encryption techniques usually rely on limited machine learning algorithms, like support vector machines or neural networks, for tasks like key generation and feature extraction. This narrow selection may impede the discovery of more sophisticated or optimized encryption methods that could more effectively manage large datasets and new security risks, limiting innovation.

Feature extraction is crucial in encryption systems, as it identifies and encodes relevant patterns from image segments. Inadequate or poorly selected features may result in encrypted data retaining identifiable information, making it susceptible to cryptographic attacks and unauthorized access. Suboptimal feature selection can also negatively impact encryption efficiency and compromise security, exposing sensitive information to potential threats. To maintain resilience against evolving risks, it is vital to implement optimal and adaptive feature selection techniques. Encryption systems face a challenging balance between accuracy and privacy. Robust privacy measures, such as encryption or data masking, often restrict the amount of processable information, reducing the precision of data-driven tasks like pattern recognition or image classification. Conversely, maximizing accuracy may require revealing more information, increasing the risk of privacy breaches. Achieving the optimal equilibrium between data protection and system performance remains a significant challenge in machine learning-enhanced encryption.

4 Conclusion

This research explored the integration of machine learning (ML) with chaos-based encryption to address the growing challenges of secure communication for devices with limited resources, such as wearable systems. The study demonstrated how ML techniques, including neural networks, support vector machines, and reservoir computing, can enhance encryption, improve synchronization, and counter-attacks. Findings revealed that ML offers a versatile and dynamic approach to overcoming the limitations of traditional chaos-based encryption methods. Despite promising results, several challenges persist, including the need for efficient real-time processing, computational complexity, and vulnerability to adversarial attacks. The study concludes that further investigation into secure and lightweight ML-assisted encryption frameworks could lead to advancements in secure communication across various cutting-edge technologies, including wireless networks, cloud computing, and the Internet of Things.

References

1. Shafique, A., Mehmood, A., Alawida, M., Elhadeif, M., Rehman, M. U.: A fusion of machine learning and cryptography for fast data encryption through encoding high and moderate plaintext information blocks. *Multimedia Tools Appl* 1–27 (2024)
2. Panwar, K., Kukreja, S., Singh, A., Singh, K.K.: Towards deep learning for efficient image encryption. *Proc. Comput. Sci.* **218**, 644–650 (2023)
3. Feng, L., Du, J., Fu, C., Song, W.: Image encryption algorithm combining chaotic image encryption and convolutional neural network. *Electronics* **12**(16), 3455 (2023)

4. Kiran Kumar, D. A., Chauhan, M., Gaurav, M., Pimple, N. S., Vyankatesh Argiddi, R., Shikalgar, A.: Role of machine learning in enhancing image encryption techniques for cybersecurity, *IJISAE*, 2024, 12(21s), 3904–3911 (2024)
5. Li, X., Jiang, Y., Chen, M., Li, F.: Research on iris image encryption based on deep learning. Li et al. *EURASIP J Image Video Process* 2018, 126 (2018). <https://doi.org/10.1186/s13640-018-0358-7>
6. Abu-Faraj, M.A., Al-Hyari, A., Obimbo, C., Aldebei, K., Altaharwa, I., Alqadi, Z., Almanaseer, O.: Protecting digital images using keys enhanced by 2D chaotic logistic maps. *Cryptography* 7(2), 20 (2023)
7. Alsafyani, M., Alhomayani, F., Alsuwat, H., Alsuwat, E.: Face image encryption based on feature with optimization using secure crypto general adversarial neural network and optical chaotic map. *Sensors* 23(3), 1415 (2023)
8. Panwar, K., Singh, A., Kukreja, S., Singh, K.K., Shakhovska, N., Boichuk, A.: Encipher GAN: an end-to-end color image encryption system using a deep generative model. *Systems* 11(1), 36 (2023)
9. Jia, H., Cai, D., Yang, J., Qian, W., Wang, C., Li, X., Yang, S.: Efficient and privacy-preserving image classification using homomorphic encryption and chunk-based convolutional neural network. *J. Cloud Comput.* 12(1), 175 (2023)
10. Liu, L., Gao, M., Zhang, Y., Wang, Y.: Application of machine learning in intelligent encryption for digital information of real-time image text under big data. *EURASIP J. Wirel. Commun. Netw.* 2022(1), 21 (2022)
11. Kumar, B.S., Revathi, R.: An efficient image encryption algorithm using a discrete memory-based logistic map with deep neural network. *J. Eng. Appl. Sci.* 71(1), 41 (2024)
12. Tian, M., Zhang, Y., Zhang, Y., Xiao, X., Wen, W.: A privacy-preserving image retrieval scheme with access control based on searchable encryption in media cloud. *Cybersecurity* 7(1), 22 (2024)
13. Cheng, Z., Wang, W., Dai, Y., Li, L.: A high-security privacy image encryption algorithm based on chaos and double encryption strategy. *J. Appl. Math.* 2022(1), 9040702 (2022)
14. Huang, Q.X., Yap, W.L., Chiu, M.Y., Sun, H.M.: Privacy-preserving deep learning with learnable image encryption on medical images. *IEEE Access* 10, 66345–66355 (2022)
15. Abu-Faraj, M.A., Al-Hyari, A., Alqadi, Z.: A complex matrix private key to enhance the security level of image cryptography. *Symmetry* 14(4), 664 (2022)
16. Liu, Y., Cen, G., Xu, B., Wang, X.: Color image encryption based on deep learning and block embedding. *Security Commun Netw* 2022(1), 6047349 (2022)
17. Hwang, J., Kale, G., Patel, P.P., Vishwakarma, R., Aliasgari, M., Hedayatipour, A., Sayadi, H.: Machine learning in chaos-based encryption: theory, implementations, and applications. *IEEE Access* 11, 125749–125767 (2023)
18. Ding, Y., Tan, F., Qin, Z., Cao, M., Choo, K.K.R., Qin, Z.: DeepKeyGen: a deep learning-based stream cipher generator for medical image encryption and decryption. *IEEE Trans. Neural Networks Learn. Syst.* 33(9), 4915–4929 (2021)
19. Maniyath, S.R., Thanikaiselvan, V.: An efficient image encryption using deep neural network and chaotic map. *Microprocess. Microsyst.* 77, 103134 (2020)
20. Ravanna, C., Keshavamurthy, C.: A novel priority based document image encryption with mixed chaotic systems using machine learning approach. *Facta Univ. Electron. Energ.* 32(1), 147–177 (2019)

AI in Healthcare and Data Privacy

Synthetic Data Usage for Healthcare Privacy Using GENERATIVE AI



Pratyush Ranjan Sahu, Alakananda Tripathy, and Alok Ranjan Tripathy

Abstract The paper highlights the generative AI techniques for generating synthetic data for healthcare. Data plays a key role in the healthcare industry for research, treatment planning, and diagnosis. Data privacy is vital, as the patient's identity should not be revealed. Synthetic data plays an alternative by providing rich datasets without hampering the confidentiality of the patient. The basic outline of this study is to identify an alternative to creating high-quality synthetic data close to real-world healthcare data with the help of generative AI. The technique involved state-of-the-art generative models to generate synthetic datasets and evaluate their fidelity and use in different healthcare applications. The metrics for evaluation include statistical similarity to accurate data and preserving the critical pattern and the effectiveness of synthetic data in training machine learning models. The key finding of the work is that synthetic data can enhance the capability of healthcare data analytics, which provides a solution for data scarcity and privacy. The outcome is the synthetic data, which maintains the integrity of patients' data, leading to significant tools for healthcare innovation. This work also seeks data generation through the GAN model, which is helpful for assessment and possible healthcare business applications. The data generated through AI is evaluated using the KSS statistics method. The paper discusses synthetic data for different diseases and the resemblance between the real and synthetic data using the KS mean value.

Keywords GAN · Synthcity · Privacy · KS · Healthcare

P. R. Sahu

Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to Be) University, Bhubaneswar, Odisha, India

A. Tripathy (✉)

Centre for AI&ML, Siksha 'O' Anusandhan (Deemed to Be) University, Bhubaneswar, Odisha, India

e-mail: alakanandatripathy@soa.ac.in

A. R. Tripathy

Institute of Management and Information Technology, BPUT, Rourkela, Odisha, India

1 Introduction

Data drives research, informs clinical decisions, and develops innovative treatments in the rapidly evolving healthcare field. However, the sensitive nature of patient information and strict privacy regulations often limit the availability of real-world healthcare data for research and development. To address these challenges, synthetic data generation using generative AI has emerged as a promising solution.

Synthetic data is deliberately generated to imitate real-world data's statistical aspects and patterns and hide the patient's identity. Researchers can create realistic synthetic datasets with advanced generative models like Variational Auto encoders (VAEs) and Generative Adversarial Networks (GANs). Since these datasets don't include any personally identifiable information from actual people, they are safe to use and distribute.

Data, essential for research, clinical decision-making, and innovative treatments, increasingly drive the healthcare industry. However, the accessibility and use of real-world healthcare data are often hindered by stringent privacy regulations and ethical concerns surrounding patient confidentiality. These barriers limit the scope of data-driven healthcare innovations and the ability to conduct comprehensive research. This study addresses these significant challenges by exploring the potential of synthetic data generated through advanced generative AI techniques. They are used to create artificial datasets that closely resemble the patient's original data without affecting patients' synthetic data privacy. This approach helps researchers and healthcare professionals conduct a study without the risk of data breaches and ensures data protection compliance.

One of the foremost challenges in healthcare data utilization is the stringent requirement to protect patient privacy. Privacy regulations are there which impose the use and sharing of personal health information. A few are HIPAA in the United States and GDPR in Europe. While essential for protecting patient confidentiality, these regulations significantly limit the availability of real-world healthcare data for research and development. The need for a solution that can provide high-quality data without compromising privacy led to the exploration of synthetic data generation.

The ethical and legal implications of using actual patient data pose significant challenges. Researchers must navigate complex ethical approvals and legal frameworks to access and use healthcare data. Synthetic data provides an ethical alternative, allowing researchers to conduct their studies without the need to handle sensitive personal information directly, thereby simplifying the compliance process.

2 Related Work

Murtaza et al. [1] discussed Synthetic data generation of data for the healthcare domain. The growing number of publications each year reflects the increasing interest in synthetic medical data. Despite the challenges posed by privacy laws that restrict

the secondary use of accurate medical data, synthetic data offers a promising solution by providing privacy-safe alternatives. Anonymization techniques often fail to ensure comprehensive privacy for high-dimensional health data, thus shifting focus to synthetic data generation (SDG). Gonzales and Smith [2] discussed that healthcare data holds significant societal and monetary value, particularly in improving public health and fostering innovations in the AI health industry. However, the sensitive nature of primary care data poses privacy concerns, leading to reluctance to release it in many countries. This has driven interest in synthetic data, which mirrors accurate data while ensuring privacy. The paper explores key issues in synthetic data generation, such as handling real-world data complexities, time for modeling, and minimizing the similarity of actual patients to synthetic data points. It emphasizes that synthetic datasets can maintain transparency and trust with appropriate modeling approaches while offering strong privacy protection. Imtiaz [3] discussed synthetic and private healthcare data generation using the GAN model. Bayesian networks are used for synthetic data generation, as mentioned by Benedetti et al. [4], which significantly advances healthcare data analysis. Bayesian networks, known for their ability to model complex probabilistic relationships, offer a promising solution for generating realistic synthetic health data. Shung [5] emphasized the generation of synthetic data for predictive analysis and provides security to healthcare data. Figueira discusses the different evaluation methods for synthetic data generation using the GAN model [6]. The studies in [7–9] discussed generating synthetic time series data using hidden Markov and regression models.

3 Methodology

The primary goal of this research is to generate synthetic data for healthcare, which can help to improve the privacy of the patient and also help to analyze the data without compromising the patient's real identity. The architecture consists of generating the data with the help of a generative adversarial network to generate synthetic data for different diseases like breast cancer, heart disease, and diabetes.

Data Collection

Different datasets are being used for this study like Diabetic dataset, Heart disease dataset, and breast cancer dataset. The dataset for diabetes is available in the National Institute of Diabetes and Digestive and Kidney Diseases. Specifically, in the dataset, all the patients are females of Pima Indian heritage and at least 21 years of age. The dataset's different features are BMI, Insulin, age, Pregnancy, Plasma Glucose, Blood Pressure, and Skin Thickness. The following data set used is the Heart Disease Dataset obtained from four different databases—Cleveland, Hungary, Switzerland, and Long Beach V. It has a total of 76 properties, which includes the anticipated attribute. However, only 14 of them are used in the studies. The patient's heart condition is marked in the “target” feature of the dataset, which signifies that zero is no disease, whereas 1 indicates disease. The last dataset used is the Breast Cancer

Dataset, where the features that describe the properties of the cell nuclei in the image were computed from a digitized image of a breast tumor.

3.1 Preprocessing

The dataset used to generate the synthetic data is cleaned before it is used to create artificial data. From the dataset, missing values and null values are removed. The dataset should be balanced based on the target values to obtain proper synthetic data. Once the data preprocessing is performed, the dataset is passed to the generative adversarial network to generate the fake data.

3.2 Generative Adversarial Network (GAN Model)

The model uses the generative adversarial network model, a deep learning model used to generate a new dataset that matches the existing dataset, like image or numeric data. The model consists of a generator and a discriminator. The discriminator is used to evaluate whether the dataset is the original dataset or generated by the generator. The generator is responsible for generating the synthetic data samples by adding noise or some random data.

The basic architecture of the Generative Adversarial Network is discussed in Fig. 1. The architecture explains how the synthetic data is generated with the help of the generator. After the continuous training, the data randomly collected from the different data sources is provided to the generator along with some noise.

Providing data for testing and analysis. The training has two objectives: to generate synthetic data and evaluate whether the data is real or fake.

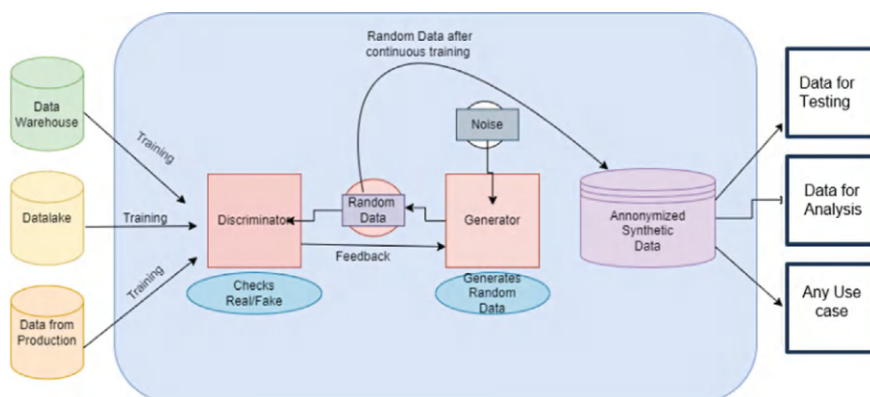


Fig. 1 Schematic layout

The fundamental objective of the training process is to find the balance between producing realistic data and the discriminator’s inability to identify it from the actual data.

The generative adversarial network uses a minimax game, where the generator tries to reduce the discriminators’ capability to identify the fake data. Table 1 depicts various methods for generating datasets.

The generator is a neural network that adds noise to the input and finds the output synthetic data. The activation function ReLU is used in the intermediate layer, and sigmoid or tanh is used in the output layer.

- The different steps of the Synthetic Data Generator involve:
- Step 1: The generator and discriminator are initialized with some weights.
 - Step 2: The discriminator is trained using accurate fake data generated by the generator.
 - Step 3: The discriminator weights are updated to improve the ability of the model to classify the data.
 - Step 4: Some random noise is added to the data in the generator by updating the weights to reduce the loss.
 - Step 5: Steps 2 and 3 are repeated until the data is generated and cannot be identified as fake.

Figure 2 explains how the synthcity is used to generate the data. The first stage is to load the data. A generator plugin is used for the training. Once the synthetic data is generated, it is evaluated based on specific metrics.

Table 1 Methods used for generating dataset

Methods	Description
Adsgan	A conditional GAN framework that generates synthetic data by hiding the identity of the actual data. It is based on the probability of re-identification by providing the combination of all the data
Pategan	The methods used in the Private Aggregation of Teacher Ensembles (PETA) framework and applied to generative adversarial networks provide differential privacy by improving the model’s performance
Ctgan	It is a conditional generative adversarial network that handles the tabular data

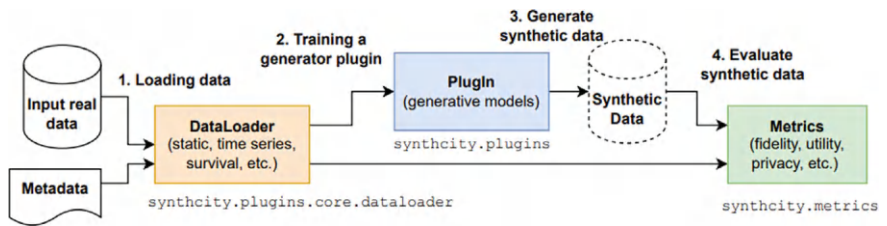


Fig. 2 Synthcity architecture used

3.3 *Synthcity*

Synthcity is required for the generation of synthetic data. It is a Python module for fake data generation. It helps generate artificial data that preserves the privacy of the patient.

It uses statistical properties of data to replicate real-world data into synthetic data and maintain the privacy of the data.

The essential features are using both tabular and time series data for generation and including numeric and categorical data. This follows a differential privacy method to maintain the security of the data.

The paper uses the generative adversarial network model to generate datasets for heart disease, breast cancer, and diabetic data.

Figure 2 explains how the synthcity works. The original dataset was used as input, basically number- or object-based data. Then, preprocessing is performed to handle the missing and null values and encode the object-based data.

Afterward, users select a model; it may be a generative adversarial network variational autoencoder or a Bayesian model.

The model then trains the data, which generates the fake data based on the statistical patterns and relationships. Differentiated privacy was used to maintain security.

The statistical similarity is measured using KL statistics. If any imbalanced data is there, then it is balanced.

4 Results and Discussion

This section discusses and presents the visualization of the results. As mentioned, the dataset is related to credit cards and was collected from Kaggle. The dataset is divided into 85% training and 15% testing samples.

4.1 *KS Statistics*

It is known as Kolmogorov Smirnov Statistics [10]. It is used to test the distribution of two values that are features of one dataset used in training and the same feature obtained in testing using different generative models.

The definition of the KS test statistic is the most significant difference between the cumulative distribution functions (also known as CDFs) of synthetic and accurate data. These cumulative distribution functions (CDFs) are known as empirical CDFs (eCDFs) in machine learning applications, typically obtained empirically from dataset samples. After calculating the KS statistic between the real and synthetic data for each feature, this function computes the mean KS statistic for all features. A lower

mean KS statistic indicates a higher similarity between the real and synthetic data distributions. The KS statistic function is shown in Eq. 1.

$$F_n(t) = \frac{\text{number of elements in the sample} < t}{n} = \frac{1}{n} \sum_{i=1}^n X_i \leq t$$

(1)

To calculate the KS statistics for each variable X, first compute the KS statistic as in Eq. 2:

$$\text{KS(Variable)} = \sup |F_{\text{real}}(\text{Var}) - F_{\text{syn}}(\text{Var})|$$

(2)

$F_{\text{real}}(\text{Var})$ and $F_{\text{syn}}(\text{Var})$ are the empirical cumulative distribution of two datasets. The mean of the KS statistics for all the variables is the sum of KS for all variables divided by the number of variables. The KS statistics for different datasets are shown in Fig. 3.

Figure 4 shows the real and synthetic data generated using Synthcity for the diabetes dataset.

Figure 6 showcases the data distribution comparison between the actual data in blue and the synthetic data in red.

The rest of the features consist of multimodal distribution, but the synthetic data is not represented correctly. Similarly, for the feature thalach, the actual data has a normal distribution compared to the synthetic data. So, the figure shows that synthetic data has accurate data trends, but the replication is complex due to the imbalanced features.

Delimiter:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFun...
1	6	148	72	35	0	33.6	0.627
2	1	85	66	29	0	26.6	0.351
3	8	183	64	0	0	23.3	0.672
4	1	89	66	23	94	28.1	0.167
5	0	137	40	35	168	43.1	2.288
6	5	116	74	0	0	25.6	0.201
7	3	78	50	32	88	31	0.248
8	10	115	0	0	0	35.3	0.134
9	2	197	70	45	543	30.5	0.158
10	8	125	96	0	0	0	0.232
11	4	110	92	0	0	37.6	0.191
12	10	168	74	0	0	38	0.537
13	10	139	80	0	0	27.1	1.441
14	1	189	60	23	846	30.1	0.398
15	5	106	72	19	175	25.8	0.587
16	7	100	0	0	0	30	0.484
17	0	118	84	47	230	45.8	0.551
18	7	107	74	0	0	29.6	0.254
19	1	103	30	38	83	43.3	0.183
20	1	115	70	30	96	34.6	0.529
21	3	126	88	41	235	39.3	0.704
22	8	99	84	0	0	35.4	0.388
23	7	196	90	0	0	39.8	0.451
24	9	119	80	35	0	29	0.263

Fig. 3 Real dataset (diabetes dataset)

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFun...
1	1	144	80	70	353	30.323912385743423	0.46951107705937767
2	7	143	39	26	458	19.68740244750032	0.6593052085875205
3	4	66	67	23	264	21.8390813573522	0.4180312509540774
4	6	31	58	46	58	44.66709948617483	0.078
5	0	196	47	38	236	9.564804758085543	0.078
6	7	66	56	31	304	22.513973587670343	0.078
7	8	96	44	31	165	22.80773769230327	0.078
8	0	107	56	0	348	24.872318487784636	0.078
9	6	79	65	13	117	23.342053214419458	0.078
10	2	156	54	44	389	24.14991156949278	0.078
11	6	120	69	38	360	27.440187432734923	0.3448998397404148
12	6	167	72	35	0	21.07914450089509	0.15216605863826171
13	7	185	63	43	254	18.19320285901064	0.22831001475993778
14	7	85	55	27	204	20.10463777834842	0.078
15	6	83	79	27	156	35.36495180151205	0.078
16	6	105	48	41	83	20.612127650132688	0.09135749809682708
17	2	74	61	29	324	18.908550883841357	0.11688249353520241
18	10	116	44	25	374	23.4134875731022	0.1489252874392739
19	7	47	59	41	308	23.28959010122547	0.078
20	1	60	65	37	497	21.642101965775897	0.078
21	3	96	64	34	395	22.002160745464685	0.07991121149976449
22	2	148	64	42	213	24.109590495118965	0.2156448451180475
23	5	102	61	29	238	28.794251742639627	0.078
24	5	83	64	48	44	22.02227865808473	0.3605737218368046

Fig. 4 Synthetic dataset (diabetes dataset)

The KS mean value is explained in Fig. 5 for different datasets like diabetes, heart disease, and breast cancer. For this, the synthetic data generated is near to the actual data. The KS mean value for the diabetes dataset depicts a moderate similarity between real and artificial data. The heart disease data contains imbalanced features, leading to a 0.46 value for KS Mean statistics. From this, it is analyzed that the synthetic data will be nearer to the actual data by doing class balancing and optimizing. Figure 6 shows the KS statistic distribution for different datasets.

Figure 7 shows the data distribution graph for different features of the heart patient dataset.

KS MEAN VALUE	
DIABETES DATASET	0.39049
HEART DISEASE DATASET	0.46801
BREAST CANCER DATASET	0.50058

Fig. 5 KS mean value for diabetes, heart disease and breast cancer datasets

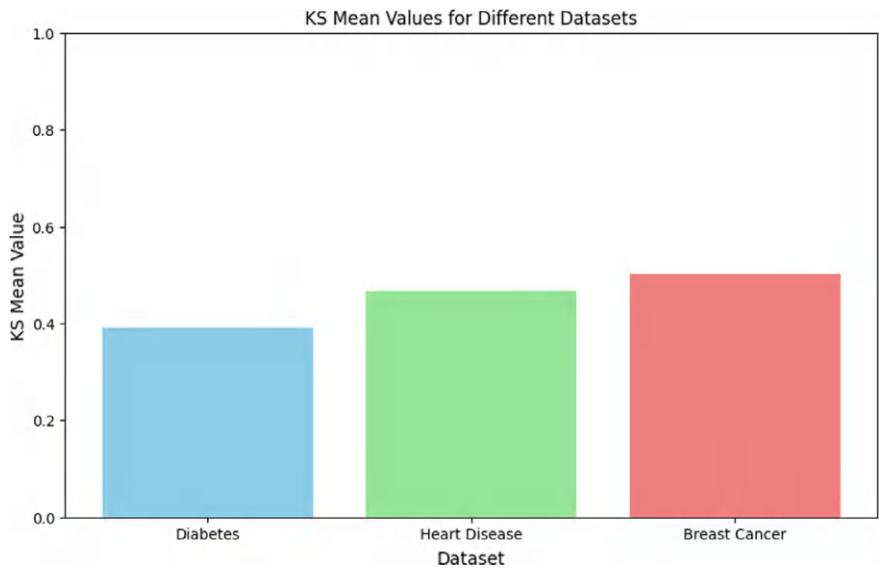


Fig. 6 Graphical representation for KS mean value for different datasets

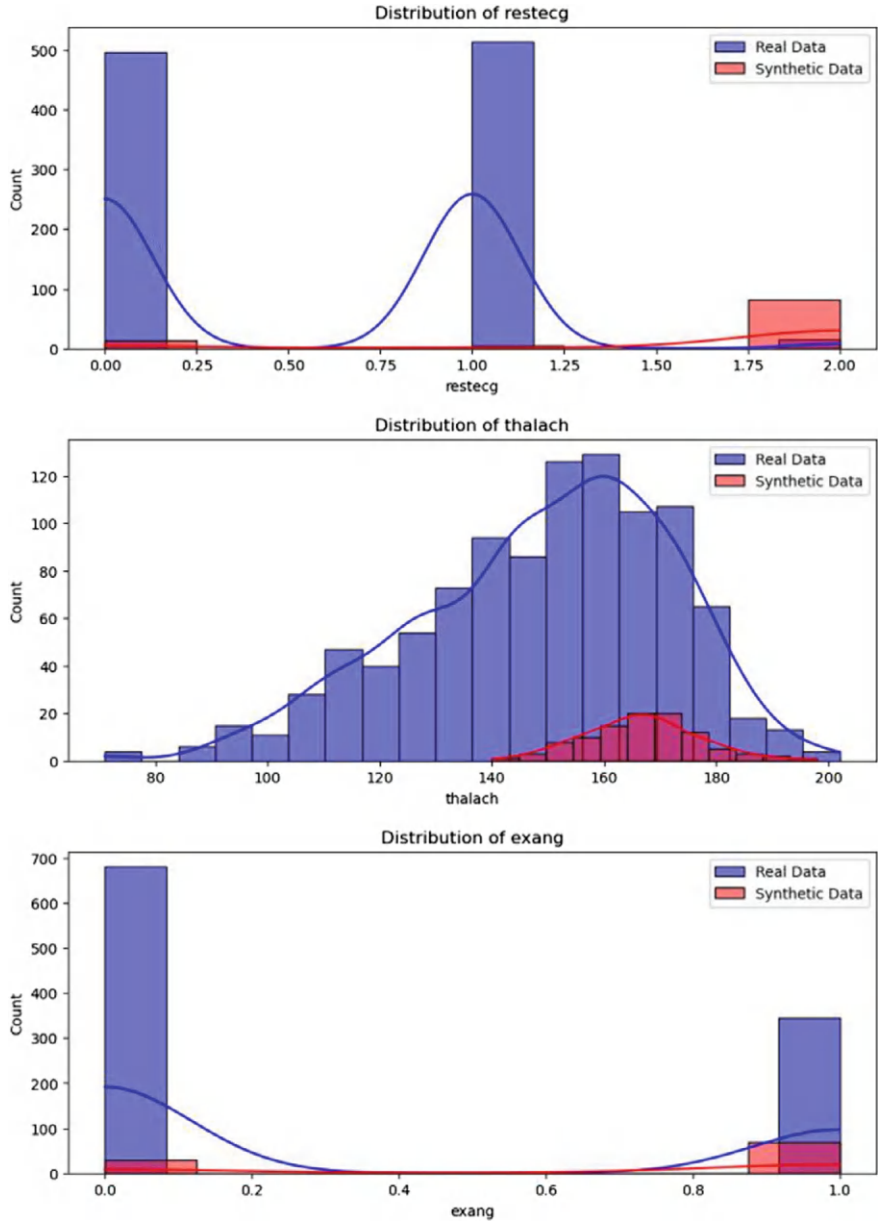


Fig. 7 Distribution graph for different features of the heart patient dataset

5 Conclusion

This study presents a comprehensive evaluation of generating synthetic data using AI, which has the potential to revolutionize healthcare research and practice. Synthetic data generation offers a robust and innovative solution by handling issues like data privacy, limited access to real-world datasets, and the inefficiencies of manual data handling. Synthetic data preserves patient privacy while enabling extensive research, complying with regulatory requirements, and reducing the risk of data breaches. Generative AI produces high-quality synthetic datasets that emulate real-world data, making valuable data more accessible and promoting inclusive healthcare research. Moreover, synthetic data simplifies ethical research by streamlining the approval process and ensuring compliance with legal standards. It also enhances the training and validation of machine learning models, contributing to advanced predictive tools that improve patient outcomes and healthcare delivery.

The paper discusses the synthetic data generated using a generative adversarial network. The data generated also maintains privacy by using differential privacy-preserving techniques. The threshold for the data generated is between 0.1 and 0.5 for diabetes and heart disease, which shows a similarity between the real and the synthetic data, but for breast cancer, the value for the KS mean is high. It may be due to imbalanced features. The value for the KS mean can be reduced by balancing the features or doing hyperparameter tuning. The application of generative AI for healthcare provides future potential. The study focuses on generating heart disease data using a generative adversarial network. The data generated almost looks like accurate data. KS statistics is used to find the accuracy of how much the fake data looks like accurate data. So, synthetic data generation using generative AI for healthcare will significantly impact the healthcare industry, improving research and maintaining patient privacy.

References

1. Murtaza, H., Ahmed, M., Khan, N.F., Murtaza, G., Zafar, S., Bano, A.: Synthetic data generation: state of the art in the healthcare domain. *Comput. Sci. Rev.* **48**, 100546 (2023). <https://doi.org/10.1016/j.cosrev.2023.100546>
2. Gonzales, A., Guruswamy, G., Smith, S.R.: Synthetic data in health care: a narrative review. *PLOS Digit. Health* **2**(1), e0000082–e0000082 (2023)
3. Imtiaz, S., Arsalan, M., Vlassov, V., Sadre, R.: Synthetic and private innovative health care data generation using GANs. In: 2021 International Conference on Computer Communications and Networks (ICCCN), July 2021, pp. 1–7. IEEE (2021)
4. de Benedetti, J., Oues, N., Wang, Z., Myles, P., Tucker, A.: Practical lessons from generating synthetic healthcare data with Bayesian networks. In: *ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020*, Ghent, Belgium, September 14–18, 2020, Proceedings, pp. 38–47. Springer International Publishing (2020)

5. Giuffrè, M., Shung, D.L.: Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *NPJ Digit. Med.* **6**(1), 186 (2023)
6. Figueira, A., Vaz, B.: Survey on synthetic data generation, evaluation methods, and GANs. *Mathematics* **10**(15), 2733 (2022). <https://doi.org/10.3390/math10152733>
7. Dahmen, J., Cook, D.: SyNSYS: a synthetic data generation system for healthcare applications. *Sensors* **19**(5), 1181 (2019). <https://doi.org/10.3390/s19051181>
8. Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., Rankin, D.: Synthetic data generation for tabular health records: a systematic review. *Neurocomputing* **493**, 28–45 (2022). <https://doi.org/10.1016/j.neucom.2022.04.053>
9. Ferreira, A., Magalhães, R., Alves, V.: Generation of synthetic data. In: *Advances in Medical Technologies and Clinical Practice Book Series*, pp. 236–261 (2022). <https://doi.org/10.4018/978-1-7998-9172-7.ch010>
10. Zhan, Y., Mechefske, C.K.: Robust detection of gearbox deterioration using compromised autoregressive modeling and Kolmogorov-Smirnov test statistic—Part I: compromised autoregressive modeling with hypothesis tests and simulation analysis. *Mech. Syst. Signal Process.* **21**(5), 1953–1982 (2007)

The Future of Healthcare Chatbots: Balancing AI Innovation with Robust Cybersecurity Practices



**Bhagyashree Shendkar, Pankaj Chandre, Ganesh Pathak,
and Madhukar Nimbalkar**

Abstract Integrating advanced artificial intelligence (AI) into healthcare chatbots has revolutionized patient care by enabling efficient, personalized, and accessible services. However, the rapid adoption of these systems underscores the critical need to balance AI innovation with robust cybersecurity measures. This paper presents a comprehensive architecture for a secure healthcare chatbot system that encompasses key components such as user interfaces, AI and natural language processing (NLP) engines, and data storage with end-to-end encryption. A dedicated security and compliance layer ensures adherence to regulations like HIPAA and GDPR while mitigating risks through multi-factor authentication and real-time threat analysis. Analytics and monitoring modules provide actionable insights and detect security incidents, while seamless integration with healthcare systems facilitates appointment scheduling, telemedicine, and medication management. Continuous learning mechanisms and model retraining enhance the chatbot's accuracy and adaptability. This framework offers a roadmap for deploying healthcare chatbots that prioritize technological advancement and data privacy, fostering trust and efficacy in patient care.

Keywords Healthcare chatbots · Artificial intelligence · Cybersecurity · Natural language processing · Electronic health records · Regulatory compliance

B. Shendkar (✉) · P. Chandre · G. Pathak · M. Nimbalkar
Department of Computer Science and Engineering, MIT School of Computing, MIT Art Design
and Technology University, Pune, India
e-mail: bhagyashree.shendkar@mituniversity.edu.in

P. Chandre
e-mail: pankaj.chandre@mituniversity.edu.in

G. Pathak
e-mail: ganesh.pathak@mituniversity.edu.in

M. Nimbalkar
e-mail: madhukar.nimbalkar@mituniversity.edu.in

1 Introduction

Chatbots for healthcare have become revolutionary tools for increasing patient participation, promoting smooth communication, and expanding access to medical care. These AI-powered solutions help with mental health assistance, prescription reminders, appointment scheduling, and symptom checks. Chatbots, which use natural language processing (NLP) and machine learning, offer 24/7 engagement and direction, greatly relieving the workload of medical staff while guaranteeing individualized patient care [1]. Its incorporation into healthcare workflows reflects a growing reliance on technology to provide scalable and practical solutions to fulfill patient demands.

Despite its promise, healthcare chatbots have significant obstacles to overcome, especially regarding data security and moral AI applications. Because chatbots manage private patient data, including medical records and personal information, they are often the focus of data breaches and hacks. Their acceptance is made more difficult by worries about AI biases, data exploitation, and lack of transparency [2]. Resolving these problems is essential to fostering user trust and guaranteeing adherence to strict healthcare laws like HIPAA and GDPR. Therefore, the objective is to develop AI-driven chatbot systems that are safe, reliable, and morally sound while giving equal weight to innovation and data security [3].

This study aims to present a thorough analysis of the state of healthcare chatbots today, highlighting the developments in AI technologies that underpin their operation. Additionally, it looks for and evaluates the cybersecurity issues related to these systems, providing information about possible weaknesses and dangers. The study will also look at best practices and preventative actions to protect chatbot systems while preserving the harmony between patient trust and technology progress. By doing this, this initiative hopes to support the future of digital healthcare by helping to establish strong, safe, and moral healthcare chatbot systems.

2 Background and Literature Review

A. Evolution of Chatbots in Healthcare: Early Implementations versus Modern AI-Driven Systems

Healthcare chatbots have experienced a substantial change from rule-based systems to AI-driven platforms. Their reliance on prewritten scripts and decision trees constrained the versatility and range of early chatbots to address various user queries. These systems mostly performed simple tasks like responding to often requested questions or offering static health information [4]. Thanks to developments in artificial intelligence (AI), contemporary chatbots today use machine learning (ML) and natural language understanding (NLU) to deliver contextually aware, dynamic, and personalized responses. These days, they play a crucial role in virtual health aides, which can help with diagnosis, symptom monitoring, mental health assistance, and

managing chronic illnesses. In addition to increasing their capabilities, this growth has brought out new difficulties, particularly in guaranteeing data security and moral AI practices.

B. Role of AI in Chatbot Development: Natural Language Processing (NLP), Machine Learning, and Deep Learning

The development of chatbots has been transformed by AI technologies, which have improved their comprehension and reaction to intricate human language. Chatbots can interpret user input, extract meaning, and produce contextually relevant responses thanks to natural language processing (NLP). The quality of interactions is further enhanced by methods like named entity recognition (NER) and sentiment analysis. While deep learning, which uses neural networks, enables chatbots to handle unstructured data, such as free-text inputs, machine learning allows them to learn from past data and gradually improve their performance [5]. These features have enabled chatbots to provide conversational experiences, adaptive responses, and more precise diagnoses. Strong design and validation procedures are necessary since integrating AI entails concerns like algorithmic bias and susceptibility to hostile attacks.

C. Overview of Cybersecurity in Healthcare: Key Threats, Challenges, and Regulations (e.g., HIPAA, GDPR)

Because chatbots manage sensitive patient data, cybersecurity in the healthcare industry is a significant concern. Unauthorized access, ransomware attacks, and data breaches are substantial risks resulting in identity theft, monetary loss, or jeopardized patient safety. An additional degree of complexity is introduced by the growing usage of AI, which carries hazards, including training data leakage and hostile manipulation of AI models [6]. Strict criteria for data protection are established by regulatory frameworks such as GDPR in the European Union and HIPAA in the United States, which emphasize secure processing, transport, and storage. Respecting these rules is essential to keeping patients' trust and avoiding legal trouble. Addressing these cybersecurity issues with proactive measures and standard compliance is crucial as chatbots become increasingly integrated into healthcare delivery.

The study [7] explored the challenges and solutions in implementing AI-based healthcare systems while preserving patient privacy. Strict privacy laws, a lack of curated datasets, and problems with data harmonization are the main obstacles to the broad use of AI in healthcare. The paper emphasizes several privacy-preserving strategies, including Secure Multiparty Computation, Federated Learning, and Homomorphic Encryption. These techniques seek to protect private patient information while facilitating cross-institutional cooperation in creating AI models. Along with discussing several privacy threats that can jeopardize AI systems, such as membership inference and model inversion, the study recommends future research directions to improve healthcare AI's resilience, scalability, and ethical compliance.

The study [8] introduced a BERT-based medical chatbot to improve healthcare communication by leveraging natural language understanding (NLU). It tackles issues that conventional medical chatbots encounter, like deciphering intricate medical jargon and giving precise answers. The chatbot can efficiently manage

multi-turn discussions thanks to BERT's bidirectional architecture, which improves contextual comprehension. When tested on datasets such as MIMIC-III and PubMed, the chatbot demonstrated strong performance metrics, including 98% accuracy, 97% precision, and a 97% AUC-ROC score. These outcomes show that it can forecast illnesses and provide tailored medical guidance. This chatbot represents a significant advancement in improving patient care and healthcare accessibility, even though computational demands and data privacy issues still exist.

The study [9] explored the evolving cybersecurity landscape in the age of AI, highlighting both opportunities and challenges. AI improves cybersecurity using predictive analytics, automated responses, and better threat detection. However, it creates new vulnerabilities that conventional defenses could find difficult to counter, like adversarial attacks and AI-enabled cyber threats. For AI to be implemented responsibly, ethical factors, including privacy, bias, and accountability, are essential. Industry standards and regulatory frameworks are crucial for guaranteeing compliance and transparency. Strong governance, interdisciplinary cooperation, and ongoing staff training are necessary to bolster defenses. Securing AI-driven systems requires a comprehensive strategy that blends theoretical understanding with valuable tactics.

The study [10] discussed the transformative role of artificial intelligence (AI) in healthcare, emphasizing the need to balance technology and human compassion. AI helps with better diagnosis, more individualized treatment plans, and more efficient healthcare operations like medical imaging and appointment scheduling. It expands access to neglected areas by enabling telemedicine and predictive analytics. Nonetheless, difficulties like algorithmic prejudice, data privacy issues, technical mistakes, and financial constraints are emphasized. Integrating AI requires careful consideration of ethical problems, such as openness and patient confidence. To ensure AI enhances human knowledge rather than replaces it, the text strongly emphasizes interdisciplinary collaboration, digital literacy, and patient-centered approaches. A future of effective and compassionate healthcare can be achieved by fusing technology and people.

The study [11] explored the transformative role of AI in healthcare, addressing its applications in hospitals and clinics while focusing on clinical decision-making, hospital operations, diagnostics, and patient care. AI improves the precision of diagnoses, streamlines hospital operations, and uses predictive analytics to tailor treatments. It facilitates telemedicine, remote patient monitoring, and medical imaging, opening up access to healthcare for more people. Case studies demonstrate how AI can transform medical procedures by assisting in the early detection of diseases and developing new drugs. Along with the significance of interdisciplinary cooperation, ethical and legal issues, including data protection, algorithmic prejudice, and adherence to regulations like HIPAA, are highlighted. The article promotes patient-centered, egalitarian, and responsible integration to optimize AI's advantages in healthcare systems.

3 Current State of Healthcare Chatbots

A. Use Cases

Healthcare chatbots are revolutionizing patient care by automating various processes and enhancing accessibility. Some of the most prominent use cases include:

Appointment Scheduling: Chatbots improve efficiency and lessen administrative strain by streamlining the scheduling process. Conversational interfaces allow patients to arrange, reschedule, or cancel appointments, reducing human error and enhancing clinic efficiency.

Symptom Checking: Many healthcare chatbots employ artificial intelligence (AI) to evaluate user-provided symptoms and suggest possible diagnoses or the best courses of action, such as seeing a doctor, getting tested, or getting emergency care [12]. This use case improves instant access to medical guidance, particularly in isolated locations.

Mental Health Support: AI-powered chatbots can give therapeutic exercises, coping strategy guidance, and instant emotional support. These systems are especially helpful in lowering the stigma associated with mental illness and provide round-the-clock access to services for mental health care, which helps close accessibility gaps.

Chronic Disease Management: Chatbots can provide daily monitoring, prescription reminders, and behavioral advice to patients with diabetes, hypertension, or asthma. Chatbots contribute to better long-term disease management and increased patient adherence to treatment plans by offering tailored interventions.

B. Technologies in Use

The effectiveness of healthcare chatbots is mainly dependent on the integration of advanced technologies:

AI Models (GPT, BERT): Chatbot systems are powered by natural language processing (NLP) models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), which allow them to comprehend and produce responses that resemble those of a human [13, 14]. These models enable chatbots to understand and analyze user input, respond with pertinent context, and conduct logical, customized discussions.

Chatbot Platforms: Healthcare chatbots are often designed, implemented, and maintained using platforms such as Microsoft Bot Framework, Dialogflow, and Rasa. These systems guarantee scalability and flexibility for developers while integrating backend databases, user interfaces, and NLP models to produce smooth user experiences.

APIs: Chatbots can access external systems and data sources, including scheduling software, diagnostic tools, and electronic health records (EHRs), thanks to APIs (Application Programming Interfaces). Richer interactions and improved functionality are made possible by chatbots' ability to pull real-time data, provide well-informed recommendations, and link with current healthcare infrastructures through APIs.

C. Limitations and Challenges

While healthcare chatbots offer significant benefits, several limitations and challenges must be addressed for them to reach their full potential:

Data Privacy: The sensitive and regulated nature of healthcare data is a significant challenge when incorporating chatbots into clinical settings. It is crucial to ensure chatbots abide by data protection laws like the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA). Another concern is protecting patient data from online attacks since security lapses could have disastrous results.

AI Bias: Healthcare chatbot AI models may be biased by the data they are trained on, which could result in unfair or erroneous treatment suggestions. For example, a chatbot trained on a non-diverse dataset might provide subpar healthcare advice for particular groups [15]. To reduce this risk, it is crucial to make sure that training datasets are inclusive and diverse.

Technical Constraints: Even with AI's advances, many healthcare chatbots still have trouble answering complicated medical questions and might not always give thorough or correct answers. Errors might result from NLP models' inability to comprehend complex medical terminology or handle uncommon illnesses, undermining user efficacy and trust.

User Trust: Particularly regarding health-related advice, patients might be reluctant to trust AI-driven chatbots completely. Creating transparent systems that clearly explain how the AI operates, what data it consumes, and the safety precautions is just as necessary as giving accurate and dependable information to gain users' trust. To keep the chatbot's knowledge base up to-date and functional, ongoing training and updates are also required.

4 Cybersecurity Challenges in Ai-Driven Chatbots

A. Data Vulnerabilities: Risks of Breaches, Unauthorized Access, and Data Misuse

AI-driven healthcare chatbots depend on private patient data to provide individualized services like diagnosis, treatment suggestions, and symptom checks. This data has serious privacy concerns, including health history, personal information, and medical records. Weak encryption, unsafe storage, and inadequate access controls are the leading causes of the vulnerabilities. Insufficient data protection during transmission or weak authentication procedures might lead to unauthorized access. Identity theft, insurance fraud, or unapproved medical procedures might arise from data misuse once it has been compromised. Privacy concerns may also be heightened by the possibility of data leaking from third-party integrations for chatbots that gather massive databases. Robust data protection measures like multi-factor authentication, end-to-end encryption, and stringent access control procedures are crucial to lessen these vulnerabilities.

B. Threat Landscape: Malware, Phishing, and Adversarial Attacks Targeting Chatbots

As AI chatbots become more prevalent in healthcare settings, they become prime targets for various cybersecurity threats. Malware poses a serious concern since it can be introduced into the chatbot system to exploit flaws, obtain unauthorized access, or interfere with normal operations. Phishing attacks may use chatbots to pose as medical professionals, tricking patients into divulging private information or clicking on harmful links. Additionally, there is rising concern about adversarial assaults, in which malevolent individuals control the chatbot's AI algorithms to give misleading information or make wrong predictions, potentially harming patients. Attackers might, for instance, trick a symptom checker into giving false medical advice by using adversarial examples. AI models must be resilient to these risks, and there must be malware detection tools, user authentication procedures, and ongoing monitoring.

C. Ethical Concerns: Transparency, Accountability, and Informed Consent in AI Usage

In the context of AI-driven healthcare chatbots, ethical concerns play a critical role in ensuring technology's safe and fair use. Transparency in how chatbots collect, store, and process personal health data is essential to build trust with users. Patients need to know how their information will be used, whether it will be shared with third parties, and how it will be protected. Another issue is accountability, mainly when chatbots assist decision-making or offer medical advice. Who bears the blame if a chatbot provides inaccurate or dangerous information that affects a patient's health? To avoid ethical and legal problems, clear criteria for accountability are required, covering the roles of developers, healthcare providers, and AI systems. Finally, when incorporating AI chatbots into healthcare, informed permission is essential. Patients must agree to use their data and be informed that they communicate with an AI rather than a human. To guarantee the responsible use of healthcare chatbots, addressing ethical issues, such as ensuring appropriate permission procedures and educating users on AI's potential and constraints, is imperative.

5 Proposed Methodology

Figure 1 illustrates the architecture of a Healthcare Chatbot System designed to interact with users and securely manage patient data. The User Interface allows communication through chat, voice, mobile apps, or web portals. The AI and NLP Engine uses intent recognition, contextual understanding, and machine learning models to interpret inputs and provide personalized responses. The Security and Compliance Layer ensures secure data access with multi-factor authentication, GDPR, HIPAA compliance, and end-to-end data encryption. The Analytics and Monitoring component tracks health insights, usage, and security incidents,

providing real-time threat analysis. Healthcare System Integration synchronizes patient data with systems like doctor’s dashboards, appointment scheduling, and telemedicine services. Data Storage and Integration manages patient profiles, medical history, and electronic health records (EHR). Finally, the AI Model Training and Feedback ensures continuous learning and model improvement.

User Interface: The user interface (UI) is the entry point for patient interactions with the healthcare chatbot system. It provides multiple platforms for communication, including text-based chat interfaces, voice interactions, mobile apps, and web portals. Patients input their queries or requests through these channels, which are then forwarded to the AI and NLP Engine for processing. This module is designed to be user-friendly, ensuring that patients of all ages and technical backgrounds can easily engage with the system. It acts as the bridge between the user and the backend systems, delivering AI-generated responses.

AI and NLP Engine: The AI and NLP Engine is the heart of the chatbot system. It utilizes advanced Natural Language Processing (NLP) techniques to understand and process user input. The engine can recognize the intent behind the queries, understand the context, and generate personalized responses using machine learning models. The engine retrieves relevant data from the Data Storage and Integration module, ensuring

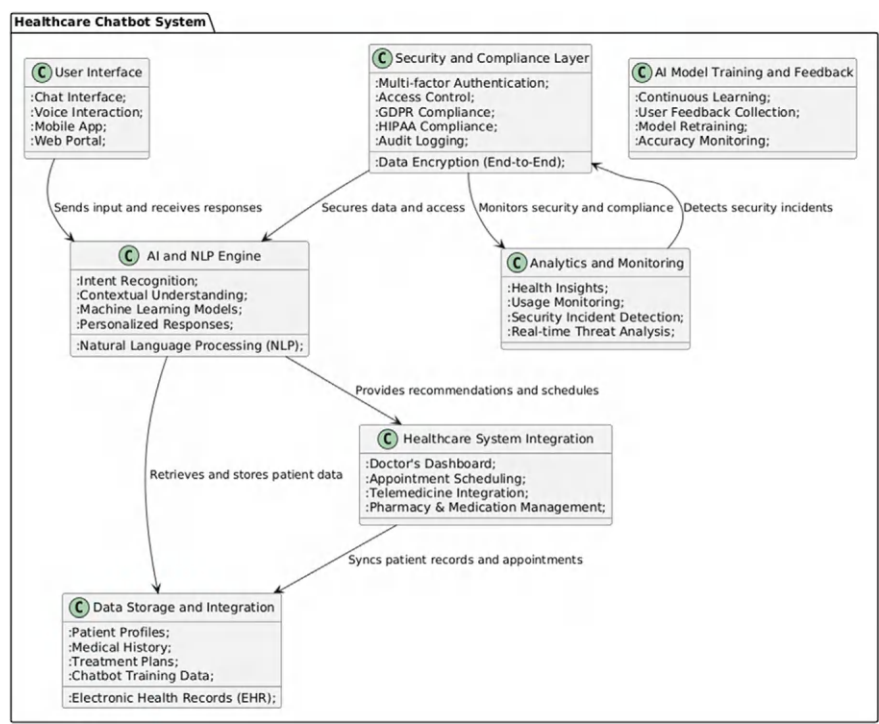


Fig. 1 An architecture of a healthcare chatbot system designed to interact with users and securely manage patient data

that responses are tailored to the patient's medical history, treatment plans, and other pertinent information. This module ensures the chatbot can converse intelligently and provide meaningful assistance, including health advice, appointment scheduling, and medication reminders.

Data Storage and Integration: This module stores all critical patient-related data, including medical history, treatment plans, patient profiles, and Electronic Health Records (EHR). It also maintains the necessary data to train AI models, such as previous chatbot interactions and user feedback. The Data Storage and Integration module ensures that all data is securely stored, managed, and easily accessible when needed by the AI and NLP Engine for generating responses. It also plays a key role in syncing patient records with other healthcare systems, ensuring real-time appointments and health data updates, which are then shared with the Healthcare System Integration module.

Healthcare System Integration: The Healthcare System Integration module connects the chatbot system to existing healthcare infrastructure, such as doctor dashboards, appointment scheduling systems, telemedicine platforms, and pharmacy/medication management systems. This ensures seamless integration between the chatbot and healthcare workflows. For instance, the chatbot can schedule appointments, manage telemedicine sessions, and access medication-related information. The integration ensures that patient data retrieved from the Data Storage and Integration module is in sync with healthcare providers' systems, facilitating better coordination and care management.

Security and Compliance Layer: To ensure that the system adheres to regulatory requirements such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation), the Security and Compliance Layer incorporates robust security measures. This includes multi-factor authentication, data encryption (end-to-end), access control, audit logging, and continuous monitoring. This layer ensures that all patient data and interactions are secure and comply with relevant privacy laws. It also prevents unauthorized access to sensitive information, safeguarding patient confidentiality while maintaining trust in the system.

Analytics and Monitoring: The Analytics and Monitoring module tracks the performance of the chatbot system, providing insights into health trends, system usage, and the detection of security incidents. Continuously monitoring these aspects ensures the system operates efficiently and securely. It helps identify potential vulnerabilities and areas for improvement in the chatbot's functionality. Furthermore, the module supports real-time threat analysis to detect suspicious activities or security breaches, which are then escalated to the Security and Compliance Layer for mitigation. This module is essential for maintaining the system's operational integrity.

AI Model Training and Feedback: The AI Model Training and Feedback module focuses on improving the chatbot's AI capabilities. It gathers user feedback, tracks the accuracy of responses, and retrains the system's machine-learning models to improve performance. This continuous learning process ensures that the chatbot remains up-to-date with the latest medical knowledge and user preferences. The module enhances the user experience by refining the chatbot's ability to provide relevant, timely, and

personalized information, improving its usability and effectiveness in healthcare contexts.

6 Proposed Security Measures for Chatbot Systems

A. Secure Data Handling: Encryption, Anonymization, and Secure Storage

To ensure the confidentiality and integrity of sensitive patient data, healthcare chatbots must adopt secure data handling practices. Data encryption guarantees unauthorized parties cannot read patient information in transit or at rest. Strong encryption methods like TLS (Transport Layer Security) for safe communication between chatbots and healthcare systems and AES (Advanced Encryption Standard) for data storage can help achieve this. Personally identifying information (PII) should be eliminated from chatbot data by using anonymization techniques. Working with big datasets for AI training helps safeguard patient privacy [16]. Furthermore, secure storage options are required to protect data from possible breaches, such as cloud services with access control policies and encrypted databases. These safeguards protect private health data and guarantee that healthcare chatbots adhere to privacy laws.

B. AI Model Security: Adversarial Training, Secure Algorithms, and Bias Mitigation

Healthcare chatbots rely on complex AI models vulnerable to adversarial attacks, where malicious actors might manipulate inputs to deceive the model. Adversarial training, in which the chatbot's AI model is exposed to possible threats during the training phase, is essential to reducing these risks since it allows the chatbot to identify and react to manipulated inputs [17]. Additionally, using safe algorithms guarantees that the AI system is resistant to assaults that try to take advantage of flaws in the model. Bias mitigation must also be addressed to ensure the chatbot functions justly and equally across various patient demographics. Biased AI algorithms that produce inaccurate diagnoses or recommendations may disproportionately affect specific populations. Healthcare chatbots can lessen prejudice and increase the reliability of their AI systems by employing strategies like fairness limits, data diversity, and transparent algorithms.

C. Regulatory Compliance: Alignment with Healthcare Data Protection Standards (HIPAA, GDPR)

Healthcare chatbots must comply with stringent data protection regulations to ensure patient privacy and the legal use of healthcare data. The Health Insurance Portability and Accountability Act, or HIPAA, establishes guidelines for the safety of patient health data in the United States and mandates that chatbots use suitable security measures such as audit trails, access limits, and encrypted data transfer [18]. Similarly, the General Data Protection Regulation (GDPR) strongly emphasizes patient

permission, the right to access data, and the right to be forgotten when regulating the gathering and use of personal data inside the European Union. While designing chatbots, these rules must be considered to guarantee that patient data is handled securely and transparently. Compliance avoids legal issues and increases patient trust by showcasing a dedication to patient rights and data protection.

D. User Education: Promoting Awareness Among Patients and Healthcare Providers

One of the often-overlooked aspects of securing healthcare chatbot systems is user education. Both patients and healthcare providers must understand how to use chatbot systems safely and securely. This involves teaching patients how to disclose private information, spot phishing scams and fraudulent bots, and their rights to privacy and data access [10]. Education for healthcare professionals should include safe patient interaction management, accurate patient identity verification, and knowledge of potential cybersecurity risks. Training sessions, thorough user manuals, and designing the chatbot with security best practices—like multi-factor authentication for sensitive tasks—can all help achieve this. Because human error or ignorance can undermine even the most substantial technical safeguards, a knowledgeable user base is crucial to healthcare chatbots' overall security and effectiveness.

7 Future Trends and Opportunities

A. Innovations in AI for Chatbots: Real-Time Analytics, Emotion Detection, and Adaptive Learning

Healthcare AI chatbots are developing to provide more complex and individualized patient interactions. These chatbots can continuously monitor patient discussions and medical data thanks to real-time analytics, giving prompt insights into patient conditions and possible hazards [19]. Emotion detection can improve chatbot interactions by examining patients' tone, language, and context to gauge their emotional states—which might be vital for mental health treatment or empathic reactions in stressful situations. Additionally, chatbots will be able to learn and get better with every encounter thanks to adaptive learning, which will let them tailor their responses and medical recommendations to each patient's unique preferences and histories. Chatbots may become more precise, responsive, and able to deliver more specialized care due to AI advancements.

B. Integration with Emerging Technologies: Blockchain for Secure Data Transactions, Quantum Computing for Advanced Encryption

New technologies have a lot of promise to improve healthcare chatbot security and effectiveness. Blockchain technology can be used to make sure that patient data is handled in a way that is safe, open, and impenetrable [20]. Blockchain can help consumers keep control of their medical information by decentralizing data storage

while granting healthcare professionals safe, authorized access. Although it is still in its infancy, quantum computing has the potential to completely transform chatbot security by providing sophisticated encryption methods that are impervious to future cyberattacks. Sensitive health data might be protected by quantum-powered encryption techniques, making it nearly hard for hackers to access data, even with powerful computers. These innovations can potentially significantly improve the productivity and reliability of AI-powered healthcare chatbots.

C. Enhanced User Trust: Building Transparency and Ethical AI Practices

A key component of using AI-powered chatbots in healthcare is user trust. It is crucial to guarantee openness in how these chatbots decide and handle private data. Patients must understand how healthcare providers use, store, and share data. Chatbots can increase user confidence by implementing explicit data handling guidelines and providing clear, understandable explanations of AI decision-making procedures. Furthermore, ethical AI methods will be essential for addressing accountability, bias, and discrimination[21]. Developers and organizations must create and adhere to moral norms to prevent AI systems from reinforcing unfair prejudices and preserve human oversight in crucial decision-making situations.

D. Scalability and Global Adoption: Addressing Challenges in Resource-Limited Settings

The key to the widespread adoption of healthcare chatbots is their scalability. There is an urgent need to create chatbot systems that can function successfully in environments with limited resources, even though big hospitals and well-funded healthcare systems can incorporate cutting-edge AI solutions. In underprivileged areas, obstacles, including poor infrastructure, a lack of technological know-how, and restricted internet connectivity, might make it challenging to implement AI-powered chatbots. However, these obstacles might be overcome by developments in offline functionality, lightweight chatbot models, and low-bandwidth optimizations. Democratization of healthcare services may also result from the affordability of AI solutions, which could increase access to healthcare for people in low-income nations. Developers can guarantee that healthcare chatbots are helpful in various international situations by creating flexible and adaptive chatbot systems.

8 Case Studies and Examples

A. Successful Implementations: Real-World Healthcare Chatbot Systems with Robust Security

Several healthcare institutions have effectively used AI-driven chatbots in recent years, focusing on security and innovation. An AI-powered chatbot, for instance, is integrated into the Mayo Clinic's "Mayo Clinic Care Companion" to help with patient engagement by sending out appointment reminders and individualized health

advice[22]. By HIPAA compliance guidelines, this system uses secure data storage techniques and end-to-end encryption to safeguard private patient information. Another example is Babylon Health, which uses AI chatbots to offer virtual health consultations. Babylon uses sophisticated machine learning models to comprehend symptoms and provide medical recommendations. The platform has implemented a thorough security strategy to protect patient interactions, including strong authentication procedures and encrypted communication routes. Furthermore, the chatbot helps reduce the risks associated with data breaches by integrating with electronic health records (EHRs) and protecting data privacy through secure API connections. These examples demonstrate how AI-powered healthcare chatbots may effectively combine cybersecurity best practices with healthcare innovation, guaranteeing users' security and usefulness.

B. Lessons Learned: Challenges Faced and How They Were Mitigated

Even though using AI chatbots in healthcare has been advantageous, several issues have been resolved. Making sure their chatbot could handle private patient data without jeopardizing security was one of the significant challenges the Mayo Clinic faced. More advanced encryption techniques and multi-factor authentication for user verification were implemented due to early worries regarding data encryption during transit. Due to problems with language and contextual awareness, the Babylon Health chatbot had trouble sustaining the accuracy of AI-driven diagnosis [23–26]. This was lessened by incorporating feedback loops, enabling human medical professionals to step in when needed and consistently enhancing the chatbot's natural language processing algorithms. Furthermore, user trust is a problem many healthcare chatbots face because individuals frequently hesitate to divulge their medical information to automated systems. Transparency in data usage and storage regulations was a top priority for Babylon Health and Mayo Clinic in solving this issue. Users were reassured that their data was managed safely through explicit consent forms and real-time notifications. Significant legal and regulatory adherence issues were also brought on by GDPR and HIPAA compliance. Still, these were resolved by incorporating legal counsel into the development process and routinely checking chatbot systems for compliance. These examples show how careful design, constant monitoring, and continual user training may help healthcare chatbots confront and overcome security and functional challenges [27, 28].

9 Conclusion

Integrating AI-powered healthcare chatbots offers transformative potential to enhance patient care, improve operational efficiency, and ensure accessibility. However, this advancement must be met with equally robust cybersecurity measures to safeguard sensitive patient data and comply with regulatory standards. The proposed chatbot architecture emphasizes a secure and efficient system design, incorporating multi-factor authentication, end-to-end encryption, and real-time threat

analysis to mitigate risks. Seamless integration with healthcare systems ensures effective scheduling, telemedicine, and medication management, while continuous learning mechanisms enhance adaptability and accuracy. This framework balances innovation with security by addressing challenges such as data privacy, system scalability, and compliance. The future of healthcare chatbots relies on striking this equilibrium, fostering trust, and ensuring ethical AI deployment to support the evolving needs of patients and providers.

References

1. Zhuo, R., Huffaker, B., Claffy, K.C., Greenstein, S.: The impact of the general data protection regulation on internet interconnection **45**(2) (2021)
2. Ravindar, K., Gupta, M., Abdul-Zahra, D.S., Subhashini, K., Maiti, N., Chawla, R.: Healthcare chatbots with NLP and cybersecurity: safeguarding patient data in the cloud. In: International Conference on Artificial Intelligence for Innovations in Healthcare Industries ICAIHI 2023, vol. 1, May 2023, pp. 1–7 (2023). <https://doi.org/10.1109/ICAIIHI57871.2023.10489713>
3. Wang, X., Wu, Y.C.: Balancing Innovation and Regulation in the Age of Generative Artificial Intelligence, vol. 14 (2024)
4. Binhammad, M., Alqaydi, S., Othman, A., Abuljadayel, L.H.: The role of AI in cyber security: safeguarding digital identity. *J. Inf. Secur.* **15**(02), 245–278 (2024). <https://doi.org/10.4236/jis.2024.152015>
5. Chen, Y., Esmaeilzadeh, P.: Generative AI in Medical Practice: In-Depth Exploration of Privacy and Security Challenges. *J. Med. Internet Res.* **26**(1), e53008 (2024). <https://doi.org/10.2196/53008>
6. Hirani, R., et al.: Artificial intelligence and healthcare: a journey through history, present innovations, and future possibilities. *Life* **14**(5), 557 (2024). <https://doi.org/10.3390/life14050557>
7. Khalid, N., Qayyum, A., Bilal, M., Al-Fuqaha, A., Qadir, J.: Privacy-preserving artificial intelligence in healthcare: techniques and applications. *Comput. Biol. Med.* **158**, 106848 (2023). <https://doi.org/10.1016/j.combiomed.2023.106848>
8. Babu, A., Boddu, S.B.: BERT-based medical chatbot: enhancing healthcare communication through natural language understanding. *Explor. Res. Clin. Soc. Pharm.* **13**, 100419 (2024). <https://doi.org/10.1016/j.rcsop.2024.100419>
9. Familoni, B.T.: Cybersecurity challenges in the age of AI: theoretical approaches and practical solutions. *Comput. Sci. IT Res. J.* **5**(3), 703–724 (2024). <https://doi.org/10.51594/csitrj.v5i3.930>
10. Sachdeva, C., Grover, V.: Balancing technology and humanity: perspectives on AI in healthcare. In: Analyzing Explainable AI in Healthcare and the Pharmaceutical Industry, pp. 1–12 (2024). <https://doi.org/10.4018/979-8-3693-5468-1.ch001>
11. Maleki Varnosfaderani, S., Forouzanfar, M.: The role of AI in hospitals and clinics: transforming healthcare in the 21st century. *Bioengineering* **11**(4), 1–38 (2024). <https://doi.org/10.3390/bioengineering11040337>
12. Li, Y.H., Li, Y.L., Wei, M.Y., Li, G.Y.: Innovation and challenges of artificial intelligence technology in personalized healthcare. *Sci. Rep.* **14**(1), 1–9 (2024). <https://doi.org/10.1038/s41598-024-70073-7>
13. Satpathy, S., Mangla, M., Sharma, N., et al.: Predicting mortality rate and associated risks in COVID-19 patients. *Spat. Inf. Res.* **29**, 455–464 (2021). <https://doi.org/10.1007/s41324-021-00379-5>
14. Satpathy, S., Nandan Mohanty, S., Chatterjee, J.M., Swain, A.: Comprehensive claims of AI for healthcare applications-coherence towards COVID-19. In: Nandan Mohanty, S., Saxena,

- S.K., Satpathy, S., Chatterjee, J.M. (eds) Applications of Artificial Intelligence in COVID-19. Medical Virology: From Pathogenesis to Disease Control. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-7317-0_1
15. Boudierhem, R.: Shaping the future of AI in healthcare through ethics and governance. *Humanit. Soc. Sci. Commun.* **11**(1), 1–12 (2024). <https://doi.org/10.1057/s41599-024-02894-w>
 16. Gadekar, B.P., Hiwarkar, T.: From business objectives to analytics and machine learning solutions: a framework for conceptual modeling. *Int. J. Adv. Res. Sci. Commun. Technol.* **3**(1), 277–285 (2023). <https://doi.org/10.48175/ijarsct-7876>
 17. Walter, Y.: Managing the race to the moon: Global policy and governance in Artificial Intelligence regulation—A contemporary overview and an analysis of socioeconomic consequences. *Discov. Artif. Intell.* **4**(1), 14 (2024). <https://doi.org/10.1007/s44163-024-00109-4>
 18. Amjad, A., Kordel, P., Fernandes, G.: A review on innovation in healthcare sector (telehealth) through artificial intelligence. *Sustainability* **15**(8), 1–24 (2023). <https://doi.org/10.3390/su15086655>
 19. Pathak, G.R., Patil, S.H.: Mathematical model of security framework for routing layer protocol in wireless sensor networks. *Phys. Procedia* **78**, 579–586 (2016). <https://doi.org/10.1016/j.procs.2016.02.121>
 20. Damre, S.S., Shendkar, B.D., Kulkarni, N., Chandre, P.R., Deshmukh, S.: Smart healthcare wearable device for early disease detection using machine learning. *Int. J. Intell. Syst. Appl. Eng.* **12**(4s), 158–166 (2024)
 21. Dhotre, D., Chandre, P.R., Khandare, A., Patil, M., Gawande, G.S.: The rise of crypto malware: leveraging machine learning techniques to understand the evolution, impact, and detection of cryptocurrency-related threats. *Int. J. Recent Innov. Trends Comput. Commun.* **11**(7), 215–222 (2023). <https://doi.org/10.17762/ijritcc.v11i7.7848>
 22. Kotwal, J., Kashyap, D.R., Pathan, D.S.: Agricultural plant diseases identification: from traditional approach to deep learning. *Mater. Today Proc.* **80**, 344–356 (2023). <https://doi.org/10.1016/j.matpr.2023.02.370>
 23. Pathak, G.R., Premi, M.S.G., Patil, S.H.: LSSCW: a lightweight security scheme for cluster based wireless sensor network. *Int. J. Adv. Comput. Sci. Appl.* **10**(10), 448–460 (2019). <https://doi.org/10.14569/ijacsa.2019.0101062>
 24. Mohapatra, S., Satpathy, S., Mohanty, S.N.: A comparative knowledge base development for cancerous cell detection based on deep learning and fuzzy computer vision approach. *Multimed. Tools Appl.* **81**, 24799–24814 (2022). <https://doi.org/10.1007/s11042-022-12824-0>
 25. Baral, S., Satpathy, S., Pati, D.P., Mishra, P., Pattnaik, L.: A literature review for detection and projection of cardiovascular disease using machine learning. *EAI Endorsed Trans. Internet of Things* **10**, 1–7 (2024)
 26. Mohapatra, S., Satpathy, S., Paul, D.: Data-driven symptom analysis and location prediction model for clinical health data processing and knowledge base development for COVID-19. In: Nandan Mohanty, S., Saxena, S.K., Satpathy, S., Chatterjee, J.M. (eds) Applications of Artificial Intelligence in COVID-19. Medical Virology: From Pathogenesis to Disease Control. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-7317-0_6
 27. Jawale, A., Warole, P., Bhandare, S., Bhat, K., Chandre, R.: Jeevn-Net: brain tumor segmentation using cascaded U-net & overall survival prediction. *Int. Res. J. Eng. Technol.* **14**, 56–62 (2020)
 28. Sahu, S., Kumar, R., Mohdshafi, P., Shafi, J., Kim, S., Ijaz, M.F.: A hybrid recommendation system of upcoming movies using sentiment analysis of Youtube trailer reviews. *Mathematics* **10**(9), 1–22 (2022). <https://doi.org/10.3390/math10091568>

AI for Secure Digital Infrastructure

Optimized Outsourced Decryption for Attribute-Based Encryption: Cost-Efficiency for Users and Cloud Servers in Green Cloud Computing



Subhash G. Rathod, Meghana R. Yashwante, Sunita Nikam,
Sushama Laxman Pawar, and Nilesh J. Uke

Abstract In green cloud computing, secure and efficient data management has become the biggest challenge. Attribute-based encryption (ABE) has emerged as a powerful cryptographic technique ensuring fine-grained access control over encrypted data. Still, the decryption process in ABE schemes is computationally intensive and challenging for users with limited resources to tackle this problem. Attribute-based encryption with Outsourced Decryption (ABE-OD) was presented, which can outsource decryption operation to the cloud server. Still, the cloud server decrypts exact cipher text repeatedly for multiple users satisfying the same access policy, which leads to inefficiency in green cloud networks. We propose a method to ensure efficient and minimum resource utilization by restricting cloud servers from performing single decryption operations for various users with the same access policy. Unlike ABE-OD, the total computation overhead of the proposed approach remains independent of several decryption requests. The proposed approach decreases the cloud servers' total workload and lowers the computational cost for users.

Keywords Green cloud computing · Data decryption · Data access control · Attribute-Based Encryption with Outsourced Decryption (ABE-OD)

These authors contributed equally to this work.

S. G. Rathod (✉) · M. R. Yashwante
Department of Computer Engineering, Marathwada Mitramandal's Institute of Technology, Pune
University, Pune, Maharashtra, India
e-mail: subhashrathod@gmail.com

S. Nikam
Symbiosis Skills and Professional University, Pune, Maharashtra, India

S. L. Pawar
Vishwakarma Institute of Information Technology, Pune University, Pune, Maharashtra, India

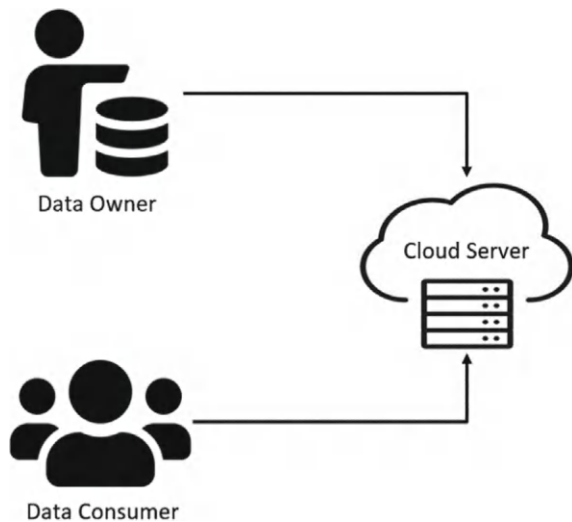
N. J. Uke
Indira College of Engineering and Management, Pune University, Pune, Maharashtra, India

1 Introduction

In modern data-sharing environments, centralized servers or cloud platforms store data, and an authorized approach is needed to access that data. However, the usage of public cloud computing has grown over time. Fig. 1 depicts the cloud-based data-sharing architecture where the data owner represents an individual or organization that generates and uploads data on a cloud server for backup or sharing. A cloud server represents a central repository that manages stored data, provides access to data as per the access permission, and provides secure transmission between parties. Data consumers represent individuals or organizations needing data stored on a cloud server. They request a cloud server to provide access to data uploaded by the data owner.

As cloud usage grows, it has introduced significant challenges to data security and user trust. Data owners store their data on cloud servers operated by third-party providers that are not entirely trustworthy. To address these challenges, [1] introduced an Attribute-Based Encryption (ABE) scheme for fine-grained access control for data sharing in cloud environments. This scheme allows data owners to encrypt their data using specific access policies, ensuring that only users with attributes that satisfy these policies can decrypt the data. This scheme is structured in two ways: Ciphertext-Policy ABE (CP-ABE) and Key-Policy ABE (KP-ABE). In CP-ABE, the access policy is embedded in the ciphertext, and users can decrypt the ciphertext only if the attributes associated with the user’s private key match the policy. In KP-ABE, the access policy gets assigned to the user’s private key, which allows decryption only if the attributes in the ciphertext satisfy the policy. Despite this state-of-the-art design of ABE, it still faces some critical efficiency challenges in the encryption

Fig. 1 Data sharing in cloud server



and decryption phases. ABE schemes execute computationally heavy modular exponentiation operations during encryption, and computational complexity increases as access policy becomes more complex, which poses difficult challenges in mobile and fog computing environments where devices have limited computational resources. It becomes impossible to share data securely without high computational devices. Researchers proposed outsourcing computationally intensive tasks to external service providers to solve these challenges.

Green cloud computing has attracted significant attention recently due to its focus on environmentally aware and efficient resource management solutions for cloud-based services. Ensuring data security and management becomes challenging when organizations increasingly depend on cloud servers for data storage and processing. One primary objective in a cloud environment is secure data sharing and fine-grained access control. When providing flexible data sharing, traditional cryptographic techniques often fall short.

As illustrated in Fig. 2, the data owner is an entity that is responsible for creating and encrypting the data using ABE. Data owners generate cipher text based on specific policy. The policy contains an access structure defining who can access the encrypted data. The data owner then uploads the encrypted data to a cloud server. End users who want to access the data can send retrieval requests to a cloud server. Each user has a set of attributes and keys defining their access privileges. If the user attributes satisfy the access policy, the cloud server partially transforms requested data for authorized users. Here, the cloud server acts as a middle entity that stores encrypted data and provides services for outsourced decryption operation, which reduces the computational load from users' resource-constrained devices. When the cloud server receives the data access request from the user, it checks the attributes of the requesting user against the access policy embedded in the ciphertext. Suppose the user attributes satisfy the access policy. In that case, the cloud server performs a transformation operation with the help of a transformation key on the requested ciphertext data to produce transformed text. Transformed text is then sent to the requesting user, who fully decrypts data with the user's secret key. Here, only the user with attributes that satisfy the access policy completes the decryption process.

The ABE scheme became popular for providing fine-grained access control over encrypted data. ABE ensures fine-grained access control by integrating access policies, cipher texts, and attributes with user keys, allowing data owners to apply complex access policies without any continuous interaction with the system. However, the computational complexity and cost of the ABE scheme are critical issues specifically for users with resource-constrained devices. So, to tackle this issue, Attribute-Based Encryption with Outsourced Decryption (ABE-OD) was proposed, which allows users to perform computationally heavy decryption processes on cloud servers by outsourcing the process to the cloud server.

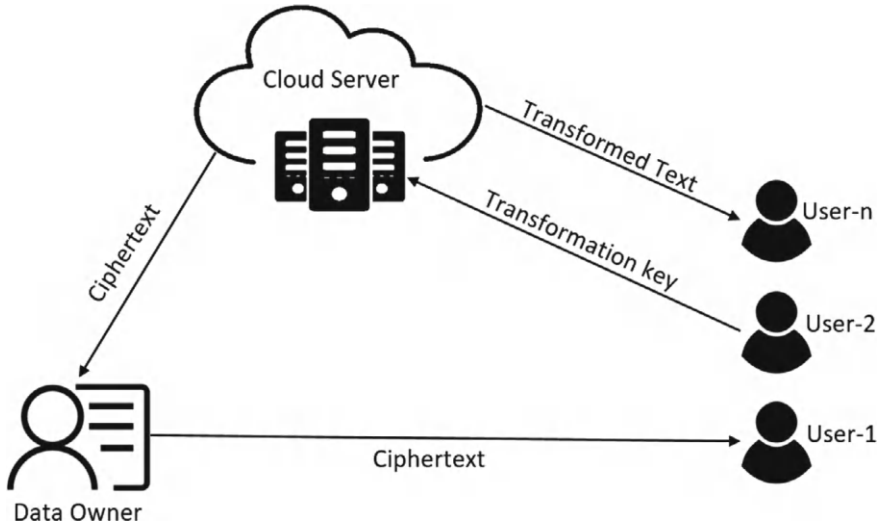


Fig. 2 ABE-OD architecture

1.1 Problem Statement

ABE-OD is designed to minimize heavy computational overhead on users. Still, it transfers overhead on the server side, introducing inefficiencies on the server side, which does not align with the principles of green cloud computing. The server performs the same decryption operations separately for multiple users with the same access policy. This repetitive decryption operation increases the computational workload on the server side.

For instance, in an organization, employees have identical access privileges to shared resources, and the server would repetitively perform the same decrypt operation for each user. This repetitive process leads to unnecessary resource consumption, reducing the overall efficiency of green cloud networks. Solving this problem is essential to achieve sustainable and scalable cloud infrastructure.

1.2 Contributions

The proposed methodology is designed to address the critical issue of unnecessary resource consumption in existing ABE-OD schemes by reducing the overhead of the cloud server and enhancing scalability and efficiency. The proposed method guarantees robust security of sensitive data even during outsourced decryption operations. Abbreviations and Notations are shown in Table 1.

The remaining sections are structured: Sect. 2 presents related work and a comparative analysis of existing data sharing and access revocation schemes. In Sect. 3,

Table 1 Abbreviation and notations

Notation	Description
Sec_{param}	Security parameter
Pu_{param}	Public parameters
U	Universal attributes
S	Attribute Set
MSK	Master secret key
M	Plain text message
$CapCapA$	Access structure
SK	Private decryption key
Tr_{key}	Transformation key
Tr_{CT}	Transformed ciphertext
Ret_{key}	Function as part of the key generation process

we have proposed a design goal that optimizes the ABE scheme in green cloud computing. In Sect. 4, we have proposed data sharing with a lightweight ABE-OD scheme. Section 5 focuses on analyzing its performance and security, respectively. Finally, the paper is concluded in the last section.

2 Related Works

Author Sahai et al. presented a concept of Fuzzy IBE, which can be used for applications such as biometric and attribute-based encryption (ABE). This concept incorporates error tolerance between the private key and public key identities. This scheme provides security under the Selective-ID model through a modified Bilinear Decisional Diffie-Hellman assumption. This paper also proposed problems like multi-authority attribute-based systems and schemes that hide the public key used during encryption. To address data security issues in hybrid cloud environments, P. Kanchanadevi et al. proposed ABE-DAS, which is designed to enhance data security by dynamically supporting attributes and securing sensitive data operations on private clouds while allowing nonsensitive data operational tasks on public clouds. This scheme focuses on privacy and data leakage in hybrid cloud setups [2]. Kapusta et al. *addressed two major challenges in cloud storage: secure data sharing and efficient access revocation*. The system integrates three approaches: encryption, AON transformation, and data dispersal. This scheme protects sensitive data and provides fast access revocation [3]. Song et al. *addressed security incidents like data leakage in the cloud*. The author proposed CSSM. This mechanism integrates data dispersion, encrypted chunking, and distributed storage. Cryptographic material leakage is prevented by combining user passwords with secret sharing [4]. Chen et al. *addressed a scenario* where a primary user may be unable to decrypt data in time. For such cases, the author presented an approach that allows alternate

semi-authorized users to decrypt ciphertexts cooperatively. This method allows independent message recovery for authorized users and ensures honest participation of semi-authorized users using an integrated access tree, resulting in low computational and storage costs [5]. Ramachandra et al. addressed privacy and security concerns for big data in cloud environments. To tackle this issue, the author proposed the TDES methodology. This methodology extended the key sizes of the traditional Data Encryption Standard (DES), achieving reduced encryption and decryption times [6]. A. Bakas. et al. *designed a hybrid encryption scheme combining ABE and Symmetric Searchable Encryption (SSE) to improve security in data storage in cloud environments*. ABE is used for user revocation, and SSE enhances security. The author also integrated Intel's SGX to design revocation and access control mechanisms for further security [7]. Rawal et al. presented a file-sharing mechanism using a Disintegration Protocol (DIP). This protocol allows seamless file sharing without sharing encryption keys across different clouds. This approach maintains data integrity and confidentiality and addresses vulnerabilities in traditional encryption methods [8]. Ma et al. proposed the PDSC system to tackle data leakage and computational inaccuracy. The author designed privacy-preserving computation protocols to ensure secure data sharing among multiple providers and computational correctness [9]. Chen et al. *developed a framework integrating Attribute-Based Access Control (ABAC), CP-ABE, and cloud-based file-sharing services*. Performance limitations in CP-ABE from Lattice (CP-ABE-L) are addressed by incorporating an optimized Small Policy Matrix (SPM) generation algorithm and Error Proportion Allocation (EPA) mechanism, which helps in minimizing error bounds and reduces computation and storage costs [10]. Xue et al. proposed a system that integrates CP-ABE. The system is designed to address Economic Denial of Sustainability (DDoS) attacks and provide resource accountability. That system allows cloud providers to verify downloaders' decryption capabilities, preventing malicious attacks that exploit cloud resources. This approach provides transparency in resource consumption [11]. Li et al. reduced the computational complexity of ABE for users and attribute authorities. The Secure Outsourced ABE scheme uses a Key Generation Service Provider (KGSP) and a Decryption Service Provider (DSP) that outsources access policy and attribute-related operations. Data verification mechanisms ensure the correctness of outsourced computations [12]. Qin et al. *designed a scheme for the concept of ABE with outsourced decryption*. The author addressed the problem of verifying the correctness of the transformation operation performed by the cloud server. The author proposed a model to verify outsourced decryption and a generic method that converts any ABE scheme into one that supports verifiable decryption [13]. Lin et al. *addressed the issues of bandwidth and computational inefficiencies in existing verifiable outsourced ABE schemes; the author proposed a symmetric-key encryption scheme based on an attribute-based key encapsulation mechanism*, which reduces bandwidth and computational costs [14]. Wang et al. *proposed a privacy-preserving mechanism that can audit data and check the integrity of shared cloud data using ring signatures*. This approach maintains signers' privacy while performing data integrity checks. The approach does not require retrieving the entire file during verification [15]. Li et al. addressed the problem of heavy cryptographic operations in Traditional

Secure Cloud Storage (SCS), which makes data outsourcing computationally expensive. To solve this problem, the author designed A lightweight Secure Auditable Cloud Storage (SecACS) scheme that provides data dynamics with less computational overhead. The proposed scheme supports multiple update operations and uses lightweight cryptographic operations, which improve efficiency [16]. Rabaninejad et al. *Proposed an ID-based data verification protocol* that provides simplified key management by not relying on Public Key Infrastructure (PKI). The Proposed protocol offers dynamic data operations, group user revocation, and privacy against Third-Party Auditors (TPAs). The protocol requires constant computational overhead regardless of group size, making it suitable for large dynamic groups. The protocol requires only one pairing operation for large groups [17]. Wang et al. *proposed an extended data access control scheme for multi-authorized cloud storage (NEDAC-MACS)*, which integrates CP-ABE to enhance security in multi-authority cloud storage. The proposed scheme enhances security by addressing collusion between servers and users and supports secure revocation [18]. Ding et al. proposed four privacy-preserving division schemes for flexible access control. Scheme provides division for integers and is extended to support fractional and fixed-point numbers. The scheme addresses the lack of privacy-preserving division operations in encrypted computations. It enhances privacy-preserving computation by supporting division and fractional/fixed-point operations [19]. Wei et al. *addressed the problem of lack of trust* in cloud servers and scalable user revocation. To solve this problem, a multi-authority CP-ABE scheme is proposed that offers data access control with scalable user revocation and public ciphertext updates. Here, multiple authorities issue secret keys independently and provide dynamic user revocation with forward and backward security [20]. Wang et al. *addressed the issues of security and privacy in cloud computing*, specifically where sensitive data is outsourced to untrusted public cloud servers. Here, the author proposed a secure data-sharing scheme to maintain the privacy of the data owner, provide security to outsourced data, and, without compromising on security, provide flexibility for data utility. The scheme is designed to manage electronic health (E-health) records, where privacy and security are critical requirements [21]. Nelmiawati et al. proposed a tool called dCloud to enhance the security of data stored in public cloud storage. The author addressed the issues of data privacy and integrity. The iCloud tool integrates Rabin's Information Dispersal Algorithm (IDA) and Shamir's Secret Sharing Algorithm (SSA) for file protection. This tool automates the secure random generation of the secret key using Rabin's IDA, and to disperse the secret key into the output files generated by Rabin's IDA, it uses Shamir's SSA [22]. Reyazulla et al. *The proposed framework is proposed to provide security against Economic Denial of Sustainability (EDoS) attacks, and it uses CP-ABE for fine-grained and flexible access control*. The system addresses the lack of visibility into resource usage and associated costs. The proposed framework mitigates DDoS attacks, reduces unnecessary resource consumption, and improves transparency in resource usage and associated expenses [23]. Lekshmi et al. proposed a solution that uses semi-trustable online proxy servers for attribute revocation. The proposed approach uses proxy re-encryption with CP-ABE, enabling multi-authority systems to revoke user attributes with minimal overhead. This approach

outsources the computational workload to proxy servers, which reduces overhead on the authority [24]. Comparative Assessment of Current Frameworks is shown in Table 2.

Table 2 A comparative assessment of current frameworks

Ref. No.	Framework/Encryption scheme used	Is cost efficient	Access revocation	Key features
[25]	ABE with outsourced encryption and decryption	Yes	No	Fine-grained data sharing with outsourced encryption and decryption, which is suitable for mobile environments
[26]	ABE outsourcing scheme for fog-enabled IoT	Yes	Yes	Efficient cipher text update with high revocation efficiency, suitable for fog computing environments
[27]	CP-ABE with attribute revocation	Yes	Yes	Selective security under q -parallel BDHE problem. Suitable for fog-enabled e-health applications,
[28]	Lightweight ABE framework with computational offloading	Yes	No	Integrated proxy servers for cryptographic operations are efficient for resource-constrained IoT devices
[29]	CP-ABE with Proxy Re-encryption (PRE) and policy versioning	No	Yes	It supports policy updates and the traceability of policy updates designed for secure outsourced personal health records (PHR)
[30]	Ciphertext-Policy Attribute-Based Signcryption	No	No	Designed for secure sharing of PHR via the cloud. Implemented outsourcing to reduce computational overhead
[31]	Privacy-preserving outsourced similarity test (PPOS) using CP-ABE and garbled circuits	Yes	No	Designed for efficient outsourced similarity testing with privacy
[32]	A hybrid encryption scheme with identity-based encryption (IBE) is integrated into OutFS	No	No	An efficient encrypted file system with outsourced data sharing for cloud storage, robust against various attacks
[33]	Revocable Multi-Authority ABE (RMA-ABE) based on elliptic curve cryptography	No	Yes	Supports fine-grained access control as well as attribute revocation, suitable for resource-constrained devices

3 Design Goals

We have proposed a design to optimize the ABE scheme in green cloud computing by considering the following design goals:

Efficiency in Decryption with Security and Privacy

There is less computational overhead for both the cloud server and users during the decryption process. Enable the cloud server to perform a single decryption operation for various users who satisfy the same access policy, thereby avoiding repetitive computations for the exact cipher text. Provide confidentiality during the outsourced decryption process. Ensure that the cloud server does not gain access to plaintext data or information about access policies.

Scalability

As the number of users increases, the system should handle decryption requests without an increase in server workload. Maintain consistent performance as the number of decryption requests increases.

Resource Optimization

Optimizing resource usage, reducing energy consumption, and minimizing operational costs by achieving green cloud computing goals. Using efficient resource management techniques to achieve sustainable and eco-friendly cloud operations.

4 Proposed Design and Implementation

The proposed model addresses the inefficiencies of the ABE-OD model. Cloud servers can perform decryption only once for several users requesting decryption under the same access policy.

The ABE-OD scheme consists of seven algorithms: *Setup*, *Encrypt*, *KeyGen*, *Gen_{Tkey}*, *Tr_{out}*, *Decrypt*, and *Dec_{out}*. Unlike some earlier models [12], this approach consists of *Gen_{Tkey}* Algorithm that creates transformation keys for improving flexibility for users. The algorithms are described as follows:

Setup (Sec_{param}, U): This algorithm initializes the system by producing a master secret key MSK and public parameters Pu_{param} , based on the security parameter Sec_{param} and the universe of attributes U .

KeyGen(MSK, Pu_{param}, S): Using the master secret key MSK , public parameters Pu_{param} , and an attribute set S , this algorithm generates a private decryption key SK For a user.

Encrypt(Pu_{param}, M, A): This algorithm encrypts the plaintext message M under an access structure A , producing the cipher text CT .

Decrypt(SK, CT): Using the private decryption key SK , this algorithm decrypts the cipher text CT if the attribute set S satisfies the access policy A .

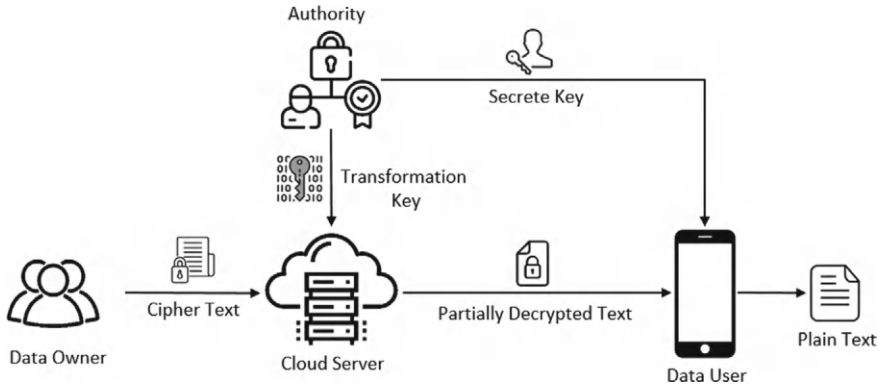


Fig. 3 Framework of outsource decryption with ABE

$Gen_{Tkey}(Pu_{param}, SK)$: This algorithm creates a transformation key Tr_{key} and a retrieval key Ret_{key} , enabling outsourced decryption by the cloud server.

$Tr_{out}(Tr_{key}, CT)$: The cloud server uses the transformation key Tr_{key} to partially decrypt CT , resulting in a transformed ciphertext Tr_{CT} .

$Dec_{out}(CT, Tr_{CT}, Ret_{key})$: This algorithm completes decryption on the client side. When the attribute set S satisfies A , it combines CT , Tr_{CT} and Ret_{key} to recover the plaintext M . If S does not meet A , the output is \perp .

As illustrated in Fig. 3, the proposed design has three main key components.

Initial Setup:

ABE-OD system (1) initializes and generates public parameters in this phase. Pu_{param} And master secret key. MSK based on security parameters Sec_{param} And attributes Atr .

$$Setup(Sec_{param}, Atr) \rightarrow (MSK, Pu_{param}) \quad (1)$$

Decryption Key:

Private decryption key (2) SK is generated for the user Usr using attribute set Usr_{atr} master secret key MSK , and the public parameters Pu_{param}

$$KeyGen(MSK, Pu_{param}, Usr_{atr}) \rightarrow SK \quad (2)$$

Decryption Key SK is mathematically related to Usr_{atr} To enforce access policy during decryption.

Encryption:

Plaintext message M is encrypted under an access structure A Which defines the policy for encryption and ciphertext CT Is generated (3).

$$\text{Encrypt}(M, A, Pu_{param}) \rightarrow CT \quad (3)$$

Transformation Key:

Transformation keys are generated for each other using their private key and attributes.

These keys help the server to determine if the users share the same access policy instead of performing decryption from scratch for every user (4)

$$Tr_{key} = Gen_{Tkey}(Pu_{param}, SK) \quad (4)$$

where,

Pu_{param} is public parameters.

SK is the private decryption key.

The transformation key ensures the server can only partially decrypt cipher texts without full decryption.

Transformed Cipher Text:

Using the transformation key, the cloud server computes a single transformed cipher text for a given cipher text. When the first request for decryption comes, the transformed cipher text is stored temporarily and reused for subsequent requests from users with the same access policy (5).

$$Tr_{CT} = Tr_{out}(Tr_{key}, CT) \quad (5)$$

where,

Tr_{key} is transformation key.

CT is cipher text from the data owner.

User's Role:

Each user completes the decryption using their lightweight computation after receiving the transformed cipher text from the cloud server, ensuring privacy and security (6).

$$M = Dec_{out}(CT, Tr_{CT}, Ret_{key}) \quad (6)$$

where,

M is decrypted plain text.

Tr_{CT} is transformation Text.

CT is cipher text from the data owner.

Ret_{key} is derived from Gen_{Tkey} Function as part of the key generation process.

5 Performance Analysis

Figure 4 illustrates the relation between the number of attributes in the access structure and the transformation time required by the cloud server. The transformation time grows linearly as the number of attributes increases, indicating that the computational cost for transformation is directly proportional to the increase in the attribute count.

Figure 5 depicts how the size of cipher text increases as the number of attributes increases in the access structure. The linear growth in cipher text size depicts that additional components for each attribute are included during the encryption process, resulting in more extensive cipher texts as the attribute count increases.

Figure 6 illustrates a comparative performance analysis of a proposed method and Green et al. [34] method in handling outsourced decryption requests on a cloud server. The proposed method processing time remains constant regardless of the number of requests, demonstrating high efficiency and scalability. The method shows a linear increase in processing time as the number of requests increases, indicating poorer scalability. Overall, the proposed approach is significantly more efficient, especially as the number of requests grows.

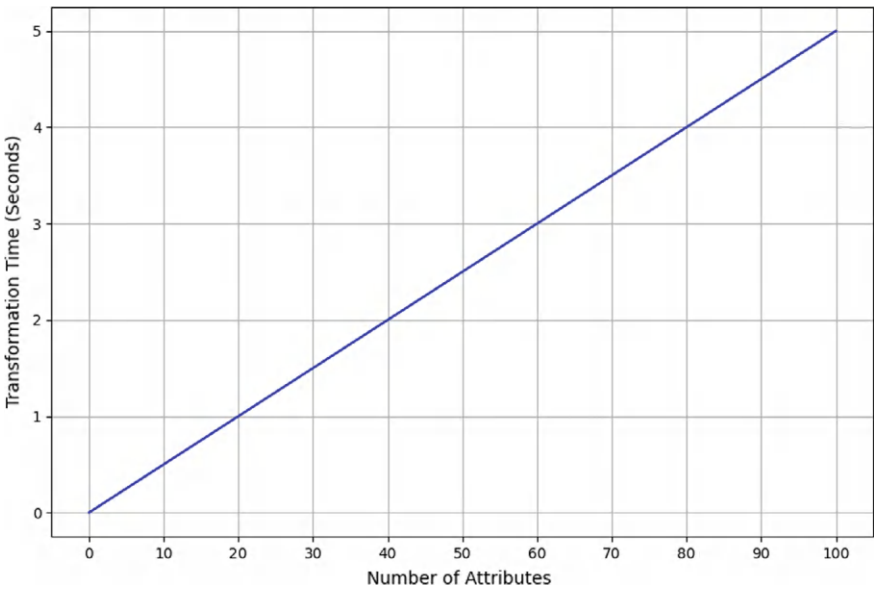


Fig. 4 Transformation time required for n number of attributes

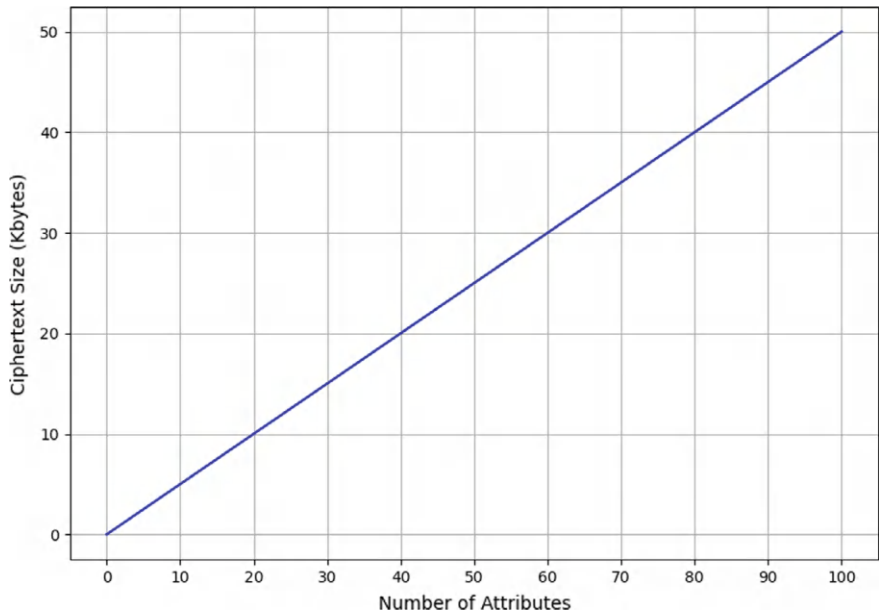


Fig. 5 Cipher text Size for n number of attributes

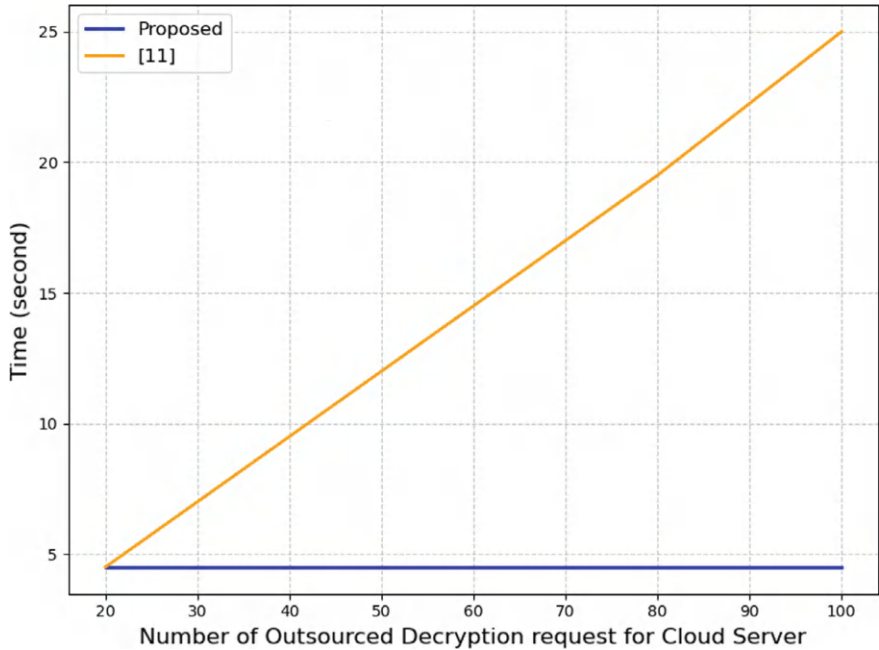


Fig. 6 Comparative analysis of time required for n number of attributes

6 Conclusion

The proposed model effectively balances user convenience, server efficiency, and security by introducing flexibility in attribute-based encryption through transformation keys and outsourced decryption. By leveraging the transformation key, the cloud server can efficiently handle multiple decryption requests, ensuring that users with the same access policy share a reusable transformed cipher text, eliminating the need for repetitive decryption operations while maintaining security.

Allowing a single decryption operation per cipher text for multiple users with the same access policy reduces computational overhead in cloud environments, thereby achieving the goal of green cloud networks.

Future research will be focused on optimizing the transformation key generation process to handle dynamic user attribute changes.

Funding There is no funding.

Data Availability No Data is associated with this research.

Conflicts of Interest On behalf of all contributors, the corresponding author confirms that no conflicts of interest exist.

References

1. Sahai, A., Waters, B.: Fuzzy identity-based encryption. In: Cramer, R. (eds.) *Advances in Cryptology – EUROCRYPT 2005*. EUROCRYPT 2005. Lecture Notes in Computer Science, vol. 3494. Springer, Berlin, Heidelberg (2005). https://doi.org/10.1007/11426639_27
2. Kanchanadevi, P., Raja, L., Selvapandian, D., Dhanapal, R.: An attribute-based encryption scheme with dynamic attributes supporting in the hybrid cloud. In: *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Palladam, India, pp. 271–273 (2020). <https://doi.org/10.1109/I-SMAC49090.2020.9243370>
3. Kapusta, K., Qiu, H., Memmi, G.: Secure data sharing with fast access revocation through untrusted clouds. In: *2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, Canary Islands, Spain, pp. 1–5 (2019). <https://doi.org/10.1109/NTMS.2019.8763850>
4. Song, H., Li, J., Li, H.: A cloud secure storage mechanism based on data dispersion and encryption. *IEEE Access* **9**, 63745–63751 (2021). <https://doi.org/10.1109/ACCESS.2021.3075340>
5. Chen, N., Li, J., Zhang, Y., Guo, Y.: Efficient CP-ABE scheme with shared decryption in cloud storage. *IEEE Trans. Comput.* **71**(1), 175–184 (2022). <https://doi.org/10.1109/TC.2020.3043950>
6. Ramachandra, M.N., Srinivasa Rao, M., Lai, W.C., Parameshachari, B.D., Ananda Babu, J., Hemalatha, K.L.: An efficient and secure big data storage in cloud environment by using triple data encryption standard. *Big Data Cogn. Comput.* **6**(4), 101 (2022). <https://doi.org/10.3390/bdcc6040101>
7. Bakas, A., Dang, H.-V., Michalas, A., Zalizko, A.: The cloud we share: access control on symmetrically encrypted data in untrusted clouds. *IEEE Access* **8**, 210462–210477 (2020). <https://doi.org/10.1109/ACCESS.2020.3038838>

8. Rawal, B.S., Vivek, S.S.: Secure cloud storage and file sharing. In: 2017 IEEE International Conference on Smart Cloud (SmartCloud), New York, NY, USA, pp. 78–83 (2017). <https://doi.org/10.1109/SmartCloud.2017.19>
9. Ma, Z., Ma, J., Miao, Y., Liu, X., Yang, T.: Privacy-preserving data sharing framework for high-accurate outsourced computation. In: ICC 2019 - 2019 IEEE International Conference on Communications (ICC), Shanghai, China, pp. 1–6 (2019). <https://doi.org/10.1109/ICC.2019.8761251>
10. Chen, E., Zhu, Y., Liang, K., Yin, H.: Secure remote cloud file sharing with attribute-based access control and performance optimization. *IEEE Trans. Cloud Comput.* **11**(1), 579–594 (2023). <https://doi.org/10.1109/TCC.2021.3104323>
11. Xue, K., Chen, W., Li, W., Hong, J., Hong, P.: Combining data owner-side and cloud-side access control for encrypted cloud storage. *IEEE Trans. Inf. Forensics Secur.* **13**(8), 2062–2074 (2018). <https://doi.org/10.1109/TIFS.2018.2809679>
12. Li, J., Huang, X., Li, J., Chen, X., Xiang, Y.: Securely outsourcing attribute-based encryption with checkability. *IEEE Trans. Parallel Distrib. Syst.* **25**(8), 2201–2210 (2014). <https://doi.org/10.1109/TPDS.2013.271>
13. Qin, B., Deng, R.H., Liu, S., Ma, S.: Attribute-based encryption with efficient verifiable outsourced decryption. *IEEE Trans. Inf. Forensics Secur.* **10**(7), 1384–1393 (2015). <https://doi.org/10.1109/TIFS.2015.2410137>
14. Lin, S., Zhang, R., Ma, H., Wang, M.: Revisiting attribute-based encryption with verifiable outsourced decryption. *IEEE Trans. Inf. Forensics Secur.* **10**(10), 2119–2130 (2015). <https://doi.org/10.1109/TIFS.2015.2449264>
15. Wang, B., Li, B., Li, H.: Oruta: privacy-preserving public auditing for shared data in the cloud. *IEEE Trans. Cloud Comput.* **2**(1), 43–56 (2014). <https://doi.org/10.1109/TCC.2014.2299807>
16. Li, L., Liu, J.: SecACS: enabling lightweight secure auditable cloud storage with data dynamics. *J. Inf. Secur. Appl.* **54**, 102545. <https://doi.org/10.1016/j.jisa.2020.102545>, ISSN 2214-2126
17. Rabaninejad, R., Sedaghat, S.M., Ahmadian Attari, M., Aref, M.R.: An ID-based privacy-preserving integrity verification of shared data over untrusted cloud In: 2020 25th International Computer Conference, Computer Society of Iran (CSICC), Tehran, Iran, pp. 1–6 (2020). <https://doi.org/10.1109/CSICC49403.2020.9050098>
18. Wang, J., Wu, K., Ye, C., Xia, X., Ouyang, F.: Improving security data access control for multi-authority cloud storage. In: 2019 IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Xiamen, China, pp. 608–613 (2019). <https://doi.org/10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00092>
19. Ding, W., et al.: An extended framework of privacy-preserving computation with flexible access control. *IEEE Trans. Netw. Serv. Manag.* **17**(2), 918–930 (2020). <https://doi.org/10.1109/TNSM.2019.2952462>
20. Wei, J., Liu, W., Hu, X.: Secure and efficient attribute-based access control for multiauthority cloud storage. *IEEE Syst. J.* **12**(2), 1731–1742 (2018). <https://doi.org/10.1109/JSYST.2016.2633559>
21. Wang, H.: Anonymous data sharing scheme in public cloud and its application in E-health record. *IEEE Access* **6**, 27818–27826 (2018). <https://doi.org/10.1109/ACCESS.2018.2838095>
22. Nelmiawati, Arifandi, W.: A seamless secret sharing scheme implementation for securing data in public cloud storage service. In: 2018 International Conference on Applied Engineering (ICAE), Batam, Indonesia, pp. 1–5 (2018)/ <https://doi.org/10.1109/INCAE.2018.8579388>
23. Reyazulla, K., Kesavulu Reddy, E.: Efficient two sided access control system in cloud storage. *Int. J. Eng. Res. & Technol. (IJERT) NCISIT* **8**(02) (2020)
24. Lekshmi, S.V., Revathi, M.P.: Implementing secure data access control for multi-authority cloud storage system using Ciphertext Policy-Attribute based encryption. In: International Conference on Information Communication and Embedded Systems (ICICES2014), Chennai, India, pp. 1–6 (2014). <https://doi.org/10.1109/ICICES.2014.7033749>

25. Li, Z., Li, W., Jin, Z., Zhang, H., Wen, Q.: An efficient ABE scheme with verifiable outsourced encryption and decryption. *IEEE Access* **7**, 29023–29037 (2019). <https://doi.org/10.1109/ACCESS.2018.2890565>
26. Li, L., Wang, Z., Li, N.: Efficient attribute-based encryption outsourcing scheme with user and attribute revocation for fog-enabled IoT. *IEEE Access* **8**, 176738–176749 (2020). <https://doi.org/10.1109/ACCESS.2020.3025140>
27. Zhao, J., Zeng, P., Choo, K.-K.R.: An efficient access control scheme with outsourcing and attribute revocation for fog-enabled E-health. *IEEE Access* **9**, 13789–13799 (2021). <https://doi.org/10.1109/ACCESS.2021.3052247>
28. Taha, M.B., Khasawneh, F.A., Quttoum, A.N., Alshammari, M., Alomari, Z.: Outsourcing attribute-based encryption to enhance IoT security and performance. *IEEE Access* **12**, 166800–166813 (2024). <https://doi.org/10.1109/ACCESS.2024.3491951>
29. Fugkeaw, S.: A lightweight policy update scheme for outsourced personal health records sharing. *IEEE Access* **9**, 54862–54871 (2021). <https://doi.org/10.1109/ACCESS.2021.3071150>
30. Deng, F., Wang, Y., Peng, L., Xiong, H., Geng, J., Qin, Z.: Ciphertext-policy attribute-based signcryption with verifiable outsourced designcryption for sharing personal health records. *IEEE Access* **6**, 39473–39486 (2018). <https://doi.org/10.1109/ACCESS.2018.2843778>
31. Yang, D., Chen, Y.-C., Ye, S., Tso, R.: Privacy-preserving outsourced similarity test for access over encrypted data in the cloud. *IEEE Access* **6**, 63624–63634 (2018). <https://doi.org/10.1109/ACCESS.2018.2877036>
32. Khashan, O.A.: Secure outsourcing and sharing of cloud data using a user-side encrypted file system. *IEEE Access* **8**, 210855–210867 (2020). <https://doi.org/10.1109/ACCESS.2020.3039163>
33. Ming, Y., He, B., Wang, C.: Efficient revocable multi-authority attribute-based encryption for cloud storage. *IEEE Access* **9**, 42593–42603 (2021). <https://doi.org/10.1109/ACCESS.2021.3066212>
34. Green, M., Hohenberger, S., Waters, B.: Outsourcing the decryption of ABE ciphertexts. In: *Proceedings of 20th USENIX Conference on Security*, San Francisco, CA, USA, August 2011, p. 34 (2011)

Document Privacy Preservation Using Information Security Methods and Create Awareness About Privacy Policies



Vidhya Gavali, Aastha Shinde, Shweta Gumaste, Vaishnavi Akul, and Ameya Kunte

Abstract Privacy is the fundamental right of individuals to control their personal information and its usage. In the context of document privacy, existing solutions primarily focus on basic encryption for secure storage, often lacking comprehensive user control over data access and usage comprehension. Data privacy is significant to avoid data breaches and protect the stability and integrity of financial transactions and digital economic activities. It helps establish trust between individuals and organizations in the businesses. Data Privacy concerns have become increasingly prevalent due to the growing reliance on digital technologies and the increasing amount of online personal data. Privacy concerns such as compliance with privacy regulations, third-party data sharing and selling, data breaches and unauthorized access, and lack of transparency in data collection underline the need for stronger data protection measures, more transparent regulations, and increased user control over personal data.

Keywords Privacy encryption · Data breach prevention · Privacy threats · Unauthorized access to personal data

1 Introduction

With the rapid expansion of Digital platforms and mobile applications, the collection and exchange of personal data have surged dramatically. However, despite increased awareness of privacy risks, user behavior often contradicts their expressed concerns—a phenomenon known as the “privacy paradox.” This paradox reflects a gap between users’ stated desire for privacy and their willingness to sacrifice it for convenience, exposing them to potential data misuse. To tackle these issues, privacy

V. Gavali (✉) · A. Shinde · S. Gumaste · V. Akul · A. Kunte

Department of Artificial Intelligence and Data Science, Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India

e-mail: vidhya.n.gavali@gmail.com

laws like the General Data Protection Regulation (GDPR) have been introduced to empower users by giving them greater control over their data. However, many users remain unaware of how their data is collected, processed, and shared. Despite understanding the risks, others prioritize the convenience of digital services over data protection. In such a complex environment, finding the optimal balance between privacy and functionality remains a significant challenge for users and businesses. This challenge makes it essential to develop tools that offer users better insights into privacy policies and secure platforms for managing sensitive information. Our project aims to address these issues by combining a secure document storage platform with an NLP-powered privacy policy analyzer, which offers a seamless way to protect data and improve transparency in privacy practices.

2 Features

1. **AES-256 Encryption for Robust Data Security:** AES-256 is one of the most advanced encryption standards available and widely trusted for safeguarding sensitive data. It ensures that even if files are intercepted or compromised, they remain unreadable without the corresponding decryption key. This level of security is commonly used by enterprises and governments, making it ideal for protecting sensitive documents. 2. **Temporary Key Access for Controlled Sharing:** The system allows users to generate JWT-based temporary keys for sharing files with specific individuals for a limited period. Once the key expires, access is automatically revoked, ensuring the shared documents are protected. This feature provides the convenience of temporary access without compromising the integrity and security of the stored data.

3 Literature Survey

Makhdoom et al. [1] introduced the PrivySeC framework, which ensures secure personal data sharing in Internet of Things (IoT) ecosystems using Distributed Ledger Technology (DLT) like Corda and Homomorphic Encryption. Their system facilitates privacy-preserving analytics on encrypted data while maintaining scalability and firm access control. However, some challenges remain, particularly in optimizing performance and handling complex data-sharing scenarios in a scalable manner. Demirer et al. [2] examine the literature to determine how privacy laws, especially the GDPR, affect data usage and firm production. Previous studies demonstrate how laws such as the GDPR impact businesses' data management procedures by raising compliance expenses. Though the long-term effects of privacy laws on production efficiency have not yet been thoroughly investigated, studies have looked at the role

of cloud computing in company production, specifically in data storage and computation. Using a difference-in-differences approach, the study expands on this by evaluating how GDPR compliance affects businesses' reliance on data and technology, specifically in EU versus US firms. The survey also touches

On the limitations of relying on data from a single cloud provider and the broader implications of privacy regulations on firm productivity and innovation. Campbell et al. [3] studied businesses' difficulties in fulfilling GDPR, particularly in online and mobile applications. Previous studies underscore the need to be unambiguous, especially about data protection by design and user consent mechanisms. Research indicates that the creation of privacy policies has been automated through the use of static code analysis and deep learning models, which has improved consistency and decreased manual labor. Limitations include the variety of data sets, tool accuracy, and user understanding of automated policies; however, these remain significant obstacles. The report emphasizes the importance of maintaining compliance through ongoing monitoring and upgrades. The research builds on these studies to propose automating privacy policy captions to enhance transparency, efficiency, and user trust.

Zhang et al. [4] aimed to give consumers more control over their data. However, research indicates that many people continue to put social demands ahead of privacy, perfecting the contradiction. According to the privacy model, users balance the advantages of service access against possible privacy risks. Peer pressure is necessary; because of social connections, individuals may follow their networks to platforms that prioritize privacy or stay on less secure ones. Learned helplessness and privacy cynicism are also covered, with users feeling resigned to privacy infractions due to complicated and inefficient privacy tools. The survey emphasizes the need for better platform design and more straightforward privacy protections to address these issues. Hiwale et al. [5] systematically reviewed integrating blockchain and federated learning for telemedicine applications. Their study demonstrates how blockchain can offer decentralized and immutable data storage while federated learning enhances privacy by facilitating collaborative model training without sharing raw data. Despite the promising combination of these technologies, the authors note that blockchain suffers from scalability and interoperability issues, and federated learning remains vulnerable to inference attacks.

Hassan et al. [6] focused on privacy protection in textual documents, utilizing word embeddings and generalization techniques. Their work proposes masking quasi-identifying terms to maintain privacy while preserving the utility of the documents. Although the method is flexible and language agnostic, it relies heavily on detailed knowledge bases, such as ontologies, which may not always be available. Additionally, performance overheads are involved in training word embeddings, and challenges arise in scenarios where multiple entities are present in the same document. Khan et al. [7] addressed data security in the healthcare sector, specifically within Industrial Internet of Things (IIoT) ecosystems. They proposed a blockchain-based solution, supported by technologies like Wireless Sensor Networks (WSNs) and smart contracts, to enhance security and privacy in E-healthcare data. Their system reduces resource consumption and prevents tampering, but it faces challenges in

managing communication between on-chain and off-chain transactions and initial weak interconnectivity of healthcare IIoT devices.

Nguyen et al. [8] explored privacy-preserving techniques such as Secure Multi-Party Computation (SMC), Homomorphic Encryption, and Differential Privacy to decentralize data training. Their research emphasizes the role of federated learning (FL) in reducing central storage risks by training models locally without sharing raw data. However, their work highlights that FL is still vulnerable to inference attacks and model poisoning, and additional privacy-preserving techniques are needed to ensure full GDPR compliance.

Chaudhari et al. [9] proposed a privacy-preserving searchable encryption scheme with fine-grained access control, enabling secure searches over encrypted cloud data. Their system also ensures scalability in a multi-sender and multi-receiver setup. However, despite the proven security against chosen-keyword attacks using a random oracle model, their approach incurs some performance overhead due to the complexity of handling multiple attributes for encryption and search.

Story et al. [10] proposed a three-tiered classification model used in the NLP component that classifies privacy activities according to three factors: modality (performed or not performed), party (first or third), and data type. This method's benefits include the system's scalability, allowing it to analyze millions of apps and enhancing privacy transparency through automating privacy issue identification. The study benefits businesses and government authorities by assisting them in more effectively identifying compliance issues. The study also Liu et al. [11] automated the annotation of privacy rules so that users may better comprehend them by utilizing machine learning and natural language processing (NLP). Using annotated corpora like OPP-115, the study uses logistic regression (LR), support vector machines (SVMs), and convolution neural networks (CNNs) to classify policy language, including First Party Data Collection, Third Party Sharing, and Data Security. The goal is to make data practices more transparent and reduce users' work to understand privacy policies. The benefits include better classification accuracy, with segment-level F1 scores as high as 0.78, which facilitates the identification of essential privacy practices. Although these techniques are a step in the right direction for automated privacy policy analysis, they still need to be more accurate and scalable.

4 System Architecture

The Dockrypt platform integrates two primary tools: a secure document storage system and an NLP-powered privacy policy analyzer, offering users a unified experience. Both components share a standard user authentication and authorization framework, with JWT tokens securing access to the platform. Role-Based Access Control (RBAC) manages permissions, ensuring users have the appropriate access based on their roles.

The platform emphasizes scalability and portability through Docker-based containerization, allowing easy deployment of cloud services like AWS or GCP.

This modular architecture ensures that the storage and privacy tools operate independently while interacting seamlessly within the same ecosystem. The backend APIs are implemented using Flask or FastAPI to handle requests, encryption operations, and NLP processing. The frontend interface is built using React.js, offering a responsive and user-friendly experience. This architecture ensures that each tool can function autonomously. Still, their integration provides a cohesive platform where users can manage secure documents and analyze privacy policies effectively, as shown in Fig. 1.

The secure document storage system safeguards sensitive information using advanced encryption and access control techniques. When uploaded, files are encrypted with AES-256 encryption, ensuring they remain confidential throughout storage. The encrypted documents are stored in a cloud database (e.g., AWS S3 or

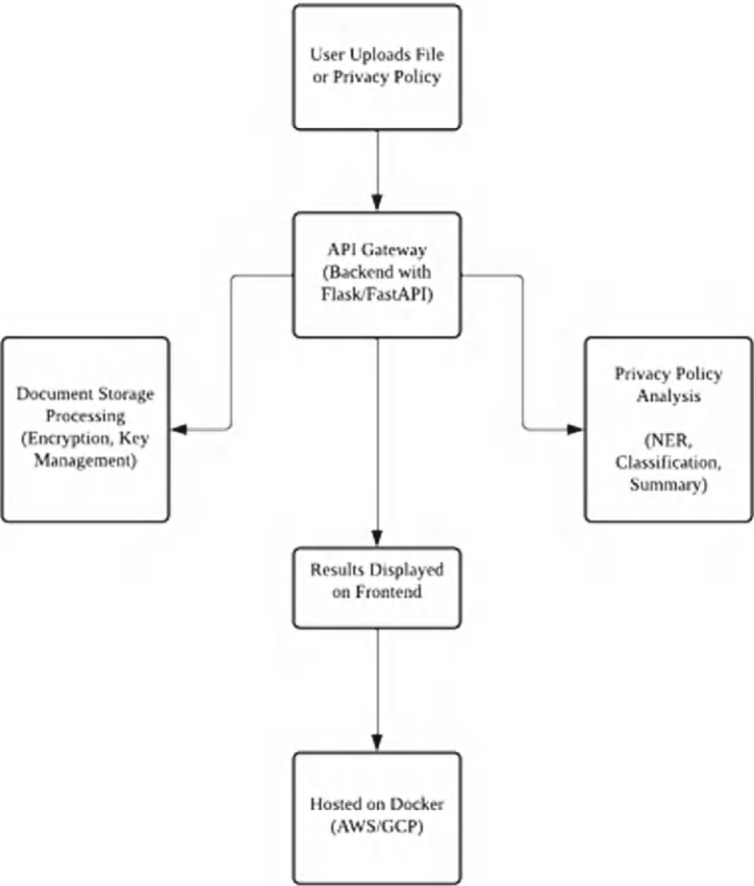


Fig. 1 Method A

MongoDB). The system uses RSA encryption to exchange encryption keys securely during authentication.

Users can generate temporary JWT-based keys to share documents for a limited time, ensuring controlled access. These keys are automatically revoked after expiration, preventing unauthorized access. This time-bound sharing mechanism provides flexibility without compromising data security. RBAC (Role-Based Access Control) governs user actions by assigning permissions based on roles, limiting certain functionalities to specific users. For example, administrators may delete files, while regular users can only view or share documents according to their access level.

The storage system leverages Docker containerization for scalability, allowing it to handle increasing storage demands and deploy quickly in various environments. Backend APIs handle encryption, decryption, and key management, ensuring seamless and secure operations. Through a React. Js-based frontend, users interact with the system to efficiently upload, access, and share documents. This setup ensures that sensitive data is protected and easily manageable, giving users greater control over their files, as shown in Fig. 2.

5 Conclusion and Future Scope

The literature survey emphasizes the importance of privacy-preserving techniques such as AES 256 encryption, JWT-based temporary keys, Role-Based Access Control (RBAC), and Natural Language Processing (NLP) for secure data management. While encryption ensures data confidentiality, RBAC streamlines access control, and JWT provides secure, time-bound access. NLP techniques, such as Named Entity Recognition (NER), enhance privacy transparency by extracting key elements from policies. However, scalability, performance overhead, and dynamic access requirements persist. A dual-platform approach integrating these techniques offers a promising solution to bridge the gap between secure data handling and improved privacy awareness. Many privacy policies are written in languages other than English, limiting accessibility for non-English-speaking users. Expanding the analyzer's capabilities to include multi-language support can address this gap. Models like mBERT or XLM-RoBERTa could be integrated to handle text in various languages, making the tool more inclusive and effective across global audiences [12]. Implementing Blockchain for Versioning and Access Audits: Blockchain technology offers a secure way to log every access and modification to stored documents, ensuring transparency and accountability. Incorporating blockchain into the storage platform would enable tamper-proof tracking of access history and document versions, preventing unauthorized changes. Additionally, it would ensure that all document modifications are recorded, providing users with an extra layer of trust [13].

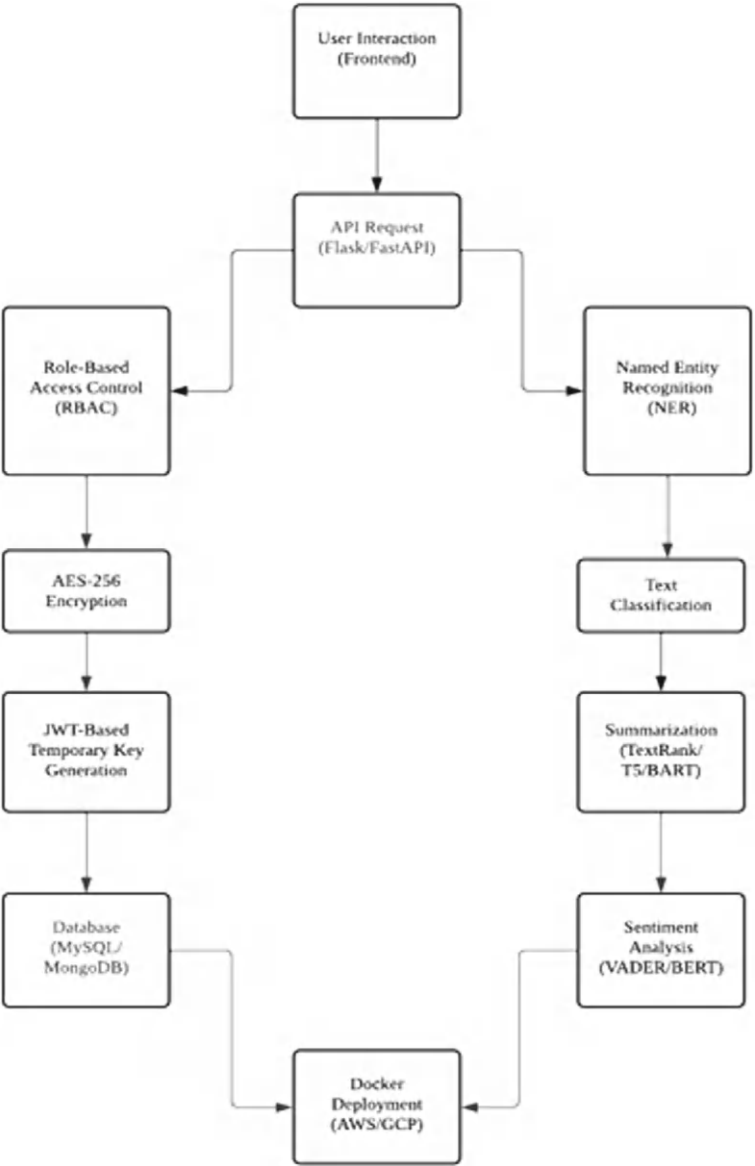


Fig. 2 Method B

References

1. Makhdoom, I., Abolhasan, M., Lipman, J., Piccardi, M., Franklin, D.: PrivySeC: a secure and privacy-compliant distributed framework for personal data sharing in IoT ecosystems **22**(2), 1–26 (2021)
2. Demirer, M., Jiménez Hernández, D.J., Li, D., Peng, S.: Data, privacy laws and firm production: evidence from the GDPR (Working Paper No. 32146). National Bureau of Economic Research (2024)
3. Campbell, T.-A., Eromonsei, S., Afolabi, O.: Efficient compliance with GDPR through automating privacy policy captions in web and mobile applications. *World J. Adv. Eng. Technol. Sci.* **12**(2), 446–467 (2024)
4. Zhang, J.H., Koivumäki, T., Chalmers, D.: Privacy vs convenience: understanding intention-behavior divergence post-GDPR. *Comput. Hum. Behav.* (2024)
5. Hiwale, M., Walambe, R., Potdar, V., Kotecha, K.: A systematic review of privacy-preserving methods deployed with blockchain and federated learning for telemedicine (2023)
6. Hassan, F., Sánchez, D., Domingo-Ferrer, J.: Utility-preserving privacy protection of textual documents via word embeddings (2023)
7. Khan, A.A., Bourouis, S., Kamruzzaman, M.M., Hadjouni, M., Shaikh, Z.A., Laghari, A.A., Elmannai, H., Dhahbi, S.: Security in healthcare industrial internet of things with blockchain (2023)
8. Salunke, M.D., Kumbharkar, P.B., Kumar, S.: Proposed methodology to prevent a ransomware attack. *Int. J. Recent Technol. Eng. (IJRTE)* **9**(1), 2723–2725 (2020)
9. Dhamdhare, P.B., Gond, S.: Semantic trademark retrieval system based on conceptual similarity of text with leveraging histogram computation for images to reduce trademark infringement. *Webology* **18**(5), 4171–4183 (2021)
10. Shinde, B., Gupta, A.: Privacy preserving ANN over cloud
11. An, S.B.: Efficient and secure mobile health system using cloud. *Webology* **18**(5), 4154–4170 (2021)
12. Satpathy, S., Swain, P.K., Mohanty, S.N., Basa, S.S.: Enhancing security: federated learning against man-in-the-middle threats with gradient boosting machines and LSTM. In: 2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Niagara Falls, ON, Canada, pp. 1–8 (2024). <https://doi.org/10.1109/AVSS61716.2024.10672589>
13. Satpathy, S., Mahapatra, S., Singh, A.: Fusion of blockchain technology with 5G: a symmetric beginning. In: Tanwar, S. (ed.) *Blockchain for 5G-Enabled IoT*. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-67490-8_3

An Enhanced Adaptive Steganography Framework Using Inverted LSB and Optimal Patterns (EASIOP)



Sheetal Agrawal and Kshiramani Naik

Abstract Numerous techniques in image steganography have been developed to enhance Stego image qualities, focusing on imperceptibility, capacity, and security. Maintaining more payload capacity without compromising security and imperceptibility is a significant challenge in steganography. To fortify the system further, an adaptive pattern is proposed during the inverted LSB substitution process to enhance imperceptibility. The Advanced Encryption Standard (AES) is implemented before embedding for better security and reliability. The method will find the best bit combination with the least error rate during the message embedding process, improving the efficiency of the inverted Least-Significant-Bit (LSB) replacement technique, which uses a two-bit LSB pattern in the cover image. Before embedding, the bits of the message and the cover image are analyzed. The error rate is calculated using the inverted LSB replacement techniques for various potential patterns. The pattern with the lowest error rate is then selected for embedding. By employing such an adaptive pattern in inverted LSB image steganography, imperceptibility is enhanced substantially. Extensive testing and comparative analysis with previous research demonstrate substantial improvements.

Keywords Image steganography · Inverted LSB · Imperceptibility · Security

1 Introduction

Society now depends far more on technology, especially the Internet, due to the speed at which technology develops. Ensuring data transmission security becomes essential in covert communication. Various techniques, such as cryptography and data hiding, have been widely employed for digital data security. Cryptography involves

S. Agrawal (✉) · K. Naik
Department of IT, VSSUT, Burla, Sambalpur, Odisha, India
e-mail: she.pce@gmail.com

K. Naik
e-mail: kshiramaninaik_it@vssut.ac.in

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
M. Yang et al. (eds.), *Demystifying AI and ML for Cyber-Threat Intelligence*,
Information Systems Engineering and Management 43,
https://doi.org/10.1007/978-3-031-90723-4_39

569

encrypting messages using a specific key and algorithm, transforming plaintext into ciphertext to prevent direct reading by unauthorized parties [1, 2]. In contrast, data hiding focuses on securing messages by concealing them within container media to deceive potential attackers. Watermarking and steganography are two widely used data concealment techniques with different goals: watermarking protects the container media itself, while steganography tries to protect the embedded message [3]. The steganography concept has been used in this discussion, where data concealment is achieved by using different container media (text, audio, video, and photos) [4, 5]. Among these, images have been extensively researched and utilized [6]. Moreover, a challenging task is to enhance the significant characteristics of security, message capacity, and imperceptibility, as these factors are interdependent and influence one another. Increased capacity may compromise imperceptibility and potentially impact security [7]. In response to these challenges, advanced steganographic methods are designed to be more adaptive, incorporating region-based, machine learning artificial intelligence or human visual system (HVS) approaches. The adaptability of watermarking methods is driven by the goal of minimizing error rates caused by message embedding in container images.

In steganography, message embedding occurs in spatial and transform domains [8]. The former, favored for its superior imperceptibility quality and embedding capacity, includes widely used techniques such as LSB Replacement, Difference Expansion, Exploiting Modification Direction (EMD), and Pixel value differences (PVD) [9, 10]. While the traditional LSB substitution method remains popular for its simplicity, ease of development, good imperceptibility, and large capacity, it is considered less secure due to its predictability. Efforts to optimize its performance include focusing on edge areas, utilizing artificial intelligence methods, and employing the inverted bit technique to reduce error ratios [11–16].

Despite the relatively straightforward computation of the inverted LSB substitution image steganography method, its adaptive development has not been explored. Reference [17] noted a significant reduction in bit error ratios and enhanced imperceptibility, but results lacked stability across different container images and messages due to a less adaptive pattern. In their earlier research, combining the inverted bit technique with AES for message randomization resulted in relatively consistent imperceptibility across different covers and messages. However, it did not exceed the outcomes of their later study.

1.1 Contributions

1. **Adaptive Pattern Selection:** Introduces a mechanism to minimize error during message embedding by optimizing three-bit combinations, improving stego image quality.
2. **Inverted LSB Technique:** Enhances security by utilizing the least significant bits more effectively, reducing predictability and increasing resistance to detection.

3. AES Encryption Integration: Strengthens security by encrypting the message before embedding, protecting it from unauthorized access.
4. Comprehensive Performance Evaluation: Assesses performance using MSE, PSNR, and SSIM to evaluate stego image quality objectively.
5. Robustness: Maintains high imperceptibility and capacity across various cover images, making it a reliable solution for secure communication.

The paper is organized into five sections. The first is the introduction, which provides the related concept. Related research, discussing relevant studies. The proposed method is explained through steps and a flowchart. Experimental results with tabulation presentation and comparative analysis of existing work and the conclusion, summarizing key results and future research directions.

2 Related Research

Numerous techniques on steganography on inverted bits are there, like a 2008 study by Yang. This research combines the Optimal Pixel Adjustment Process (OPAP), introduced by Chan and Cheng in 2004, with the Inverted Pattern (IP) LSB method. The IPLSB technique divides the container image and the message into 26–211 equal parts. Each segment of the stego image is analyzed by comparing MSE values on the inverted and normal message bits. If the MSE of the inverted bits is lower, it is classified as an inverted pattern. This inverted pattern is key during message extraction, effectively reducing the error rate and improving imperceptibility by over 1dB per the Peak Signal-to-Noise Ratio (PSNR).

In a study, Akhtar et al. [1] approached an inverted LSB method with a shorter pattern than Yang's work. This method incorporates the AES cryptographic technique for heightened message security. The manipulation of bits in four two-bit patterns (00, 01, 10, and 11) is necessary for the LSB inversion. The approach determines the changes in LSB caused by embedded messages by segmenting the cover picture into these patterns according to the 6th and 7th bits. Despite its simplicity, this technique outperforms the previous approach in terms of performance.

A later advancement by Akhtar et al. [2] involved extending the method with the last three-bit inverted pattern without incorporating the AES cryptographic method. This modification resulted in an improvement of 3dB in the value of PSNR. Where the individual container image leads to the variation in terms of imperceptibility. To adopt Akhtar et al.'s [2], Khadim et al. [3] attempted an inverted LSB method using the image as a complemented message but achieved marginally better outputs than the random LSB and not superior to inverted LSB.

The study of Hussain et al. [4] presents a more recent approach that emphasizes identifying the optimal similarity value to reduce the error ratio. This is achieved through a genetic algorithm (GA) that explores various LSB embedding patterns, including direction, bit order, and color channel options. The proposed method, tested

by embedding in 256×256 -pixel medical images, a 1000-character text message produced PSNR values between 60 and 66 dB.

Miri and Faez [10] introduced a method of bit flipping to embed the eight patterns. This innovative technique segments the cover image into blocks containing two pixels each. A three-bit message pattern is used to flip bits during the embedding process, where the value of PSNR is 47dB using 1.5 bits per pixel capacity (BPP). Another similar LSB inversion method was introduced by Rafrastara et al. [17], featuring a slightly altered pattern that includes the 5th, 6th, and 7th bits, which utilizes the chaotic cryptographic method to enhance the security of the message.

3 Proposed Methodology

The Enhanced Adaptive Steganography Framework (EASIOP) embeds a secret message within a cover image by converting both into binary arrays and encrypting the message using AES with a secret key. A specific three-bit combination, with the 8th bit of the cover image, embeds bits on the least significant bits (LSB) of selected pixel patterns. The algorithm calculates errors for different combinations and chooses the one with the most minor mistake for embedding. For extraction, the stego image is processed to retrieve the LSBs based on the extraction key. Hence, an encrypted message has been decrypted using a secret key, allowing the original message to be reconstructed and sent to the recipient.

3.1 AES (Advanced Encryption Standard) Overview

A symmetric encryption algorithm protects data by transforming plaintext into ciphertext with a fixed 128-bit block size of 128, 192, or 256 bits of key lengths [18]. For instance, consider a plaintext message such as “Hello World” converted into its binary representation. While using a 128-bit AES key, the algorithm executes a series of changes, including byte substitution (Sub Bytes), row shifting (Shift Rows), and column mixing (Mix Columns) across multiple rounds (10 rounds for a 128-bit key). After processing, the resulting ciphertext appears as a random string of bits, making it nearly impossible to reverse-engineer without the key. This potent method guarantees confidentiality, as the shared key is used for encryption and decryption. This enables authorized users to retrieve the original message while safeguarding it from unauthorized access. AES’s efficiency and strong security make it a global standard in various applications, from securing online communications to encrypting sensitive files.

Example:

Key: 2b7e151628aed2a6abf7158809cf4f3c

Plaintext: 3243f6a8885a308d313198a2e0370734

The process transforms the plaintext into ciphertext.

Ciphertext Output: 39f23369a9d9bacfae0c0a5a6c3d1d4f

3.2 Proposed Embedding Algorithm (EASIOP)

Step 1: Read Cover Image (C)

Step 2: Convert Cover Image into an array $C[i, j]$, then to bits $C_0 \dots C_{n-1}, C_n$.

Step 3: Read Secret Message(M), convert it into an array $M[i, j]$ and then to bits $M_0 \dots M_{n-1}, M_n$.

Step 4: Enter Secret Key(K) for message encryption using AES algorithm to generate ciphertext Ct), then to bits $Ct_0 \dots Ct_{n-1}, Ct_n$.

Step 5: Choose any three-bit combination that must include the 8th bit of the Cover Image, i.e. C_6, C_7, C_8 or C_6, C_5, C_8 likewise.

Step 6: Create 16 variables for counting the eight patterns ranging from 000 to 111

- The first 8 variables are for counting the no. of pixels that changes their values during message embedding in LSB for each ($P000, P001, \dots P111$), i.e. P
- Another 8 Variables are for counting the no. of pixels which does not change their values during message embedding in LSB for each ($P'000, P'001 \dots P'111$), i.e., P'.

Step 7: Using Eqs. (1) and (2), embed the message to get P and P':

$$P = \sum_{i=1}^n ci \otimes M \quad (1)$$

$$P' = \sum_{i=1}^n ci \otimes M' \quad (2)$$

where i in C is the LSB Index of each pixel in the cover image, n is the number of bits, and i in M is the message bit Index is a message that is reversed. \otimes And bit performs an XOR operation.

Step 8: Using Eq. (3), calculate error summation(e) for those eight-bit patterns of three-bit combination where PS is the minor error on each pattern and "i" is the no. of the pattern.

$$e = \sum_{i=1}^8 P_s \begin{cases} P, P' < P \\ P', P' > p \end{cases} \quad (3)$$

Step 9: Repeat steps 3–8 for the other 21 3-bit combinations, ensuring that the 3rd bit corresponds to the 8th of the cover image, as per the LSB technique.

Step 10: Select the best three-bit pattern with the slightest error, such as C6, C7, or C8.

- Embed the pattern data, which is K, the extraction key having the optimal three-bit combination,
- Use an eight-bit pattern ranging from 000 to 111, where 111 signifies inverted, and 000 signifies not inverted.

Step 11: Convert them into decimals, then reshape the Stego array to get the Stego Image (S).

All these steps are shown in Fig. 1.

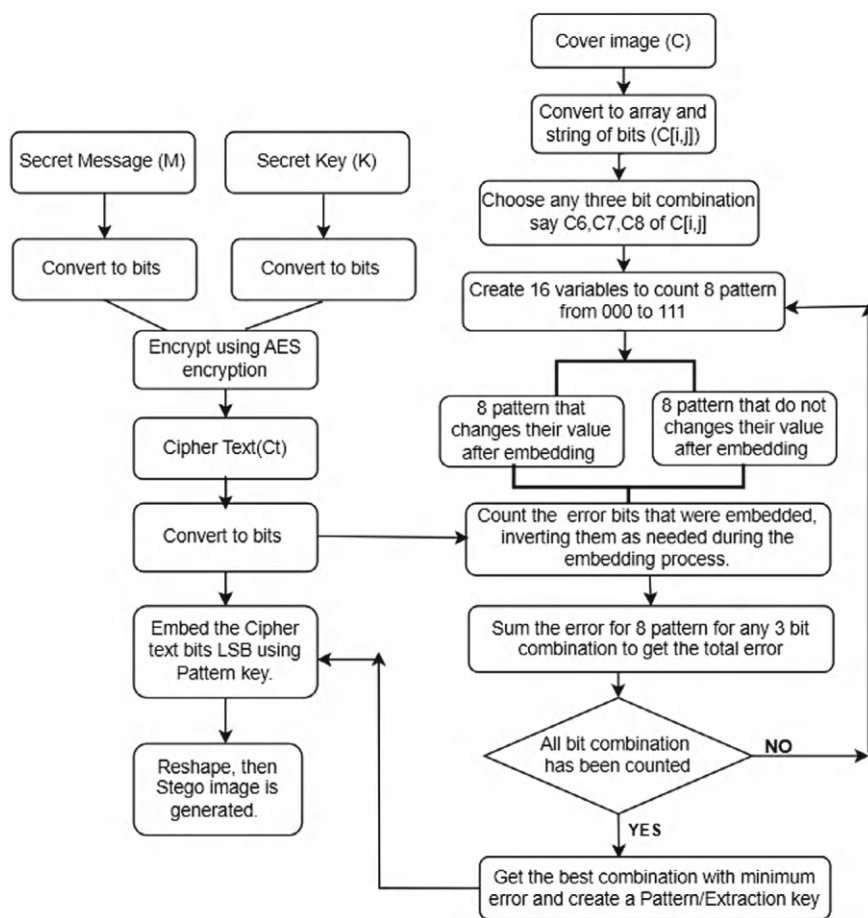


Fig. 1 Flowchart of proposed embedding process

After the completion of the embedding process, the receiver will receive the embedding message, and the extraction process begins. These inputs should be there for extraction: the stego image(S), the variable key K as the extraction key, and the Secret Key as the decryption key.

3.3 Proposed Extracting Algorithm (EASIOP)

Step 1: Read Stego Image(S).

Step 2: Convert Stego Image into an array $S[i, j]$, then to bits $S_0 \dots S_{n-1}, S_n$.

Step 3: Read the Extraction Key (K) to get the bit combination and inverted bit pattern.

Step 4: Extract data using the Stego Image (S) LSB based on the Extraction Key (K).

Step 5: To generate an encrypted message, convert each eight-bit message into an ASCII number.

Step 6: To generate the decrypted Secret message, input the Secret Key and use the AES method.

Step 7: The decrypted message is converted into characters and sent to the intended recipient as a plain text message.

All the above steps are shown in Fig. 2.

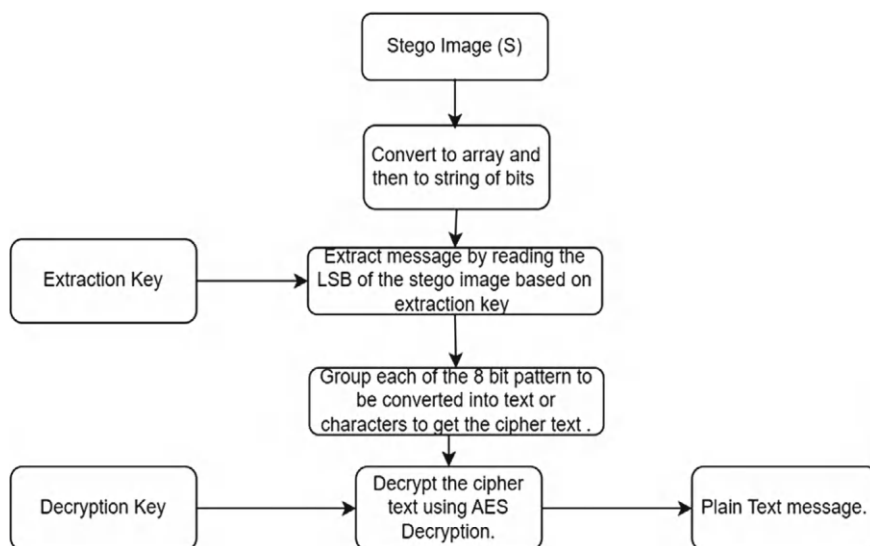


Fig. 2 Flowchart of proposed extraction process

4 Experimental Results

The proposed method was assessed with four gray images, each of size 256×256 pixels from the database of SIPI Image. as illustrated in Fig. 5 as (a) Cameraman, (b) Baboon, (c) Lena, and (d) Pepper. The data sizes used were 1024 bytes and 2048 bytes from the page of lipsum.com. The proposed algorithm (EASIOp) was implemented using MATLAB R2013a, 64-bit version.

4.1 Perspective of Results

The results demonstrate that the proposed adaptive steganography framework significantly enhances the stego image qualities while maintaining superior levels of security and imperceptibility. The following aspects were analyzed:

Imperceptibility: The qualities of the stego images are assessed against various cover images to ensure that the embedded message will not change the image's visual appearance. This is vital in steganography to avoid detection, as shown in Figs. 3, 4 and 5.

Security: AES encryption before embedding the message ensures that even if the stego image is intercepted, the hidden message remains secure. The randomness leads to an extra layer of security.

Error Minimization: An adaptive pattern selection process aimed to minimize the error ratio during message embedding to maintain the qualities of the stego image.

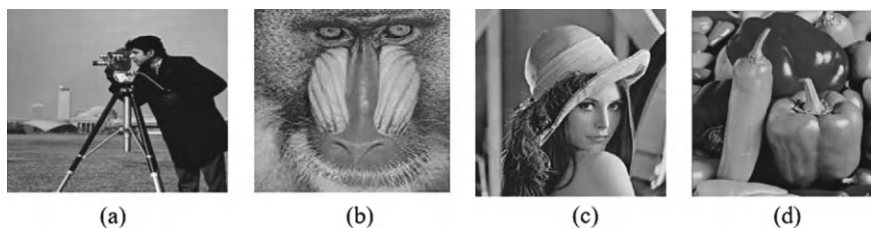


Fig. 3 The original cover images **a** cameraman, **b** baboon, **c** lena, **d** pepper

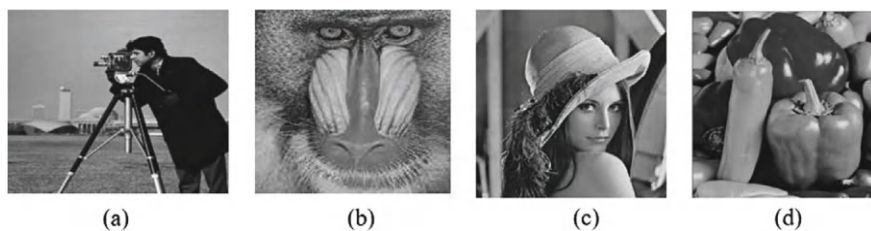


Fig. 4 The stego images after AES encryption and embedding application to the cover images

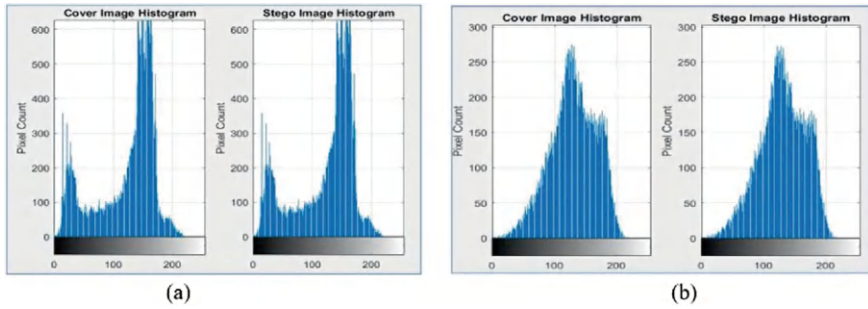


Fig. 5 Histograms of cover images and stego images **a** cameraman **b** baboon

4.2 Metrics Used for Evaluation

The following metrics were calculated to evaluate the proposed method quantitatively.

Mean square error (MSE): It measures the average squared difference between pixel values of the cover and stego images. Equation 4 is the general equation for the calculation of MSE:

$$MSE = \frac{1}{M \times N} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} [S(i, j) - C(i, j)]^2 \quad (4)$$

here, M and N are rows and column numbers, and S and C are stego and cover images respectively

The below 10 are excellent quality, and the lower values indicate better imperceptibility.

Peak Signal to Noise Ratio (PSNR): The ratio of the maximum possible value of a signal to the noise that impacts its quality is expressed in decibels (dB). Equation 5 represents the general formula for calculating PSNR.

$$PSNR = 10 \times \log_{10} \left(\frac{cmax^2}{MSE} \right) \quad (5)$$

$Cmax$ is the maximum possible value of the original image. MSE is the Root Mean Square difference between the two images. Higher values indicate better image quality and lower distortion. Above 40 is ideal, indicating high image fidelity.

Structural Similarity (SSIM) Index: Evaluates the perceived quality by considering changes in structural information, luminance, and contrast. Equation 6 is the general calculation of SSIM.

$$SSIM(I_C, I_S) = \frac{(2\mu_{I_C}\mu_{I_S} + C_1)(1\sigma_{CS} + C_2)}{(\mu_{I_C}^2 + \mu_{I_S}^2 + C_1)(\sigma_{I_C}^2 + \sigma_{I_S}^2 + C_2)} \quad (6)$$

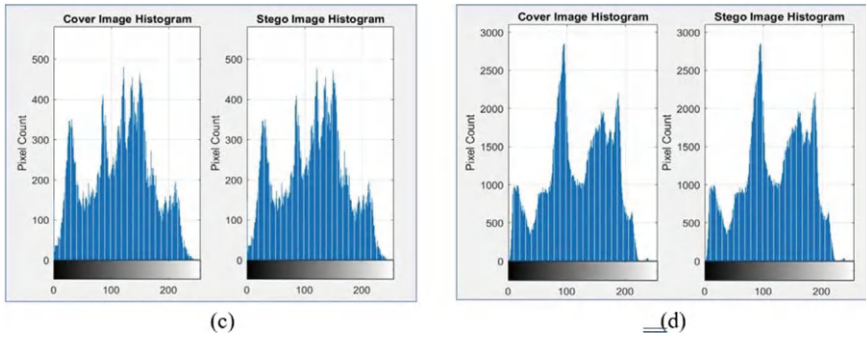


Fig. 6 Histogram of cover images and stego images **c** lena and **d** pepper

where $C1 = (k_1 L)^2$ and $C2 = (k_2 L)^2$, $L = 255$ for 8-bit images), $k_1 = 0.01$, $k_2 = 0.03$ μ_C and μ_S are the mean intensity values of Image C and S, σ_C^2 and σ_S^2 are the variance of C and S, respectively, σ_{CS} is the covariance of C and S.

Values range from 0 to 1; closer to 1 indicates higher similarity and better imperceptibility.

4.3 Histogram Analysis

Histograms of the cover and the stego images were analyzed to assess the impact after message embedding on the pixel distribution. A histogram plots the frequency of pixel values, allowing us to compare the intensity distributions visually. Ideally, the histogram of both should resemble, indicating minimal distortion and a high level of imperceptibility, as shown in Figs. 5 and 6, respectively.

The results in Figs. 5 and 6 show that the histograms of the stego images exhibit slight variations compared to the cover images, which confirms our adaptive pattern selection's effectiveness in maintaining the stego images' overall quality.

5 Summary of Results

Tables 1, 2 and 3 present the mentioned metrics (MSE, PSNR, and SSIM) values for the given cover image for message sizes of 4096 and 1024 and a comparison of PSNR with existing works, respectively. The embedding of messages is done with the adaptive pattern, as shown in the column named the optimal combination bit, which reduces the errors. When embedding messages into the cover image, utilizing a three-bit combination offers more options to minimize mistakes while searching for the most adaptive pattern. The PSNR value shows that the stego image quality grows directly with the decrease in error. Interestingly, the suggested method also raises the

Table 1 Evaluation of imperceptibility with message size of 4096 bytes

Image	Proposed method			
	MSE	PSNR (dB)	SSIM	The best combination
Baboon	0.128211	57.017684	0.998645	3,1,8
Cameraman	0.116065	57.445671	0.998071	7,4,8
Lena	0.345789	52.497342	0.996712	6,5,8
Peppers	0.116783	57.418932	0.998275	5,1,8
Average	0.176712	56.094907	0.997925	

Table 2 Evaluation of imperceptibility with a message size of 1024 bytes

Image	Proposed Method			
	MSE	PSNR (dB)	SSIM	The best combination
Baboon	0.010917	69.337186	0.998988	5,1,8
Cameraman	0.006568	68.312346	0.989678	6,2,8
Lena	0.020687	64.140863	0.998771	6,3,8
Peppers	0.012853	65.553844	0.987914	6,2,8
Average	0.012175	66.833437	0.983837	

Table 3 Comparison of PSNR (dB) with existing works

Image	Huang (2010)	Akhtar et al. [1]	Akhtar et al. [2]	Bhardwaj and Sharma [13]	Karakus and Avci [8]	Proposed method
Baboon	40.876	43.892	45.872	41.789	60.982	68.759
Cameraman	41.500	43.800	45.500	42.500	66.438	69.832
Lena	41.342	44.679	46.567	42.123	61.435	69.337
Peppers	42.000	44.500	46.200	43.000	65.876	70.216

SSIM value because the human visual system—rather than error or noise—is used to measure SSIM. Also, Table 3 shows that the proposed method achieves superior PSNR values across all test images, with a maximum PSNR of 70.216 dB for the Pepper image and 69.337 dB for the Lena image.

6 Conclusion and Future Scope

The results demonstrate that our methods will closely match stego and cover images. The proposed approaches achieve the highest PSNR and the lowest MSE, allowing sensitive information to be concealed without significantly affecting the image’s perceived quality. With the application of the adaptive pattern, inverted LSB was

performed, and the optimal bit combination with the minor error was selected during message embedding. This optimal bit combination may vary depending on the cover media and embedded message. As a result of the communication channel's numerous susceptibilities, which will be used for stego image communication, future work can expand by considering attacks and sounds. Moreover, experiments with color images will be conducted, and optimization will be done using AI methods such as blockchain technologies, as it is fundamentally based on the AES concept utilized in the proposed algorithm.

References

1. Akhtar, N., Johri, P., Khan, S.: Enhancing the security and quality of LSB-based image steganography. In: Proceedings - 5th International Conference on Computational Intelligence and Communication Networks, CICN 2013, pp. 385–390 (2013)
2. Akhtar, N., Khan, S., Johri, P.: An improved inverted LSB image steganography. In: Proceedings of the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT 2014, pp. 749–755 (2014)
3. Kadhim, I.J., Premaratne, P., Vial, P.J., Halloran, B.: A comprehensive survey of image steganography: techniques, evaluations, and trends in future research. *Neurocomputing* **335**, 299–326 (2019). <https://doi.org/10.1016/j.neucom.2018.06.075>
4. Hussain, M., Wahab, A.W.A., Idris, Y.I.B., Ho, A.T.S., Jung, K.-H.: Image steganography in spatial domain: a survey. *Signal Process. Image Commun.* **65**, 46–66 (2018)
5. Cheddad, A., Condell, J., Curran, K., Mc Kevitt, P.: Digital image steganography: survey and analysis of current methods. *Signal Process.* **90**, 727–752 (2010). <https://doi.org/10.1016/j.sigpro.2009.08.010>
6. Li, Z., He, Y.: Steganography with pixel-value differencing and modulus function based on PSO. *J. Inf. Secure. Appl.* **43**, 47–52 (2018). <https://doi.org/10.1016/j.jisa.2018.10.006>
7. Luo, W., Huang, F., Huang, J.: Edge adaptive image steganography based on LSB matching revisited. *IEEE Trans. Inf. Forensics Secur.* **5**, 201–214 (2010)
8. Karakus, S., Avci, E.: A new image steganography method with optimum pixel similarity for data hiding in medical images. *Med. Hypotheses* **139**, 109691 (2020)
9. Maniriho, P., Ahmad, T.: Information hiding scheme for digital images using difference expansion and modulus function. *J. King Saud Univ. - Comput. Inf. Sci.* **31** (2019)
10. Miri, A., Faez, K.: Adaptive image steganography is based on a transform domain via a genetic algorithm. *Optik (Stuttg)* **145**, 158–168 (2017)
11. Ardiansyah, G., Sari, C.A., Setiadi, D.R.I.M., Rachmawanto, E.H.: The hybrid method uses 3-DES, DWT, and LSB to secure the image steganography algorithm. In: 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), pp. 249–254 (2017)
12. Bai, J., Chang, C.-C., Nguyen, T.-S., Zhu, C., Liu, Y.: A high payload steganographic algorithm based on edge detection. *Displays* **46**, 42–51 (2017)
13. Bhardwaj, R., Sharma, V.: Image steganography based on complemented message and inverted bit LSB substitution. *Procedia Comput. Sci.* 832–838 (2016)
14. Chan, C.K., Cheng, L.M.: Hiding data in images by simple LSB substitution. *Pattern Recogn.* **37**, 469–474 (2004)
15. El-Emam, N.N., Al-Zubidy, R.A.S.: A new steganography algorithm is used to conceal a large amount of data. *J. Syst. Softw.* **86**, 1465–1481 (2013)
16. Fridrich, J., Goljan, M., Du, R.: Detecting LSB steganography in color and gray-scale images. *IEEE Multimed.* **8**, 22–28 (2001)

17. Rafrastara, F.A., Prahasiwi, R., Setiadi, D.R.I.M., Rachmawanto, E.H., Sari, C.A.: Image steganography using inverted LSB based on 2nd, 3rd and 4th LSB pattern. In: 2019 International Conference on Information and Communications Technology, ICOIACT (2019)
18. Biryukov, A., Wagner, D.: Security evaluation of AES: a mathematical and computational overview. J. Cryptol. (2011)

Ethical Considerations and Future Perspectives in AI Security

AI-Powered Grief Technology: The Ethical Implications of AI in Privacy



Sonal Mahapatra

Abstract The rise of grief technology, including AI-driven tools that simulate interactions with deceased loved ones, offers a novel approach to coping with loss, especially during times when traditional support systems, such as counseling or family networks, may be inaccessible. While digital avatars and chatbots provide a sense of connection and comfort to the bereaved, especially during crises like the COVID-19 pandemic, these technologies also raise significant psychological and ethical concerns. AI-powered grief technology can offer temporary emotional relief and a sense of closure by recreating familiar behaviors and conversations about the deceased. However, its potential to impede natural grieving processes or create emotional dependency highlights the need for careful consideration. Ethical issues regarding data privacy, consent, and the commercialization of mourning also complicate using these AI tools. This paper explores the benefits and limitations of AI-powered grief technology, critically examining its role in the mourning process, its psychological impact, and its ethical implications. The study addresses the delicate balance between AI technology's potential to assist with grief and the risks of commodifying personal loss through a combination of psychological frameworks, case studies, and ethical analysis. Ultimately, it argues for a mindful approach to AI-powered grief technology that respects the profoundly personal nature of mourning while acknowledging the potential for innovation in emotional support.

Keywords AI · Privacy · Grief technology · Bereavement · Emotional support · Ethics · Ethical implication

S. Mahapatra (✉)

North Carolina School of Science and Mathematics, Morganton, NC, USA

1 Introduction

1.1 What is AI-Based Grief Technology

In recent years, the emergence of AI-powered grief technology has introduced novel approaches to coping with the profound emotional distress associated with the loss of a loved one. Traditionally, grief support has been facilitated through face-to-face counseling, religious rituals, and familial networks, all providing mourners with a sense of community and structured avenues for emotional expression. However, as technological advancements continue to influence various aspects of human life, the rise of AI-based digital tools has begun to reshape bereavement processes, offering alternative methods for individuals to navigate their grief. These digital solutions leverage artificial intelligence, natural language processing, and machine learning to generate interactive experiences that simulate the presence of the deceased, allowing mourners to engage in ongoing conversations with AI-driven recreations of lost loved ones [1]. AI grief technology relies on vast amounts of personal data, including text messages, social media interactions, voice recordings, and video archives, to construct digital personas that reflect the deceased's personality traits, communication styles, and behavioral patterns. Advanced machine learning algorithms analyze this data to generate responses that mimic how the individual would have spoken, allowing for interactions that feel eerily familiar to those grieving the loss. Some platforms, such as HereAfter AI and Replika, provide users with AI-generated avatars capable of maintaining ongoing conversations, responding to prompts with personalized language, and adapting based on user engagement over time. These digital recreations present a new form of memorialization, offering mourners a way to preserve emotional connections with lost loved ones in ways that traditional forms of remembrance, such as photographs, letters, or recorded messages—cannot replicate [2]. The growing reliance on AI-powered grief technology signifies a paradigm shift in how society conceptualizes and processes loss. In contrast to conventional grief practices, which emphasize acceptance and the gradual integration of loss into one's life, digital grief tools blur the boundaries between past and present, allowing individuals to engage with representations of the deceased in real-time. While this innovation may provide immediate emotional relief by mitigating feelings of isolation and loss, it raises critical questions regarding the long-term psychological, ethical, and social implications of integrating AI into the mourning process [3]. From a psychological perspective, AI-driven grief technology aligns with the *continuing bonds* theory, which posits that individuals do not necessarily sever ties with the deceased but reshape their relationship with them over time. By enabling mourners to interact with digital representations of their loved ones, these technologies may serve as tools for maintaining meaningful connections during the grieving process. However, the risk of psychological dependency cannot be overlooked. Excessive reliance on AI recreations may inhibit the natural trajectory of grief by fostering attachment to artificial simulations rather than encouraging acceptance of loss. Additionally, the authenticity of AI-generated interactions remains questionable. At the same time, these digital

entities can mimic language patterns and behavioral traits; they lack the consciousness, emotional depth, and unpredictability that characterize human relationships, potentially leading to a distorted sense of connection that fails to provide genuine closure [4]. Beyond psychological concerns, AI-driven grief technology presents complex ethical dilemmas regarding data privacy, consent, and commercialization. Creating digital personas necessitates access to deeply personal information, often without explicit permission from the deceased. This raises concerns about who has the right to authorize using a person's digital legacy and whether such data should be utilized to construct AI-driven simulations. Furthermore, grief technology has increasingly become a commercial enterprise, with companies offering subscription-based models or tiered pricing structures to access digital avatars. This commodification of grief raises fundamental ethical questions about how much mourning should be monetized and whether profit prioritization may compromise users' emotional well-being [5]. As AI-powered grief technology evolves, critically examining its potential benefits and inherent limitations is imperative. While these tools offer innovative ways to process loss, their integration into contemporary mourning practices necessitates ongoing discourse on their psychological impact, ethical ramifications, and societal consequences. Ultimately, adopting grief technology must be cautiously approached, ensuring that preserving emotional connections does not come at the expense of moral responsibility and long-term emotional healing [6].

1.2 Relevance of Grief Technology

Grief technology is relevant in addressing some pressing challenges faced during the mourning process, especially in an era of limited access to traditional mental health services. The COVID-19 pandemic, for example, exacerbated the already significant shortage of mental health professionals, leaving many individuals without adequate support during critical moments of grief. For some, interacting with a digital avatar of the deceased, through text, voice, or video, can offer a semblance of connection, enabling them to revisit cherished memories, express unresolved emotions, and experience a continued relationship with the departed. Despite the potential benefits, grief technology raises significant psychological and ethical concerns. While it may provide temporary emotional relief, there is the risk that excessive reliance on digital re-creations may impede the natural progression of grief and hinder long-term emotional healing.

1.3 Objective

This paper examines AI-powered grief technology's impact, exploring its potential benefits and inherent limitations. Analyzing their psychological, ethical, and societal implications is essential as AI-driven tools, such as grief chatbots and digital

afterlife simulations, become increasingly integrated into mourning practices. While these technologies may offer comfort by simulating conversations with deceased loved ones or providing grief support, they also introduce complex questions about authenticity, emotional well-being, and the evolving nature of human remembrance. Through an analysis of relevant psychological frameworks, such as continuing bonds theory, which examines how individuals maintain connections with the deceased and the dual process model of coping with bereavement, this paper will evaluate how AI influences the grieving experience. Additionally, it will consider critical ethical concerns, particularly regarding data privacy, user consent, and the commercialization of mourning. As AI grief tools rely on personal data, including voice recordings, messages, and social media history, the risks of data exploitation and consent ambiguity must be carefully examined. Furthermore, the rise of grief technology as a commercial enterprise raises concerns about whether these tools serve the best interests of the bereaved or prioritize profit over emotional well-being. By assessing the psychological and ethical dimensions of AI-powered grief technology, this study aims to analyze its role in contemporary mourning practices comprehensively. It will explore whether such technologies enhance or hinder the bereavement process, offering insights into how AI reshapes emotional coping mechanisms in the digital age. Ultimately, this paper will consider the broader implications of grief tech, questioning whether the integration of AI into mourning rituals represents a meaningful evolution in how society processes loss or commodification of one of the most profound human experiences.

2 End-To-End Process

Grief technology represents an outlet for addressing bereavement, acknowledging the highly individualized nature of grief experiences. When an individual experiences the unexpected death of a loved one, the resulting absence often produces intense feelings of depression and psychological distress, mainly due to unresolved interpersonal issues and unmet expectations. Grief technology addresses these challenges by offering digital closure through virtual representations of the deceased [7]. These digital constructs, developed through advanced algorithms and data analytics, simulate aspects of the deceased's personality and behavioral patterns, enabling interactive communication. The idea of living and not being able to get closure with the dead may put the person currently alive in a depressive state. The overwhelming emotion is often complex for many, and those who want to manage it usually need support. A therapist could provide this support; however, their availability might not align [8]. With more people realizing the importance of mental health, the demand for licensed mental health therapists has skyrocketed in recent years, especially after the COVID-19 pandemic began. Unfortunately, the United States is facing a critical shortage of these professionals, leaving many individuals and families in distress without the help they need. This shortage is creating a ripple effect throughout the

healthcare system, with it becoming increasingly difficult for people with mental health issues to access the care they need [9].

2.1 Positive Grief Tech Assistance

When people grieve, especially a loss, they need immediate help. Grief technology does not aim to restore the deceased to life. Still, it facilitates a simulated interaction that mirrors past communications, such as text messages or voice recordings, in a more dynamic and personalized format. By providing a means for users to engage with these virtual representations, grief technology offers a potential avenue for emotional processing and psychological comfort. A digital representation of the deceased loved one may contribute to a sense of ongoing connection, thereby assisting individuals in managing their grief and navigating the emotional adjustment process. Henle's interview with Vox shows the positive impact of grief tech after losing a loved one. 'If I'm having a tough day, it does give me better advice than Google. It seems like it takes all the best bits and puts great wisdom into one place, like a great friend or therapist,' says Henle, whose grief experience was expensive and disappointing. While some people have good experiences with grief counselors, Henle did not. "ChatGPT felt more human to me than this therapist," she says. The avenue of support this technology provides leaves a strong mark on Henle. The comfort provided to her in a time of need by the grief tech did not go unnoticed [2]. The experience of going down a rabbit hole largely depends on the individual. Each situation is unique, and people have distinct ways of processing emotions and navigating challenges, particularly in grief. A person's coping mechanisms for losing a loved one, as well as their understanding of death, are profoundly shaped by their personality and past experiences. For some, grief can be an overwhelming and consuming process, while for others, it might be something they learn to navigate with time and support. However, there is a delicate balance to be struck. While grief tech may offer valuable support in the immediate aftermath of loss, there is the potential for people to become overly attached to these digital representations of the deceased. Instead of accepting the loss and finding ways to move forward, they might develop a dependency on this form of virtual communication. This can make it difficult for them to recognize when it's time to let go and begin to heal. According to Bresco de Luna [4], ... the two-way interaction made possible by grief bots, combined with the material quality of the messages left by the digital version of the loved one, might be problematic, particularly in mourners with avoidance/denial patterns or complicated grief symptoms. While griefbots might be helpful as part of grieving rituals, especially in the initial moments after death, as a way of communicating with the deceased one last time... problems could arise if the virtual relationship with the dead becomes a chronic coping strategy of denial.

2.2 *The Fine Line*

There's a fine line between utilizing grief tech as a temporary support system and allowing it to become an emotional crutch. If a person leans too heavily on these digital recreations, they risk blurring the boundary between reality and the virtual world. This attachment can interfere with their ability to process grief and move forward healthily fully. For example, someone might become emotionally dependent on receiving messages from an AI version of their loved one. If, at any point, the AI fails to provide the expected comfort or mimics the deceased in an unsettling way, the individual could find themselves upset or frustrated. They may experience renewed emotional pain and hurt over something the AI said or did, even though it's not their loved one. Ultimately, while grief tech can be incredibly helpful in easing the initial sting of loss, individuals need to recognize its limitations. It is not a substitute for the grieving process or a way to avoid the natural, necessary healing journey [10–12]. Understanding the boundaries of life and death is not a recent phenomenon. While grief tech has approached grief from a different angle, it has long been explored by influential figures such as Sigmund Freud. Freud's classic psychotherapy theory grapples with the complexities of life and death, offering valuable insights that enrich our understanding of grief tech. Sigmund Freud's concept of the "uncanny" delves into a psychological space where familiarity and unfamiliarity overlap, creating a feeling of discomfort or eeriness. This term, derived from the German "Heimlich," carries a dual meaning that adds to its complexity. On the one hand, "Heimlich" refers to something familiar or homely that should evoke comfort. On the other hand, it also hints at something hidden and mysterious. "The German word *unheimlich* is the opposite of *heimlich*, *heimisch*, meaning "familiar," "native," "be-longing to the home"; and we are tempted to conclude that what is "uncanny" is frightening precisely because it is *not* known and familiar" (Freud 2). Freud used this idea to explain why certain experiences, images, or technologies might feel familiar and alien, producing an eerie sensation. When applied to grief technology, the concept of uncanny offers a powerful lens to understand why such tools provoke a blend of comfort and discomfort. Grief tech, which uses AI to simulate conversations with deceased loved ones, creates a peculiar emotional space. On the surface, these interactions may provide comfort because they restore the familiar—allowing users to reconnect with those they've lost. Yet beneath that surface is an unsettling layer: the person is no longer alive, and the interaction is artificial, based on data, not true human presence. This duality is at the heart of the uncanny. Freud's interpretation helps explain why grief tech can feel so eerie. It taps into a sense of homecoming as users "revisit" their relationship with the deceased, but it simultaneously clarifies that something is amiss. The person is no longer there, yet they are present in a digital, fragmented form. The conversations generated by AI often reflect memories, patterns of speech, and data collected while the person was alive. However, this simulation is never truly the person. It is both them and not them—a facsimile that can evoke strong emotional reactions precisely because it blurs the line between life and death, presence and absence. Moreover, Freud's notion of the uncanny also

connects to unspoken, unfinished, or repressed feelings, especially in grief. Many people question the most emotionally charged moments of their relationship after losing a loved one. Grief tech offers a simulated space to revisit these moments. The user can say the things they never had the chance to say, creating a sense of closure. Yet, this sense of closure comes with its tension: the comfort of finally expressing those feelings is undermined by the knowledge that the conversation is, ultimately, with an AI, not the actual person. This creates an emotional paradox, like the uncanny—where something familiar is repackaged in a new way. The familiarity comes from the echoes of the person's voice, their mannerisms, and shared memories, while the eeriness comes from knowing it is all artificially constructed. The AI responses may also be incomplete or slightly off, further heightening the uncanny sensation, as it makes the user aware that something crucial, the human consciousness, is missing. This aspect of grief tech also taps into future-oriented anxieties. For instance, what happens when AI simulates a person so convincingly that it becomes hard to distinguish between what is real and what is artificial? The uncanny feelings tied to grief tech may grow even more pronounced as technology advances, pushing the boundaries of real emotional connection. Freud's concept suggests that the more human-like these technologies become, the deeper the sensation of the uncanny grows, providing comfort while simultaneously emphasizing what has been lost forever. Using sophisticated technology allows the situation to be replicated realistically, sometimes leaving the user questioning this reality. "Then he laughed—and for a moment, I forgot I wasn't speaking to my parents, but to their digital replicas" (Jee). In this way, grief tech becomes not only a tool for coping with loss but also a profound reflection of human psychology and our relationship with death, memory, and technology. Freud's uncanny ability helps explain why such technologies stir comforting and unsettling emotions, as they challenge our deepest perceptions of life, death, and human connection.

2.3 *The Ethical Side*

While grief technology, such as AI-driven chatbots that simulate conversations with deceased loved ones, can offer comfort and connection, its commodification raises ethical concerns. The emergence of AI therapists, particularly in the form of grief chatbots, quickly demonstrated its profitable potential. "Digital afterlife companies are, after all, a profit-seeking industry based on the use of digital remains and the monetization of the digital afterlife of Internet users, something that may not necessarily be in the best interests of the bereaved or suited to their psychological needs during their grieving process" (qt. In de Luna [4]). This commodification represents a troubling shift in how society approaches mourning, turning the deeply personal experience of grief into a marketable product. Instead of allowing individuals to process loss in an organic and meaningful way, grief technology risks transforming mourning into a transactional experience. Websites like HereAfter AI exemplify this shift, offering AI-driven grief support through tiered pricing plans:

\$3.99 per month for an essential subscription, \$7.99 for unlimited access, and one-time payments ranging from \$99 to \$199 for varying levels of service (“HereAfter AI—Plans and Pricing”). By reducing grief to a paid subscription service, akin to entertainment platforms like Netflix or Spotify, such models risk diminishing the significance of loss, replacing human remembrance and emotional depth with algorithmic interactions. Beyond concerns about commercialization, grief tech raises fundamental questions about authenticity. Genuine mourning is a profoundly human process involving complex emotions, personal reflection, and meaningful connections. When grief becomes mediated through AI, it risks flattening into scripted conversations that mimic but fail to replicate actual emotional presence. While these technologies may offer momentary comfort, they also risk detaching people from the communal and relational aspects of mourning, disrupting the natural healing process through human connection. Furthermore, reliance on grief technology could inadvertently discourage individuals from seeking support through traditional means, such as therapy, support groups, or personal reflection, reinforcing a culture that prioritizes convenience over genuine emotional processing. The commodification of grief tech also invites a broader discussion on the balance between technological innovation and preserving the human aspects of grief. While AI can serve as a tool for support, it should not come at the cost of commercializing loss or prioritizing profit over emotional well-being. Grief is a profoundly personal experience that should remain sacred and free from corporate interests. The introduction of AI into this space poses the question: to what extent should technology intervene in the grieving process? Is it ethical to use the sadness of individuals to generate profit? While grief technology may offer solace, its increasing monetization demands critical reflection, ensuring that the pursuit of profit does not overshadow the profoundly human need for authentic mourning and remembrance [13, 14].

3 Psychology

Grief, as a deeply emotional and psychological experience, involves significant changes in brain function and structure that influence how individuals process loss. Recent studies on the psychology of grief reveal that mourning can profoundly affect the brain, particularly in regions responsible for memory, emotional regulation, and decision-making. According to the American Psychological Association, “grief can significantly alter brain function, particularly in areas related to memory, emotional regulation, and self-control” (APA). This shift in brain activity can make the grieving process emotionally challenging and cognitively demanding, as individuals may struggle with maintaining focus, decision-making, and managing day-to-day tasks. The changes in brain function during grief are particularly notable in the hippocampus and prefrontal cortex. The hippocampus, a region linked to memory processing, is significantly impacted by grief, which can result in intense emotional memories of the deceased flooding the individual’s consciousness. As the APA notes, “grief can

be seen as a disruption in memory, where memories of the loved one can feel especially vivid or intrusive” (APA). This disruption in memory processing can make it difficult for individuals to detach from the intense emotions tied to their loss, which may hinder their ability to move through the stages of grief healthily. Furthermore, the prefrontal cortex, which plays a critical role in emotional regulation and decision-making, is also affected during the grieving process. The APA explains that “grief can make it harder for people to manage their emotions or make everyday decisions because the prefrontal cortex becomes less active during mourning” (APA). This decrease in emotional regulation capacity can lead to feelings of overwhelm, confusion, and difficulty navigating daily life. Grieving individuals may experience fluctuations in their mood, struggle to make decisions, or find it hard to regulate intense emotions, which are natural consequences of the brain’s response to loss.

The physiological changes brought about by grief can be further exacerbated when an individual relies heavily on external tools, such as grief technology, to cope with their emotions. AI-powered digital recreations of the deceased could reinforce these neural responses by triggering intense emotional reactions that may stall the grieving process. While digital tools may offer temporary comfort, they could also perpetuate intrusive memories and emotions, further disrupting the brain’s natural attempts at emotional regulation and memory processing. As such, grief technology must be approached cautiously, as it may support and hinder the complex cognitive processes involved in grieving. Understanding grief’s psychological and neurological effects is essential for evaluating the role of grief technology. As individuals navigate their mourning process, it is crucial to recognize that the changes in brain function brought about by grief are complex and multifaceted. Technology, while offering potential benefits, must be used responsibly to ensure that it does not exacerbate the psychological challenges of mourning. Instead, it should support individuals in a way that complements the brain’s natural healing processes, allowing for a more balanced and healthier journey through grief, as shown in Fig. 1.

4 The Technology

Integrating artificial intelligence (AI) into bereavement practices has introduced a new paradigm in how individuals process loss, reshaping the traditional methods of grieving through digital simulations of deceased loved ones. Grief technology employs advanced computational techniques, including natural language processing, generative AI, and deep learning, to construct highly detailed virtual personas capable of mimicking human speech, mannerisms, and facial expressions. By drawing upon an individual’s digital footprint, comprising emails, text messages, social media interactions, voice recordings, and videos, these AI-driven systems recreate a sense of presence, offering users the opportunity to continue engaging with their loved ones posthumously. However, while this technology presents a novel approach to coping with grief, it also raises a series of ethical, psychological, and philosophical complexities that challenge the nature of mourning, identity, and digital agency. At the core

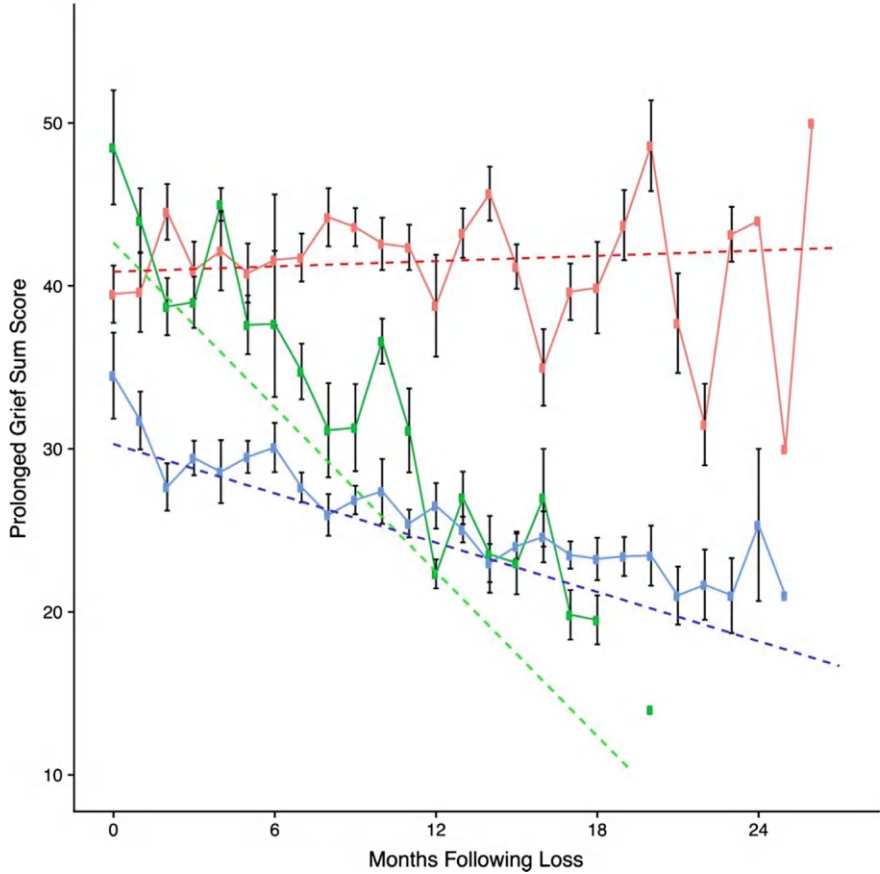


Fig. 1 Line chart shows the Prolonged Grief Sum Score over 24 months following loss. The chart includes three colored lines (green, blue and pink) with error bars, each representing different data sets. The x-axis labeled “Months Following Loss” and the y-axis labeled “Prolonged Grief Sum Score”. Dotted trend lines indicate overall trends for each data set. Overall used to represent grief that is observed and the pattern followed by those suffering. [1]

of grief technology is the extensive collection and analysis of personal data, which is the foundation for crafting a convincing digital representation. AI algorithms meticulously process and analyze these datasets, identifying linguistic patterns, conversational tendencies, and behavioral nuances to generate responses that align with the speech and personality of the deceased. This process is far more intricate than simple chatbot interactions, as it involves reconstructing past dialogue and predicting how the deceased might respond to new topics or evolving circumstances. While this can offer comfort by simulating familiar conversations, it also introduces a profound question: can AI truly replicate a person’s essence, or is it merely an illusion built on statistical probabilities? The complexity of this question lies in the distinction between mechanical mimicry and genuine presence. No matter how advanced AI

becomes, its responses are ultimately derived from pre-existing data rather than lived experiences. As Van Hooijdonk notes, “AI systems are now able to mimic speech patterns, facial expressions, and even the unique mannerisms of the deceased, making the experience of interacting with a digital avatar incredibly lifelike” (Van Hooijdonk). While technologically impressive, this lifelikeness risks blurring the boundary between reality and simulation, potentially fostering psychological dependencies that prevent individuals from moving through the natural grieving process. Beyond textual and conversational capabilities, grief technology incorporates voice synthesis and visual modeling to enhance the realism of digital recreations. Utilizing deep learning techniques, voice cloning software can generate speech that mirrors the deceased’s tone, pitch, and inflection with remarkable precision, often requiring only a few minutes of audio data for training. Similarly, deepfake technology, powered by generative adversarial networks (GANs), allows for creation of dynamic video avatars that simulate facial expressions and gestures in real-time. “Deepfake technology is now being used to create realistic avatars of the deceased, allowing people to see and hear their lost loved ones again” (Van Hooijdonk). While these advancements enable an unprecedented level of immersion, they also raise critical ethical concerns regarding consent. Since deceased individuals cannot provide explicit permission for their data to be used in such a manner, families must grapple with whether resurrecting a digital likeness aligns with their loved one’s wishes. Moreover, the potential for misuse, such as unauthorized recreations or commercial exploitation of digital identities, poses an additional layer of ethical ambiguity.

Furthermore, the psychological implications of grief technology remain deeply complex. While interacting with AI-generated versions of the deceased can provide temporary solace, it may also lead to emotional entanglement that complicates the grieving process. Traditional grief models emphasize the importance of acceptance and emotional detachment as necessary steps toward healing; by contrast, grief technology offers a form of continued engagement that could delay or even disrupt these natural psychological transitions. The ability to “speak” with the deceased at any time through AI-powered simulations risks fostering an artificial sense of continuity that may inhibit individuals from fully processing their loss. In some cases, this could result in heightened dependency, where users become reliant on digital recreations rather than seeking support through human relationships and therapeutic interventions. At the same time, grief technology does offer an alternative coping mechanism, particularly for those who struggle with unresolved emotions or lack access to traditional mental health resources. The question then arises: does this technology serve as a tool for emotional healing, or does it merely prolong attachment to an entity that no longer truly exists? The introduction of AI-driven grief technology ultimately reflects broader societal shifts in how death and mourning are conceptualized in the digital age. It challenges long-held beliefs about the finality of death, raising the possibility that individuals may continue to exist in some form through technological means. This shift is not merely technological but also philosophical, prompting new discussions about what it means to preserve someone’s essence beyond their physical existence. The more AI advances, the more it questions the nature of identity itself, if a person’s speech, thoughts, and behaviors can be replicated convincingly,

at what point does the distinction between the biological and the artificial become ambiguous? While grief technology offers valuable innovations for those seeking comfort, its ethical and psychological implications necessitate careful consideration, ensuring that it remains a tool for support rather than distorting the human experience of loss.

5 Privacy

Another concern related to grief technology pertains to the critical issues of privacy and managing data concerning the deceased. While these innovations can provide solace by allowing individuals to maintain a connection with lost loved ones, they simultaneously introduce substantial risks regarding protecting sensitive personal information. These technologies typically require extensive access to an individual's digital footprint, including texts, emails, photographs, videos, social media posts, and other online interactions, to generate an AI-powered simulation of the deceased. While this capability can be comforting for those in mourning, it raises serious ethical and legal questions about data ownership, consent, and the commercialization of digital legacies [15]. One of the most pressing moral dilemmas is the extent to which companies developing grief technology can access and utilize deeply personal aspects of a person's life, often without explicit consent from the deceased. "An April 2023 study in *Computer Law and Security Review* highlights the legal and ethical concerns of grief tech, including the lack of consent of the deceased individual" [2]. Unlike physical possessions, which are passed down through wills or next-of-kin arrangements, a person's digital footprint exists in a gray area of ownership. Most individuals do not leave clear directives regarding the posthumous use of their digital data, making it difficult to determine who—if anyone—has the right to authorize its use. This ambiguity creates an ethical minefield where technology companies, rather than families or the deceased themselves, dictate how digital legacies are preserved, modified, and even monetized. Beyond consent issues, reducing an individual's life to a dataset presents profound concerns about the erosion of privacy and the commodification of deeply personal memories. Transforming a lifetime of interactions into an AI model risks stripping away the emotional depth and personal agency associated with those memories, reducing them to algorithmic outputs that can be stored, analyzed, and potentially exploited for financial gain. This raises unsettling questions about whether an individual's most intimate moments, private messages, heartfelt emails, or personal photos, should ever be used to create a digital replica, particularly when the deceased has no opportunity to refuse or control how they are represented. In addition to these ethical considerations, security risks associated with grief technology must not be overlooked. As with any technology reliant on large-scale data storage, grief AI platforms are susceptible to data breaches, hacking, and unauthorized access. The potential exposure of such sensitive information could lead to devastating consequences for the bereaved, who may find their loved one's

digital legacy manipulated, misused, or even leaked. The mere possibility that an AI-generated representation of a deceased individual could be altered, shared without consent, or used for unintended purposes adds another layer of vulnerability to an already fragile emotional experience.

6 Conclusion

Like any new technology, grief tech was designed to offer support, providing a means of connection and an avenue to continue a relationship with the deceased. However, unlike other technologies, grief tech holds a unique power, the ability to bridge the gap between life and death by allowing individuals to interact with digital recreations of their loved ones. This technological advancement offers opportunities for solace, helping people find closure or re-establish connections with those who have passed. Yet, for all the potential benefits, it also introduces several significant challenges and risks that must be addressed. One of the major concerns surrounding grief tech is the potential for addiction. The ease with which individuals can engage with digital representations of their deceased loved ones could encourage prolonged reliance on these technologies. This dependency could, in turn, impede the natural grieving process, delaying or even preventing emotional healing. Grief, a deeply personal and inherently transient experience, could become entangled with the artificial nature of digital interactions, potentially trapping individuals in an endless cycle of digital engagement rather than allowing them to process their emotions and move forward. Further complicating the role of grief technology is the growing commodification of mourning. While these technologies can provide meaningful support, they also represent a monetized industry that profits from individuals' most profound emotional vulnerabilities. Grief tech often involves subscription-based services, with pricing models ranging from monthly fees to high-cost one-time payments. This commercialization of grief can be unsettling, as it turns something deeply personal, mourning the loss of a loved one, into a financial transaction. The idea that the grieving process could be monetized raises ethical questions about whether technology companies prioritize their profit over users' emotional well-being. Sometimes, this commodification could overshadow the authentic emotional experiences that grieving individuals seek to process. Equally concerning is the breach of privacy inherent in grief technology. These digital recreations are built from an individual's online presence, including private conversations, photos, and other intimate data. The challenge here lies in obtaining consent from the deceased to use their digital remains in such a way. In many cases, the deceased have not had the opportunity to grant permission for their data to be used to create digital avatars, creating a significant ethical dilemma. Moreover, the sensitive nature of this data raises concerns about the potential for misuse or unauthorized access, especially if it is stored or shared inappropriately. With these risks in mind, companies developing grief tech must adhere to strict data privacy standards and ethical guidelines to ensure the integrity and confidentiality of

individuals' digital legacies. While grief technology can undoubtedly provide valuable support, it is essential to recognize that its role in the mourning process must be carefully managed. Acknowledging the complexities of grief, both psychological and ethical, will be necessary to ensure that these innovations are used responsibly and with respect for the sanctity of personal loss. Technology should serve as a tool for healing, not a substitute for the human connection and emotional processing central to the grieving journey. As we navigate this new frontier in the digital age, ensuring that the human aspects of mourning remain front and center will be crucial. By prioritizing genuine emotional connection and maintaining respect for personal privacy, we can harness the potential of grief tech while safeguarding the authenticity and sacredness of the mourning process. In doing so, we will ensure that technology enhances, rather than diminishes, the deeply personal and transformative experience of grief.

References

1. Djelantik, A.A.A.M.J., Robinaugh, D.J., Boelen, P.A.: The course of symptoms in the first 27 months following bereavement: a latent trajectory analysis of prolonged grief, posttraumatic stress, and depression. *Psychiatr. Res.* **311**, 114472 (2022). <https://doi.org/10.1016/j.psychres.2022.114472>. (<https://www.sciencedirect.com/science/article/pii/S0165178122000865>) (Image 1), ISSN 0165-1781
2. Agarwal, M.: Can AI help with grief? "Grief tech" is silicon valley's solution to mourning. *Vox*, 21 November 2023. <https://www.vox.com/culture/23965584/grief-techghostbots-ai-strutups-replika-ethics>. Accessed 27 Oct 2024
3. American Psychological Association. Grieving changes the brain. *Speaking of Psychology*, 9 February 2022. www.apa.org/news/podcasts/speaking-of-psychology/grieving-changes-brain
4. de Luna, I.B.: Griefbots. a new way of communicating with the dead? ResearchGate. https://www.researchgate.net/publication/359255833_Griefbots_A_New_Way_of_Communicating_With_The_Dead. Accessed 27 Oct 2024
5. Freud, S.: The "Uncanny"1. MIT (1919). <https://web.mit.edu/allanmc/www/freud1.pdf>. Accessed 28 Sept 2024
6. HereAfter AI—Plans & Pricing. HereAfter AI. <https://www.hereafter.ai/pricing>. Accessed 11 Oct 2024
7. van Hooijdonk, R.: Grief tech: redefining death in the age of AI. Richard van Hooijdonk.com. <https://blog.richardvanhooijdonk.com/en/grief-tech-redefining-death-in-the-age-of-ai/>
8. Jee, C.: Technology that lets us "speak" to our dead relatives has arrived. Are we ready? MIT Technol. Rev. (2022). <https://www.technologyreview.com/2022/10/18/1061320/digital-clones-of-dead-people/>. Accessed 7 Oct 2024
9. Phillips, L.: A closer look at the mental health provider shortage. American Counseling Association (2023). <https://www.counseling.org/publications/counseling-today-magazine/article-archive/article/legacy/a-closer-look-at-the-mental-health-provider-shortage>. Accessed 28 Sept 2024
10. Satpathy, S., Mangla, M., Sharma, N., et al.: Predicting mortality rate and associated risks in COVID-19 patients. *Spat. Inf. Res.* **29**, 455–464 (2021). <https://doi.org/10.1007/s41324-021-00379-5>
11. Satpathy, S., Nandan Mohanty, S., Chatterjee, J.M., Swain, A.: Comprehensive claims of AI for healthcare applications-coherence towards COVID-19. In: Nandan Mohanty, S., Saxena, S.K., Satpathy, S., Chatterjee, J.M. (eds.) *Applications of Artificial Intelligence in COVID-19*.

- Medical Virology: From Pathogenesis to Disease Control. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-7317-0_1
12. Mohapatra, S., Satpathy, S., Mohanty, S.N.: A comparative knowledge base development for cancerous cell detection based on deep learning and fuzzy computer vision approach. *Multimed. Tools Appl.* **81**, 24799–24814 (2022). <https://doi.org/10.1007/s11042-022-12824-0>
 13. Baral, S., Satpathy, S., Pati, D.P., Mishra, P., Pattnaik, L.: A literature review for detection and projection of cardiovascular disease using machine learning. *EAI Endorsed Trans. Internet Things* **10**, 1–7 (2024)
 14. Mohapatra, S., Satpathy, S., Paul, D.: Data-driven symptom analysis and location prediction model for clinical health data processing and knowledge base development for COVID-19. In: Nandan Mohanty, S., Saxena, S.K., Satpathy, S., Chatterjee, J.M. (eds.) *Applications of Artificial Intelligence in COVID-19*. Medical Virology: From Pathogenesis to Disease Control. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-7317-0_6
 15. Satpathy, S., Swain, P.K., Mohanty, S.N., Basa, S.S.: Enhancing security: federated learning against man-in-the-middle threats with gradient boosting machines and LSTM. In: 2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Niagara Falls, ON, Canada, pp. 1–8 (2024). <https://doi.org/10.1109/AVSS61716.2024.10672589>

A Comparison of Interdependent Deep Learning Models and Exponential Smoothing Method for Predicting Bitcoin Price



Nrusingha Tripathy, Sarbeswara Hota, Debahuti Mishra, Meera Dash, Soumyarashmi Panigrahi, and Subrat Kumar Nayak

Abstract A virtual or digital currency, cryptocurrency, is based on blockchain technology and uses encryption to ensure security. Currently, over 2000 different coin types are available on the cryptocurrency market, with a massively unequal number of transactions and circulation. Anticipating the inclination of cryptocurrency prices to fluctuate is essential since the investment risk attached to them is greater than that of conventional goods. Four different models were utilized in this work to address the limitations of traditional production forecasting: Long Short-Term Memory (LSTM), Facebook Prophet (FB-Prophet), Silverkite, and Bidirectional LSTM. The FB Prophet and Silverkite both support the exponential smoothing method, where the LSTM and Bi-LSTM are the deep learning models. Silverkite is the main algorithm used in the Python library Graykite by LinkedIn. We examined the models using historical Bitcoin data over the previous nine years, from January 2012 to March 2021. The Bi-LSTM model provides a root mean squared error (RMSE) score of 3.415 and a mean absolute error (MAE) score of 7.539. The Bi-LSTM model detects the variations that might attract attention and prevent potential issues.

Keywords Financial data analysis · Long short-term memory · FB-prophet · Silverkite · Bi-LSTM

N. Tripathy (✉) · D. Mishra · S. Panigrahi · S. K. Nayak
Department of CSE, ITER Siksha ‘O’ Anusandhan (Deemed to Be University), Bhubaneswar, Odisha, India
e-mail: nrusinghatripathy654@gmail.com

D. Mishra
e-mail: debahutimishra@soa.ac.in

S. Hota
Department of CA, ITER, Siksha ‘O’ Anusandhan (Deemed to Be University), Bhubaneswar, Odisha, India
e-mail: sarbeswarahota@soa.ac.in

M. Dash
Gandhi Institute for Education and Technology (GIET), Baniatangi, Odisha, India
e-mail: mdash@giyet.edu.in

1 Introduction

Rapid information flow, many players with a wide range of investment horizons, and numerous feedback mechanisms are characteristics of modern financial markets. These factors cause complex phenomena, such as speculative bubbles or collapses, together. As a result, they are regarded as one of the most intricate systems in existence. The recent remarkable growth of the cryptocurrency market, which went from being utterly peripheral to capitalization at roughly the size of an intermediate-sized stock exchange, offers a rare chance to see its development quickly [1]. High-frequency data is readily available, enabling sophisticated statistical analysis in developments on cryptocurrency exchanges from the beginning to the present. This creates an opportunity to measure the evolutionary shifts in the complexity features that go hand in hand with the birth and maturation of markets [2].

Bitcoin is a digital currency that is meant to be exchanged. Users may interact with each other autonomously and transparently by exchanging native tokens, which are commonly referred to as “Bitcoins,” and collaboratively verifying the transactions. The foundational technology consists of a shared global ledger, or blockchain, among participants and a Bitcoin reward system that encourages users to manage the transaction network. About 1500 more cryptocurrencies have been released since the launch of Bitcoin in 2009; about 600 of them are now being traded frequently. While most cryptocurrencies operate on separate transaction networks, they all share the same incentive system and blockchain technology. Many are essentially Bitcoin clones but with disparities in supply, transaction validation times, and other aspects [3]. Others have resulted from more critical advancements in the blockchain’s underlying technology.

Considering how sensitive financial markets are and how common cryptocurrency is in trade, predicting Bitcoin values presents several security issues and difficulties. The well-known Bitcoin keeps track of every transaction in a distributed, append-only public database called the “blockchain.” The incentive-compatible consensus-based distributed process, which is operated by users known as “miners,” is a significant component of Bitcoin’s security. The miners are expected to utilize the blockchain honestly in return for the reward. The Bitcoin economy, which was first introduced in 2009, has expanded rapidly and is currently valued at over 40 billion dollars. Adversaries are encouraged to forecast future trends and take advantage of vulnerabilities for financial gain by the exponential rise in the market value of Bitcoin. Miners of Bitcoin are essential to preserving the network’s integrity and security. Their efforts guarantee that the blockchain, which supports Bitcoin’s decentralized and untrustworthy character, operates as intended. Miners use their expertise in solving intricate cryptographic riddles to verify and log transactions in the blockchain. By doing this, fraud and double-spending bouts are evaded, and only legal transactions are added to the blockchain. Because of its devolution, the Bitcoin network is immune to control and single points of fiasco. It is very tough to change previous transactions since

miners legalize transactions through several blocks. This method daunts spam transactions by demanding attackers to pay high costs to excess the network. Block space constraints guarantee that only valid and urgent transactions are fingered.

We liken the FB-Prophet, Silverkite, LSTM, and Bi-LSTM models to this work. Bi-LSTM models' bidirectional nature allows them to consider data from previous and subsequent time steps when producing predictions. This may recover the model's understanding of the context around vicissitudes in the price of Bitcoin. The Bi-LSTM model classifies the aberrations that might draw courtesy and deter problems.

2 Research Methodology

Predicting Bitcoin prices, frequently skilled by smearing various analyses and predictive models, is helpful to several participants in the cryptocurrency ecosystem [4]. Users can decrease possible losses by practicing risk management techniques and being aware of potential market growth. This might involve expanding portfolios, putting stop-loss orders in place, or adjusting investing plans in light of estimates. By predicting price blows, precise projections can aid in identifying possible profit opportunities. To enhance profits, traders and investors can take benefit of advantageous market situations and act promptly.

2.1 *Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Statistical Analysis*

The stationarity of the time series data set is assessed using a statistical test identified as the KPSS test. A time series' statistical assets, such as its alteration and mean, are expected to endure constant throughout time according to stationarity, which is fundamental to time series scrutiny. Scrutiny and modeling of unequal time series data can be problematic due to trends or other designs. To achieve the KPSS test, first evaluate a model comprehending a trend component and then check the stationarity of the residues [5]. When using the KPSS test, the stationary nature of the time series information along a deterministic trend is the null hypothesis. The null hypothesis is debunked, and the data is suggested to be non-stationary if the test statistic is higher than a crucial value. However, if the test parameter is smaller than the critical value, the null hypothesis remains valid, implying that the data is static. Analysts can select appropriate frameworks for predicting or other analytical uses based on the time series stationarity properties. For example, differencing may bring the non-stationary data into stationarity before using specific models. Figure 1 shows the consequence of the KPSS statistical analysis.

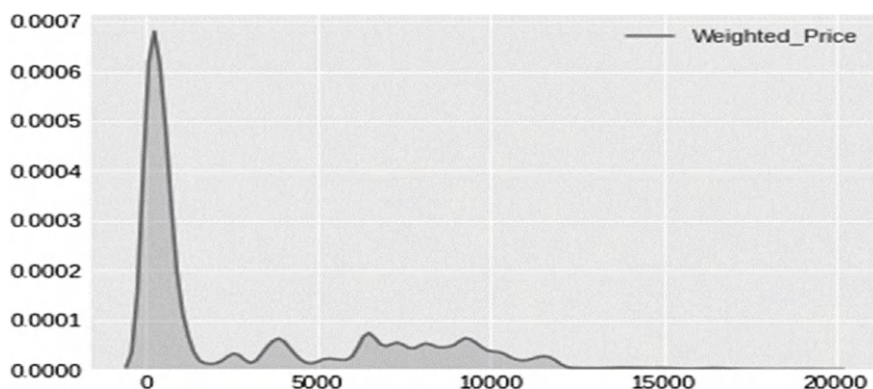


Fig. 1 KPSS statistical analysis plot

2.2 Time Series Decomposition and Statistical Test

A time series data collection may be divided into parts using a time series decomposition technique. These components usually include patterns, seasons, and remnant components. Making sense of the fundamental trends and oscillations over time series, forecasting, and decision-making are all aided by this research. Recognizing the general trajectory of the data over a long period requires understanding the trend. Seasonal patterns frequently happen on daily, monthly, or annual schedules. Visual examination or statistical analysis can be used to find these recurring patterns. Identifying the regular cycles in the data and recognizing seasonality can help you make better predictions and decisions [6]. The residual component represents the time series data's unpredictable and irregular fluctuations that are not explained by seasonality or trends. Figure 2 shows the time series decomposition plot. Reducing the residual component from the trending and seasonal components of the time series. This makes it easier to investigate the erratic variations in the data in more detail.

3 Proposed Model

Cryptocurrency models for forecasting can assist in evaluating and managing risk. Comprehending probable price swings empowers interested parties to execute risk-reduction plans and establish suitable stop-loss thresholds [7]. Algorithmic trading systems frequently incorporate predictive models, enabling automatic trade execution based on predetermined criteria. Algorithmic trading gives an advantage over competitors by reacting swiftly to changes in the market. Precise forecasts can help arrange the timing of market operations, such as purchasing or disposing of assets, according to anticipated price changes. This can lessen the effects of adverse market conditions and increase profitability. Figure 3 shows the entire workflow diagram of

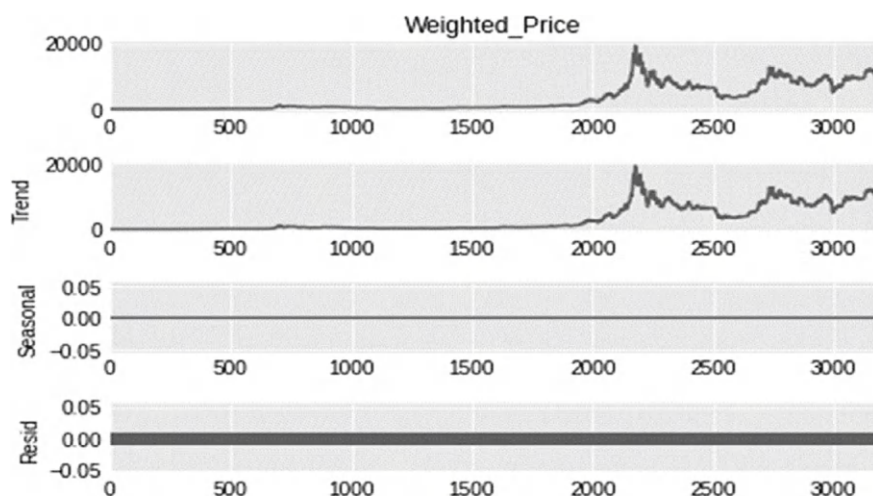


Fig. 2 Time series decomposition conspiracy

this work. Several things, including mistakes in data gathering or network problems, can cause missing values in financial datasets. Pre-processing ensures that the simulation has all the data it needs to function by assisting with imputing these missing variables. Features with different sizes, such as market capitalization, price, and trading volume, may be included in Bitcoin price data. Normalization and standardization are examples of feature scaling that guarantee every feature contributes pretty to the model and keeps any one parameter from dominating because of its size. Holidays, trends, and seasonal patterns are automatically identified and modeled using the Silverkite model. This is especially helpful for Bitcoin, as there may be regular trends at specific periods of the day, week, or year.

3.1 Long Short-Term Memory (LSTM)

The value of each data point in a sequence of cryptocurrency prices depends on its chronological context. Because LSTMs are built to capture patterns and long-term relationships in sequential data, they are an excellent choice for modeling the time-series nature of Bitcoin pricing [8]. Long-term dependencies may be problematic for conventional time series models to depict, mainly when there are significant delays between pertinent events and how they affect pricing. Long-term dependencies are well captured and learned by LSTMs, enabling a more precise representation of complicated interactions in the data [9]. The marketplaces for cryptocurrencies frequently show dynamic, nonlinear patterns that may be difficult for linear models to represent adequately. Because LSTMs can learn and express complicated nonlinear

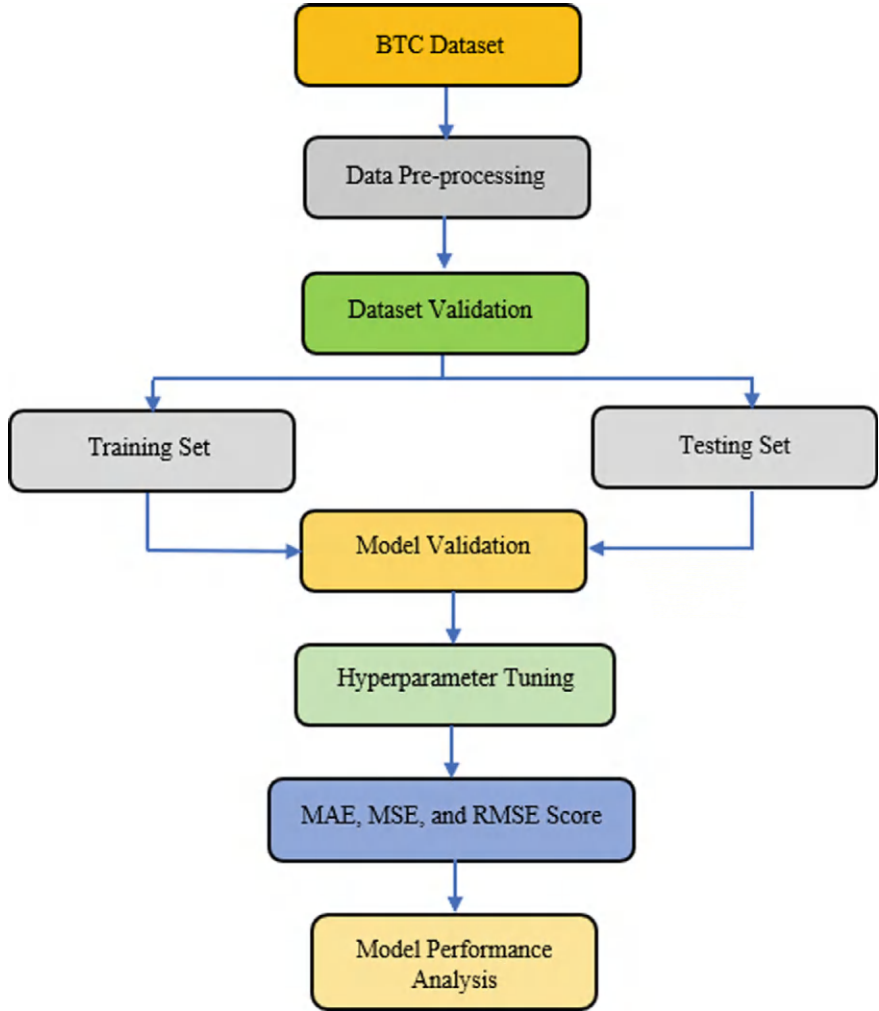


Fig. 3 Workflow diagram of this work

connections in data, they are a good fit for capturing cryptocurrency price fluctuations' dynamic and varied character. LSTMs reduce the vanishing gradient issue that is prevalent in conventional RNNs. This makes it possible to train the model more successfully on lengthy data sequences, guaranteeing that it can absorb knowledge from and adjust to the complete historical context of Bitcoin values [10]. The cryptocurrency market is prone to abrupt fluctuations and may be pretty active. Because LSTMs are flexible and may change with the market, they are helpful for price prediction when patterns and trends are subject to sudden fluctuations [11].

3.2 Facebook Prophet (FB-Prophet)

Facebook Prophet is a user-friendly, frank time series foretelling tool. Prophet is easy to use and needs slight modification of its hyperparameters. Users with diverse knowledge may apply time series forecasts for Bitcoin values due to their ease. The Prophet can routinely classify and comprise daily, weekly, and annual seasonality designs in the data [12]. This is valuable for identifying recurrent trends and designs in Bitcoin pricing, counting weekly market cycles, or daily trading designs. Prophet is tough to time series abnormalities and missing data. Because of its ability to achieve data abnormalities, it is suitable for Bitcoin price data that could contain gaps or strange fluctuations. Users may add the influence of exact occurrences or holidays on Bitcoin values by modeling holiday influences using Prophet [13]. Because of its flexibility, the model may be tailored to detention particular features in the data. Prophet crops deconstructed mechanisms that are simple to comprehend, such as trend and seasonality. This openness whitethorn is helpful for predictors and other stakeholders looking to learn the variables affecting cryptocurrency price estimates. In Eq. 1,

$$S_y(t) = \sum_{n=1}^N a_n \cdot \cos\left(\frac{2\pi nt}{P}\right) + b_n \cdot \sin\left(\frac{2\pi nt}{P}\right) \quad (1)$$

$$S_w(t) = \sum_{n=1}^N c_n \cdot \cos\left(\frac{2\pi nt}{T}\right) + d_n \cdot \sin\left(\frac{2\pi nt}{T}\right) \quad (2)$$

3.3 Silverkite

Prices for cryptocurrencies are one type of time series data that regularly shows seasonality tendencies. By considering recurrent patterns in the data, methods like Silverkite can dependably achieve and include seasonality fluctuations to help yield more accurate predictions. Automatic feature assortment may be in specific foretelling programs like Silverkite [14]. This might benefit an upsurge in the accuracy of Bitcoin price predictions by automatically locating and leveraging pertinent features from the input data. Often, accurate Bitcoin prediction involves considering outside variables, including macroeconomic statistics, regulatory developments, and market emotion. The model's capacity to represent the effect of these variables on cryptocur- rency prices may be better if Silverkite permits the incorporation of external variables [15]. Forecasting outcomes that are comprehensible and easy to understand are essential for Bitcoin decision-makers. Users can make better judgments if Silverkite offers detailed evidence about the forecasts made by the model and the variables affecting them. The marketplaces for cryptocurrencies are volatile and subject to unexpected

fluctuations. For traders and investors who want to respond quickly to changes in the market, tools that provide real-time forecasting or fast updates might be helpful [16].

3.4 *Bidirectional LSTM (Bi-LSTM)*

Future and previous market circumstances have an impact on cryptocurrency values. Because of their bidirectional architecture, which captures both forward and backward temporal dependencies, bi-LSTM models can anticipate future price movements by considering a wider range of factors [17]. The trends in cryptocurrency markets are frequently complicated and nonlinear. Bi-LSTM models are well-suited to simulating the dynamic and ever-changing character of Bitcoin price fluctuations since they are excellent at capturing complex nonlinear correlations in the data. Bi-LSTM models do not require human feature engineering since they automatically extract specific characteristics from the input data [18]. This is helpful when working with complicated and high-dimensional datasets, including those that contain several time series or different factors that affect the price of cryptocurrencies. Time series modeling is made more thorough by Bi-LSTM models, which consider past and future information. In the context of Bitcoin markets, this comprehensive viewpoint on the temporal features of the data helps to provide more precise and nuanced predictions [19]. Equation 3 reflects the fascinating formulation of time series.

$$y_t = a_0 + \sum_{i=1}^p a_i y_{t-i} + \sum_{j=0}^q \beta_j \varepsilon_{t-j} \quad (3)$$

where y_t It is the most recent reflection of fascinating time series y_{t-1} ($i = 1, 2, \dots, p$) is the past observations and ε_{t-j} ($j = 0, 1, 2, \dots, q$) is random mistakes with a finite variance and a zero mean. The Bayesian Information Criterion (BIC) rule chooses the order signified by p and q , respectively [20].

4 Result Analysis

By modifying asset allocations in response to anticipated market fluctuations, investors may utilize Bitcoin forecasts to enhance their portfolios. This contributes to the development of a varied and well-balanced investment portfolio. By focusing investments on assets with attractive predicted returns, cryptocurrency forecasts assist users in making more effective capital allocation decisions. This is especially crucial in the vibrant and often changing Bitcoin industry. Forecasting algorithms for cryptocurrencies are being developed and improved, which is helping to progress machine learning, data science, and algorithmic trading. This ongoing innovation

is advantageous to the larger technical and financial environment. Figure 4 shows the LSTM-predicted BTC price, and Fig. 5 shows the FB-prophet prophesied BTC price.

Figure 6 shows the Silverkite predicted BTC price, and Fig. 7 displays the Bi-LSTM predicted BTC price. The deep learning models outperform the exponential smoothing approach in terms of performance. The Bi-LSTM model yields an MAE score of 7.539 and an RMSE score of 3.415. Different amounts of historical data may be accessible for study in cryptocurrency data. Variable-length sequences may be handled by Bi-LSTM models, giving them flexibility in handling various time intervals and tolerating data that is sporadically sampled. Interpretability is one of Silverkite's primary characteristics. It offers insights into how the forecasts affect factors like seasonality, external variables, or recent price changes. This openness helps in comprehending the forces that influence Bitcoin values. The model may be

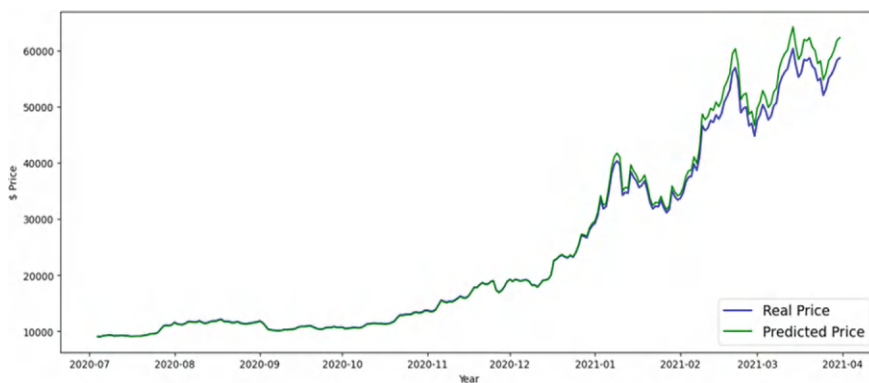


Fig. 4 LSTM predicted BTC price

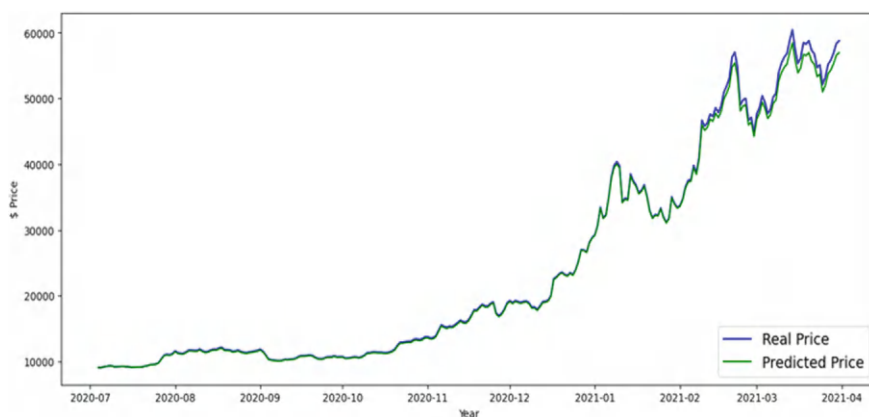


Fig. 5 FB-prophet predicted BTC price

easily tuned and adjusted to meet the Bitcoin market’s unique features by changing any of its components. Prices for Bitcoin frequently show irregular seasonality and nonlinear tendencies. The Silverkite model can effectively capture the intricacies of trend and seasonality due to its capacity to integrate intricate and adaptable components.

In contrast to conventional LSTM models, which take historical data into justification, Bi-LSTM models analyze input data forward and backward. This increases the forecast potential for Bitcoin prices, which can be wedged by patterns that emerge before and after a precise time point, by allowing the model to include correlations not just from past time steps but also from following steps. Because of many market circumstances, the correlations among Bitcoin prices are typically complex and non-linear. Equated to simpler models, the Bi-LSTM model produces more precise forecasts because of its deep learning design, which is excellent at taking

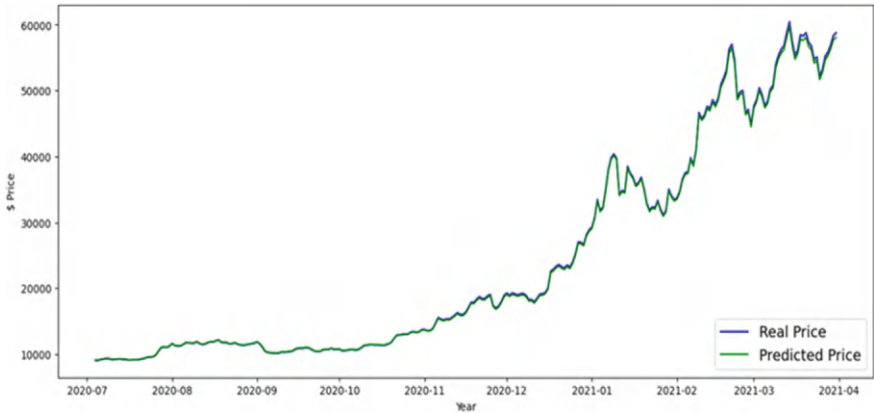


Fig. 6 Silverkite predicted BTC price

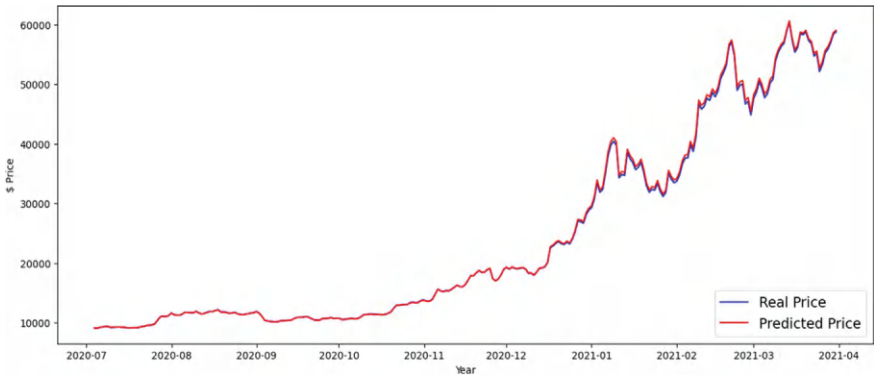


Fig. 7 BI-LSTM predicted BTC price

these complex patterns. Overall, the Bi-LSTM model is a good option for envisaging the price of Bitcoin because of its suppleness, pliability, and capacity to detention intricate chronological patterns and dependencies. This is especially true in a market where non-linear interactions and high volatility are prevalent.

Table 1 displays each model’s MAE, MSE, and RMSE scores. Figure 8 Customs a histogram plot to show each model’s recital metric. Bi-LSTMs were determined to extract long-term additions from sequential data. This is particularly important for time series forecasting, as past data is frequently cast off to forecast future values. The model can consider data from both the past and the future because of its bidirectional nature.

Accurate forecasts enable investors to reduce risk by anticipating possible downturns or volatility and making appropriate portfolio modifications. Long-term investing plans may benefit significantly from using predictive models to identify emerging trends and patterns in the market. Predictions are necessary for automated trading algorithms to be developed to handle trades based on predetermined rules and market circumstances. Thanks to predictive algorithms, high-frequency traders can profit from tiny price swings that happen over brief periods in fast-paced situations. Sentiment research and price prediction combined can shed light on the psychology of the market and how news and events may affect pricing. Accurate price projections can enhance market efficiency by reducing the difference between the market price and the inherent value of the asset. Price forecasting techniques can advance the area of financial economics by assisting scholars in comprehending the fundamental elements influencing the price of Bitcoin.

Table 1 MAE, MSE, RMSE score of models

Model name	MAE	MAP	RMSE
LSTM	15.632	27.423	5.742
FB-Prophet	9.492	16.375	4.013
Silverkite	8.323	14.371	3.971
Bi-LSTM	7.539	12.513	3.415

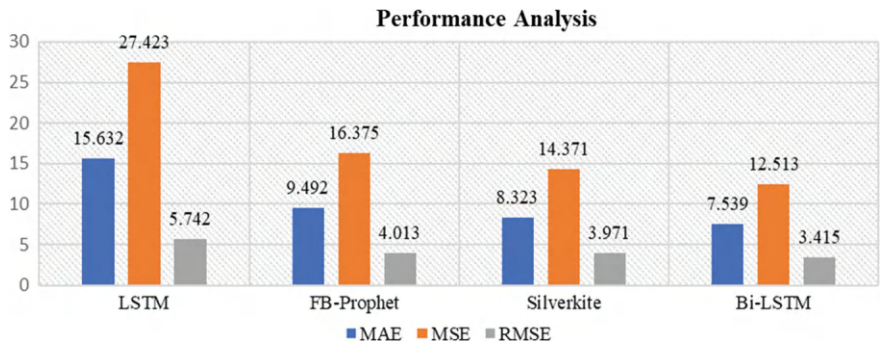


Fig. 8 Model performance analysis through histogram plot

Miners' pooled computing power is a defense against possible attacks from hostile organizations, such as big businesses or governments. An attacker can rarely outrun honest miners with a high hash rate. The decentralized and untrustworthy nature of the network is strengthened by its resiliency. The core of Bitcoin's security system is its miners. They give the Bitcoin network stability, decentralization, and integrity through proof-of-work. Bitcoin is one of the safest and most reliable blockchain systems because its incentives guarantee ongoing involvement and consistency with the network's objectives.

5 Conclusion

Cryptocurrency prediction makes it possible to recognize and evaluate the risks related to price volatility. This knowledge is essential for implementing efficient risk management techniques to reduce potential losses. Investors may utilize projections to optimize their Bitcoin portfolios by modifying asset allocations based on anticipated market changes. The Bi-LSTM model yields an MAE score of 7.539 and an RMSE score of 3.415. By identifying deviations that might draw attention, the Bi-LSTM model helps to avoid possible problems. This contributes to the creation of balanced and diverse portfolios. Algorithmic trading systems frequently incorporate predictive models, enabling automatic trade execution according to present standards. Algorithmic trading stretches benefits over competitors by responding swiftly to variations in the market. Sentiment research approaches may be included in predictive models to measure market sentiment gleaned from media coverage, social networking sites, and other sources. Comprehending the feelings of the market adds more context to forecasting. Prolonged progressions and novelties in deep learning layouts, counting more intricate neural network configurations or hybrid approaches, can raise the precision and effectiveness of Bitcoin price forecasts. Researchers may examine novel strategies or modifications to existing models to seize complex patterns and connections.

References

1. Lamon, C., Nielsen, E., Redondo, E.: Cryptocurrency price prediction using news and social media sentiment. *SMU Data Sci. Rev.* **1**(3), 1–22 (2017)
2. Tripathy, N., Hota, S., Mishra, D., Satapathy, P., Nayak, S.K.: Empirical forecasting analysis of Bitcoin prices: a comparison of machine learning, deep learning, and ensemble learning models. *Int. J. Electr. Comput. Eng. Syst.* **15**(1), 21–29 (2024)
3. Chen, W., Zheng, Z., Ma, M., Wu, J., Zhou, Y., Yao, J.: Dependence structure between bitcoin price and its influence factors. *Int. J. Comput. Sci. Eng.* **21**(3), 334–345 (2020)
4. Fosso Wamba, S., Kala Kamdjoug, J.R., Epie Bawack, R., Keogh, J.G.: Bitcoin, blockchain and fintech: a systematic review and case studies in the supply chain. *Prod. Plan. & Control* **31**(2–3), 115–142 (2020)

5. Tripathy, N., Hota, S., Prusty, S., Nayak, S.K.: Performance analysis of deep learning techniques for time series forecasting. In: 2023 International Conference on Advances in Power, Signal, and Information Technology (APSIT), pp. 639–644. IEEE, New York (2023)
6. Manujakshi, B.C., Kabadi, M.G., Naik, N.: A hybrid stock price prediction model based on pre and deep neural networks. *Data* **7**(5), 51 (2022)
7. Hamayel, M.J., Owda, A.Y.: A novel cryptocurrency price prediction model using GRU, LSTM, and bi-LSTM machine learning algorithms. *AI* **2**(4), 477–496 (2021)
8. Shintate, T., Pichl, L.: Trend prediction classification for high-frequency Bitcoin time series with deep learning. *J. Risk Financ. Manag.* **12**(1), 17 (2019)
9. Kim, G., Shin, D.H., Choi, J.G., Lim, S.: A deep learning-based cryptocurrency price prediction model that uses on-chain data. *IEEE Access* **10**, 56232–56248 (2022)
10. Heo, J.S., Kwon, D.H., Kim, J.B., Han, Y.H., An, C.H.: Prediction of cryptocurrency price trend using gradient boosting. *KIPS Trans. Softw. Data Eng.* **7**(10), 387–396 (2018)
11. Tripathy, N., Hota, S., Mishra, D.: Performance analysis of bitcoin forecasting using deep learning techniques. *Indones. J. Electr. Eng. Comput. Sci.* **31**(3), 1515–1522 (2023)
12. Fauzi, M.A., Paiman, N., Othman, Z.: Bitcoin and cryptocurrency: challenges, opportunities, and future works. *J. Asian Financ. Econ. Bus.* **7**(8), 695–704 (2020)
13. Tripathy, N., Balabantaray, S.K., Parida, S., Nayak, S.K.: Cryptocurrency fraud detection through classification techniques. *Int. J. Electr. Comput. Eng.* **14**(3), 2918–2926 (2024)
14. Lahmiri, S., Bekiros, S.: Deep learning forecasting in cryptocurrency high-frequency trading. *Cogn. Comput.* **13**, 485–487 (2021)
15. Tripathy, N., Nayak, S.K., Prusty, S.: A comparative analysis of silverkite and inter-dependent deep learning models for bitcoin price prediction. *Front. Blockchain* **7**, 1346410 (2024)
16. Mensi, W., Gubareva, M., Ko, H.U., Vo, X.V., Kang, S.H.: Tail spillover effects between cryptocurrencies and uncertainty in the gold, oil, and stock markets. *Financ. Innov.* **9**(1), 92 (2023)
17. Trabelsi, N.: Are there any volatility spill-over effects among cryptocurrencies and widely traded asset classes? *J. Risk Financ. Manag.* **11**(4), 66 (2018)
18. Zhao, D., Rinaldo, A., Brookins, C.: Cryptocurrency price prediction and trading strategies using support vector machines (2019). [arXiv:1911.11819](https://arxiv.org/abs/1911.11819)
19. Awoke, T., Rout, M., Mohanty, L., Satapathy, S.C.: Bitcoin price prediction and analysis using deep learning models. In: *Communication Software and Networks: Proceedings of INDIA 2019*, pp. 631–640. Springer, Singapore (2020)
20. Ranjan, S., Kayal, P., Saraf, M.: Bitcoin price prediction: a machine learning sample dimension approach. *Comput. Econ.* **61**(4), 1617–1636 (2023)

Recognition and Classification of ARP Spoofing and DDoS Attack Using Machine Learning Approach



Saswati Chatterjee, Suneeta Satpathy, and Deepthi Godavarthi

Abstract Despite the perceived security of wireless networks to be secure from several threats, they are vulnerable to more serious threats such as DDoS attacks. This research explores the feasibility of using artificially generated network datasets to train predictors to predict continuous threats targeting wireless networks. Over and over again, hackers take advantage of these networks' complex and constantly evolving nature, with varied weaknesses originating from various causes. Several sectors are critical infrastructures, often the first to be attacked because of their significance and the devastating impacts they bring when shaken. Among the various network security threats, two have been selected in this paper: DDoS attacks and ARP Spoofing because of their high prevalence and impact. DDoS attacks suddenly flood the network with constant traffic; thus, the network and computer resources cannot respond to genuine users. ARP Spoofing enables the attacker to intercept or alter messages between two devices in the same broadcast domain. These attacks cause interference with the actual functioning of the network and make other confidential information in the network open for further attacks. Thus, to mitigate these threats, the current study utilizes a Random Forest algorithm combined with the Principal Component Analysis (RFPCA). This way, only the most relevant features are included, enhancing efficiency and increasing the detection Model's performance. Due to its ability to handle complex data structures in conjunction with PCA for dimensionality reduction, Random Forest is a good framework for identifying such patterns of attacks. Moreover, an entropy-based classification method is also incorporated to fine-tune the detection precision to identify the inconsistency in the network traffic patterns. Therefore, this study reveals an extensive process for using the said approaches in threat identification and prevention in wireless networks. The

S. Chatterjee
FET, Sri Sri University, Cuttack, Odisha, India

S. Satpathy
SoA Deemed to Be University, Bhubaneswar, Odisha, India

D. Godavarthi (✉)
School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India
e-mail: saideepthi531@gmail.com

results are valuable in illustrating how machine learning algorithms may be applied to enhancing the robustness of networks against such threats as DDoS and ARP Spoofing by offering insights for further investigation and a catalog for real-world applications of defending wireless communication systems.

Keywords Security threat · DDoS · Random forest · ARP spoofing

1 Introduction

The Internet of Things (IoT) establishes conventional communication methods between physical objects while forwarding commands to link various systems on the network. Smart connections made possible by IoT technology surpass the capabilities of traditional sensor networks [1]. The processing and sharing of CPS data occur through cloud computing networks [2]. A cyber-physical manufacturing structure serves as a manufacturing structure that enables decentralized decision control for virtual worldwide production networks [3]. The Smart Factory utilizes data within learning algorithms, which generates innovative and long-distance operable technologies [4, 5]. High networked and dynamically distributed environments increase the total system complexity while providing potential entry points for attackers [6]. The rise of the Internet of Things in critical system operations requires enhanced Smart Factory protection measures since dangerous operations directed toward the system could lead to catastrophic outcomes. Internet service providers' control over Cloud security customization reaches only up to their infrastructure management of information facilities [7]. The "digital production" system operates with standardized Cloud configurations throughout machine networks; therefore, a security breach at one facility could spread to all other sites [8]. A manufacturer faces challenges in IoT construction because existing policies and best practices remain unclear, which includes strong password protection that can be attacked by brute force methods [9]. The limited power capacity, constrained processors, and minimal memory storage options within sensor nodes prevent them from implementing advanced protection systems. Better awareness about safety guarantees the prevention of troubles caused by misconfiguration within the IoT ecosystem [10]. A DDoS attack is a cyber-attack that overloads a specific internet service through traffic from multiple sources until the service becomes completely disabled. Attackers target a variety of assets, including government websites and banks. They create "botnets," which are networks of computers that have been infected by attackers through the distribution of malicious software using various methods. 2008, the first software-defined networking (SDN) system was implemented due to successful research [11]. The implementation of SDN technology made software-defined networking the essential structure for contemporary network infrastructure. Traditional networks, together with SDNs, remain at risk of experiencing DDoS attacks. The present network architecture becomes vulnerable to Distributed Denial of Service (DDoS) attacks because of their

current weaknesses. The rise of weak security-linked Internet sites with high bandwidth connections will most likely increase the severity of DDoS attacks. The fast growth of computer network utilization and applications makes secure environments necessary [12, 13]. The benefits of technological advancements to simplify everyday procedures create parallel opportunities for attackers to exploit these systems, which benefits users. User organizations face multiple options through which attackers can generate harm. This attack infrastructure presents different threats that may or may not trigger authorities to take action [14]. The necessity of deploying intrusion detection techniques becomes evident because of this situation. Network security and defense uses machine learning methods to detect and analyze attacks, including Smurf, UDP, and HTTP flooding [15].

The research uses machine learning to develop a mechanism for detecting DDoS attacks and ARP Spoofing attacks aiming to boost enterprise network security measures. The ML techniques use network flow data to extract detailed information, which helps them identify delicate patterns and understand DDoS attacks. All research experiments conclude that classification approaches yield better results than learning techniques. The following paper structure consists of Relational Work in Sect. 2, followed by Proposed Methodology in Sects. 3 and 4 describes the Distributed Denial-of-Service (DDoS) attack with simulation parameters and Entropy method with different experimental results. In Sect. 5, the findings are wrapped up with a discussion of further work.

2 Related Work

Attacks on distributed denial of service (DDoS) aim at server malfunction by abusing commonly exploited vulnerabilities in IoT devices [16]. Too much incoming data from the assault leads to target resource exhaustion, decreased performance, and depletion of network bandwidth [17]. The attacker executes the attack after obtaining confidentiality data and damages fundamental infrastructure [18]. These destructive attacks demand numerous machines already contaminated by computer viruses or malware to create proper botnets. These devices operate under remote controller commands to execute a synchronized offensive operation simultaneously [19] yet remain challenging to defend. Sensor networks and Internet of Things systems operate with minimal supervision; thus, their previous protection systems have disappeared [5]. IoT networks such as Mirai have become the main targets of malware [6]. After deleting the binary file, all infected devices stay undetected, according to research [7]. Continuously infecting vulnerable devices becomes straightforward when a security breach lasts across multiple system restarts. The limited amount of data collected from infected devices becomes significantly larger when it reaches the gateway connection. The paper in [8] evaluates damage elimination effectiveness by implementing learning-based algorithms. These classifiers demand a model-training process with substantial network traffic data from the IoT environment to develop their performance. IoT networks support sending behaviors among their

sensor nodes, which operate as heterogeneous elements [9]. SVM, KNN, and RF are some of the classifiers typically used for DDos [20–22]. The classifiers utilize datasets to develop classifications and make predictive outcomes. ARP enables data connection layer communication while initiating local networks via the universally adopted Address Resolution Protocol. The router must consult the MAC mappings in the ARP table to initiate communication. The protocol encounters significant challenges because it has no dependable mechanism to transmit the ARP packet. The attack remains hidden within the excessive ARP network traffic, which leads to detection difficulties. The security risk arises when attackers use an adopted assault as a basis to execute man-in-the-middle (MITM) or denial-of-service (DoS) attacks. The defense mechanism against ARP spoofing through static ARP updates will generate extra work for users and consume excess processing power. Packet filtering is an ARP protection method that examines ARP packets and eliminates suspected malicious ones. The connection of IoT data to other IT assets produces service and application diversity, which increases maintenance difficulties for existing physical components [23]. The attacker using available tools can guess the packet header identifier even when using a TCP connection, as it represents an essential fragment aspect [24]. Basic prevention of splitting operations might result in more extended time requirements for communications [25]. The study confirms that the TCP protocol lacks effectiveness against enemy attacks based on ICMP spoofing, potentially triggering TCP segment fragmentation against the source device [26]. In this research paper, another potential attack, viz Internet Protocol (IP) fragmentation attack, is considered, along with the protection assessment for better accuracy. According to the model's performance, adding more characteristics to the training process made the model more accurate but may also increase computation time. An attack must be detected quickly to be stopped, and good classification models help safeguard systems and lower maintenance costs. High F1 scores indicate a good performance level regarding correct DDos attack classification. This research presents a framework and experimental findings that resolve existing limitations in former studies.

3 Proposed Methodology

3.1 *Random Forest*

Random forest [27] represents a machine-learning technique applying numerous decision trees to achieve results. Each tree belongs to the forest term without connecting to other trees. The decision tree nodes received their partitioning criteria from random selection during unpruned operation. Random sampling with replacement functions within the bootstrap method when available training data becomes insufficient to develop the model. The average forest output is the basis for the prediction shown in Fig. 1. The maximum voting approach delivers a strong potential to obtain correct solutions during data analysis.

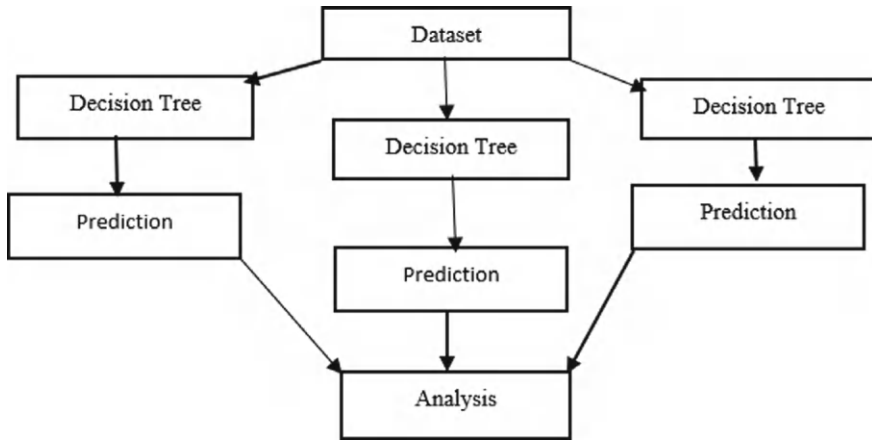


Fig. 1 Proposed architecture

3.2 Principal Component Analysis

The principal component analysis (PCA) reduces dimensionality in the system [28]. The original data follow a transformation process through independent principal components before the arrangement according to the descending order of variance between components. After applying PCA, PCs convert the dataset to principal components corresponding to the maximum variable variances. Among all selected PCs, the first has the most significant impact on data variance.

The eigenvector U of matrix A represent the scalar eigenvalue for matrix A .

$$AU = U\Lambda$$

3.3 Entropy-Based Detection

The measurement method that determines data variability and dispersion is Shannon's entropy, explained through Eq. 1 [29, 30]. PCA provides a successful means to detect network attacks. A network attack's detection speed becomes possible by combining gathered data and features that identify abnormalities in an analysis of situations.

$$H(X) = -\sum_{i=1}^n (p(x_i)p(x_i)) \quad (1)$$

The proposal detects network anomalies through entropy H evaluation (Eq. 1).

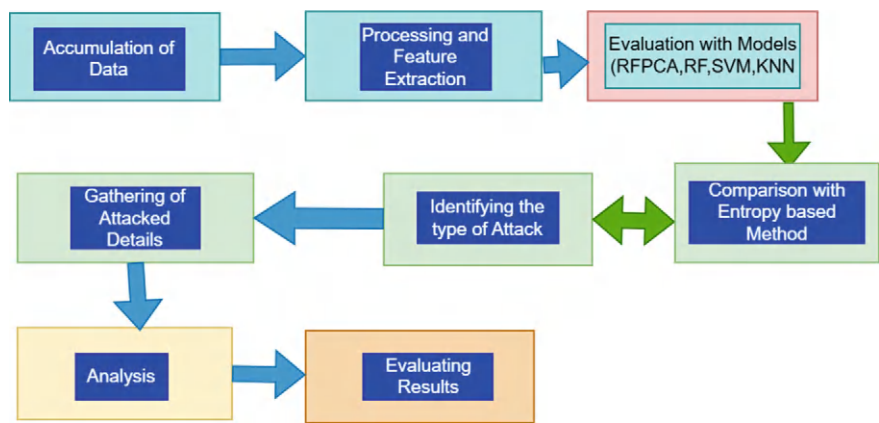


Fig. 2 Proposed detection technique

A set of packets from Fig. 2 are contained in S during the experiment. The value of x_i shows packet transmission numbers at a given source—the total number of packets sent during the period I determines $p(x_i)$. A value of b represents the logarithmic base that calculates the result after the log. The defined threshold was used for comparative analysis and assessment during packet reception. The entropy property used here represents the switch or access point (AP) to indicate the underlying probability distribution, revealing how precisely the attacker’s actions can be appropriately detected. During an attack emergence, the collected features confirm the proper identification of the abnormal behavior.

3.4 One-Time Code and Timestamp

The suggested approach for preventing the fragmentation attack by a predictable identifier is to include a timestamp and one-time code (OTC) in the fragment. The OTC and timestamp consume 40 bytes in the option fields shown in Fig. 3.

3.5 Dataset

A randomly generated dataset emerged from six feature measurements with the properties mentioned in Table 1. This record serves to simulate DDoS attacks. The classifiers construct a prediction model for future attackers using this provided dataset [31].

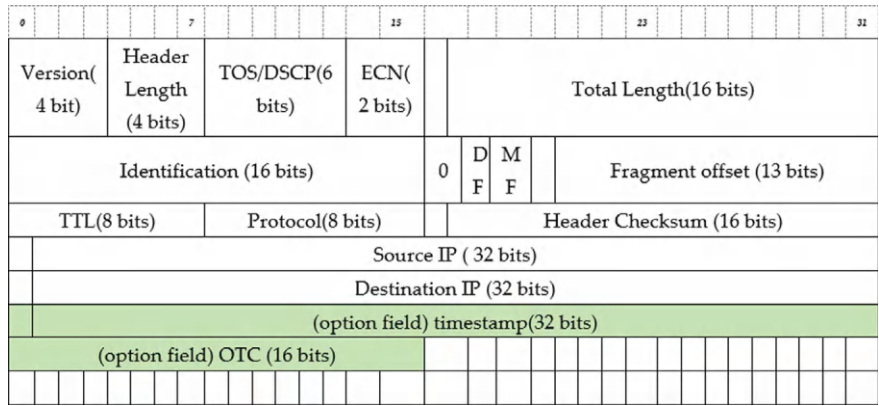


Fig. 3 Enhanced IPv4 header incorporating timestamp and OTC

Table 1 Dataset attributes

No	Field name	Data type	Description
1	Tcp. sport	Numerical	Source port from the sender (incoming)
2	Frame.len	Numerical	Describes the frame length
3	Tcp. Flags. push	Numerical	The PSH flag deals with sending data
4	Ip.flags.df	Numerical	The df flags cannot be fragmented,
5	Packets	Numerical	Total outgoing Data Pkts
6	Bytes	Numerical	Data bytes (incoming)

4 Results and Analysis

This dataset contains 100 nodes of generated traffic, with 62 percent benign traffic and 38 percent malicious traffic. Attack traffic only uses UDP data, whereas benign traffic employs TCP and UDP data. After creating the variable, we load the source IP and TCP SYN flag values into a DataFrame and get the value count on the field to see the top 100 results. The command loads the TCP and SYN flags in hexadecimal for this example, as shown below; the column names in the DataFrame are used in Table 1. All these data types exist within the numerical data range. In short, based on the criteria mentioned above, the dataset is formed. There should be no properties or features with wrong or missing data. In the second step, a correlation matrix is used to validate the features, as shown in Fig. 4. This technique helps mitigate the curse of dimensionality during model training by ensuring that no features have a strong correlation (≥ 0.8).

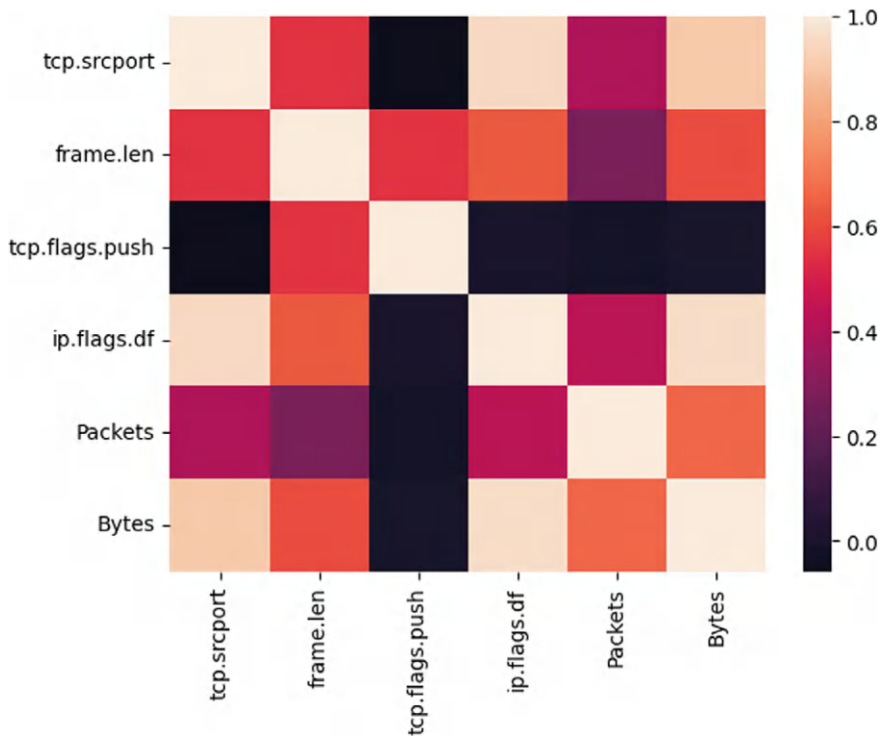


Fig. 4 Correlation matrix of the dataset attributes

4.1 DDoS Detection

Attack synchronization in DDoS functions by creating simulated infected devices known as bots for targeting servers. Each regional network computer device or gateway distributes infected data to adjacent machines [32]. The server-side router faces an overload of data, leading to a decline in service performance. The recovery process of the flow uses detection and protection protocols in the simulation. The evaluation relies on particular mathematical formulas that deduce model assessment results. The simulation runs based on the configuration settings described in Table 2.

It can be formulated as TP = true positive, TN = true negative, FP = false positive, FN = false negative. The formulas adopted in the evaluation of classifiers are shown below.

Precision = TP/(TP + FP) (2)

Recall = TP/(TP + FN) (3)

Table 2 DDoS simulation parameters

Parameter	Value
No. of legitimate nodes	62
No. of attacker	38
No. of router	1
No. of server	1
Attacker target	Server
Simulation time	1 s
Attack duration	0.3–0.4 s
Sending interval (legitimate node)	TCP = 0.2 s

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

4.2 Performance Evaluation

The computations for precision and recall are demonstrated through Eq. (2) and (3). The research data was separated into training data, taking up 75%, while test data consumed 25%. The identification of DDoS depends on how well the classification methods achieve the convergence standards. Additional features in the system bring about improved prediction accuracy levels, according to Table 3. The F1 score in the three features model reached high levels.

The simulation presents the DDoS classification problem using the models shown in Fig. 6. The entropy highlights the varying detection rates of the classifiers in traffic analysis. Model performance indicates that training with more features improves accuracy but may also increase processing time. Entropy values are kept in the range of 5.41–5.55 in the 0–0.5 s before the attack. The attack started at 0.5–0.7 s, and as entropy values decreased, it became clear that the attack was focused on the network and that different classifiers had varying detection rates. Figure 5 illustrates the performance measurement of the root mean squared error.

Table 3 Evaluation of classifiers

Three features				
Classifier	F1	Accuracy	Precision	Recall
RFPCA	91.18	94	93.94	88.57
RF	89.55	93	93.75	85.71
SVM	79.02	83	69.57	91.43
KNN	81.01	85	72.72	91.43

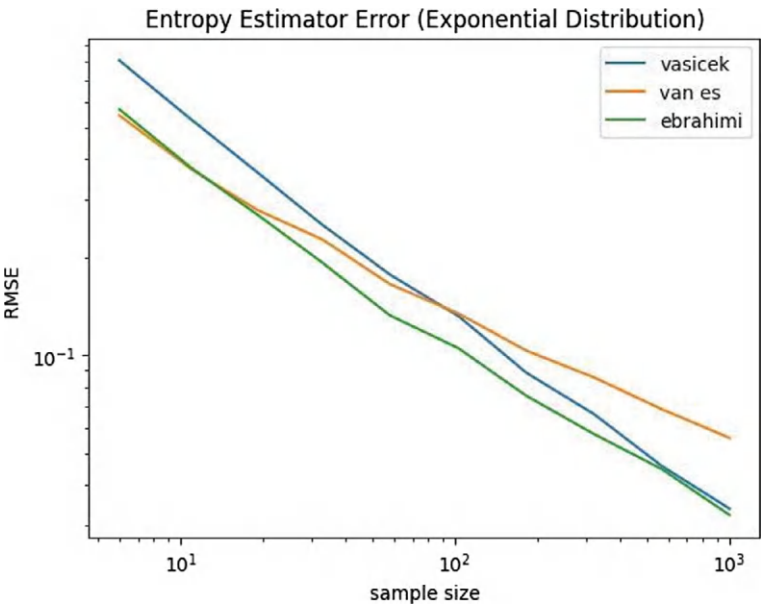


Fig. 5 Entropy calculation

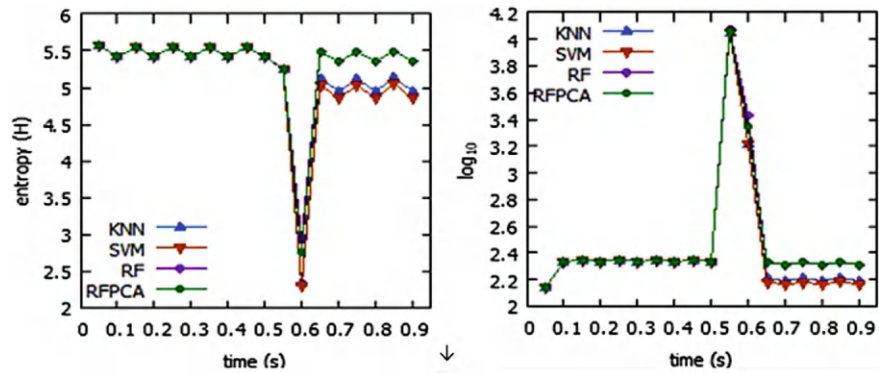


Fig. 6 Comparison of several models trained on three features for attack detection and network protection. **a** Entropy-based analysis. **b** Log10 results with different models' implementation

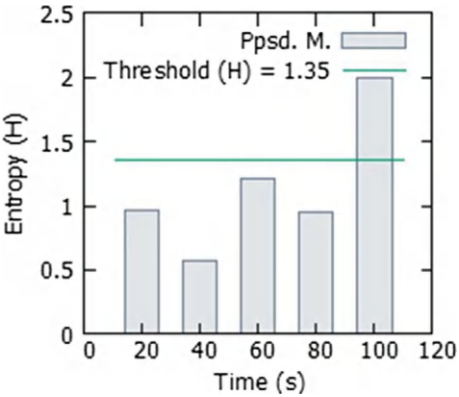
4.3 ARP Spoofing

A hacker uses ARP spoofing to generate an artificial version of a local network connecting wired and wireless network interfaces. Different types of attacks serve as the basis for evaluating detection and prevention efficiency.

Table 4 ARP spoofing simulation parameter

Parameter	Value
Nodes in the local network	1
No. of server	1
Attacker target	Router
Simulation time	100 s
Attack session	15–65 s

Fig. 7 Entropy of the proposed framework



4.4 Performance Analysis

The simulation analysis configuration is shown in Table 4 if an attack occurs. The proposed method employs an entropy cut-off of $H = 1.35$. When an intruder attempts conventional ARP spoofing to scan the network for active IP addresses, the system detects the spoofing activity, activates the proposed technique, and enforces rules to block ARP packets associated with the attacker’s MAC address.

A time window is employed to aggregate ARP packets. Within the 0–80 s interval, entropy values fell below the threshold, illustrating its effectiveness in distinguishing attack traffic from regular activity. As depicted in Fig. 7, victims responded to ARP requests during the ARP process, with entropy values registering at 1.22. The proposed method successfully detected ARP spoofing and halted the malicious traffic within 40 s.

5 Conclusion and Future Scope

The document examines and discusses all aspects of the proposed defense strategy against DDoS attacks. Security threats that target inside and outside networks exist alongside unseen malicious invasion attempts. Organizational infrastructure faces

significant risks from these frequent severe attacks and severely damaging operational structures. The methods used to address these issues contain different detection rates and levels of accuracy, together with their benefits and limits. The proposed RF classifier working with PCA features is a trainable system that produces a defendable model that adapts to changing operational conditions. The system's protective measures must be of sufficient quality to spot online attacks immediately after they happen, and this protection must not harm other operational features. The analytical results validate that random forest (RF) combined with principal component analysis (PCA) is suitable for detecting DDoS attacks. The model maintains excellent outcome accuracy because it correctly separates benign Internet traffic from malicious behavior. The system efficiency improves through PCA integration because it optimizes measurement dimensions by retaining essential features that affect classifier behavior during network dynamism. Our experimental study proves that this proposed system detects threats effectively with minimal false positive occurrences, therefore maintaining its ability to identify threats early. This system shows two crucial capabilities: its ability to handle dynamic attack methods and protect networks without affecting other operational features. The proposed RF-PCA method demonstrated its suitability and successful performance in developing an offensive strategy against DDoS attacks, thus creating conditions for enhancing cyber protection capabilities in complex, large-scale systems.

The RF-PCA-based system shows promising potential as a DDoS detection method, while various enhancement strategies will improve its operational capabilities. Implementing the presented model within real-time IDS systems and testing with different data sources will validate its effectiveness in multiple use cases. Modern approaches like LSTM or CNNs combined with feature selection investigations can boost accuracy rates in customer behavior prediction. The system's implementation scope could be enlarged to detect more cyber threats, including IoT and ARP Spoofing attacks. The operational stability of the model can be achieved by using explainable AI for model interpretability and dynamic threat detection mechanisms to handle emerging security risks. Working with stakeholders for practical implementation will help better evaluate the system's practical applicability.

References

1. Anand, G., Prathiba, S.B.: Detection of man in the middle attacks in Wi-Fi networks by IP spoofing. In: 2018 Tenth International Conference on Advanced Computing (ICoAC), December 2018, pp. 319–322. IEEE (2018)
2. Ghafir, I., Kyriakopoulos, K.G., Aparicio-Navarro, F.J., Lambbotharan, S., Assadhan, B., Binsalleeh, H.: A basic probability assignment methodology for unsupervised wireless intrusion detection. *IEEE Access* **6**, 40008–40023 (2018)
3. Staddon, E., Loscri, V., Mitton, N.: Attack categorization for IoT applications in critical infrastructures, a survey. *Appl. Sci.* **11**(16), 7228 (2021)
4. Lee, Y.J., Chae, H.S., Lee, K.W.: Countermeasures against large-scale reflection DDoS attacks using exploit IoT devices. *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije* **62**(1), 127–136 (2021)

5. Li, D., Yu, C., Zhou, Q., Yu, J.: Using SVM to detect DDoS attacks in SDN networks. *IOP Conf. Ser.: Mater. Sci. Eng.* **466**(1), 012003. IOP Publishing (2018)
6. Lohachab, A., Karambir, B.: Critical analysis of DDoS—an emerging security threat over IoT networks. *J. Commun. Inf. Netw.* **3**, 57–78 (2018)
7. Butun, I., Österberg, P., Song, H.: Security of the internet of things: vulnerabilities, attacks, and countermeasures. *IEEE Commun. Surv. & Tutor.* **22**(1), 616–644 (2019)
8. Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., Durumeric, Z., Halderman, J.A., Invernizzi, L., Kallitsis, M., Kumar, D., Lever, C., Ma, Z., Mason, J., Menscher, D., Seaman, C., Sullivan, N., Thomas, K., Zhou, Y.: Understanding the Mirai botnet. In: 26th USENIX Security Symposium (USENIX Security 17), pp. 1093–1110 (2017)
9. Sinanović, H., Mrdovic, S.: Analysis of Mirai malicious software. In: 2017 25th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), September 2017, pp. 1–5. IEEE (2017)
10. Aytaç, T., Aydın, M.A., Zaim, A.H.: Detection of DDOS attacks using machine learning methods (2020)
11. Mittal, M., Kumar, K., Behal, S.: DDoS-AT-2022: a distributed denial of service attack dataset for evaluating DDoS defense system. *Proc. Indian Natl. Sci. Acad.* **89**(2), 306–324 (2023)
12. Rodríguez, E., Valls, P., Otero, B., Costa, J.J., Verdú, J., Pajuelo, M.A., Canal, R.: Transfer-learning-based intrusion detection framework in IoT networks. *Sensors* **22**(15), 5621 (2022)
13. Xia, Y., Xu, Y., Mondal, S., Gupta, A.K.: A transfer learning-based method for cyber-attack tolerance in distributed control of microgrids. *IEEE Trans. Smart Grid* (2023)
14. Kawish, S., Louafi, H., Yao, Y.: An Instance-based transfer learning approach, applied to intrusion detection. In: 2023 20th Annual International Conference on Privacy, Security and Trust (PST), pp. 1–7 (2023)
15. Chuang, H.-M., Ye, L.-J.: Applying transfer learning approaches for intrusion detection in software-defined networking. *Sustainability* **15**(12), 9395 (2023)
16. Mittal, M., Kumar, K., Behal, S.: DL-2P-DDoSADF: deep learning-based two-phase DDoS attack detection framework. *J. Inf. Secur. Appl.* **78**, 103609 (2023)
17. Lu, H., Zhao, Y., Song, Y., Yang, Y., He, G., Yu, H., Ren, Y.: A transfer learning-based intrusion detection system for zero-day attack in communication-based train control system. *Cluster Comput.* 1–16 (2024)
18. Effah, E.Q., Osei, E.O., Tetteh, A.: Hybrid Approach to classification of DDoS attacks on a computer network infrastructure. *Asian J. Res. Comput. Sci.* **17**(4), 19–43 (2024)
19. Chatterjee, S., Satpathy, S., Paikaray, B.K.: Forecasting DDoS attack with machine learning for network forensic investigation. *Int. J. Reason.-Based Intell. Syst.* **16**(5), 352–359 (2024)
20. Chatterjee, S., Satpathy, S., Nibedita, A.: Digital investigation of network traffic using machine learning. *EAI Endorsed Trans. Scalable Inf. Syst.* **11**(1) (2024)
21. Wang, S., Gomez, K., Sithamparanathan, K., Asghar, M.R., Russello, G., Zanna, P.: Mitigating DDoS attacks in sdn-based IoT networks leveraging secure control and data plane algorithm. *Appl. Sci.* **11**(3), 929 (2021)
22. Sudar, K.M., Beulah, M., Deepalakshmi, P., Nagaraj, P., Chinnasamy, P.: Detection of distributed denial of service attacks in SDN using machine learning techniques. In: 2021 International Conference on Computer Communication and Informatics (ICI), January 2021, pp. 1–5. IEEE (2021)
23. Satpathy, S., Swain, P.K., Mohanty, S.N., Basa, S.S.: Enhancing security: federated learning against man-in-the-middle threats with gradient boosting machines and LSTM. In: 2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), July 2024, pp. 1–8. IEEE (2024)
24. Pei, J., Chen, Y., Ji, W.: A DDoS attack detection method based on machine learning. *J. Phys.: Conf. Ser.* **1237**(3), 032040. IOP Publishing (2019)
25. Aly, M., Khomh, F., Guéhéneuc, Y.G., Washizaki, H., Yacout, S.: Is fragmentation a threat to the success of the internet of things? *IEEE Internet Things J.* **6**(1), 472–487 (2018)
26. Feng, X., Li, Q., Sun, K., Xu, K., Liu, B., Zheng, X., Yang, Q., Duan, H., Qian, Z.: PMTUD is not a panacea: revisiting IP fragmentation attacks against TCP. In: Proceedings of the Network &

- Distributed System Security Symposium (NDSS), San Diego, CA, USA, April 2022, pp. 24–28 (2022)
27. Suci, I., Vilajosana, X., Adelantado, F.: An analysis of packet fragmentation impact in LPWAN. In: 2018 IEEE Wireless Communications and Networking Conference (WCNC), April 2018, pp. 1–6. IEEE (2018)
 28. Dai, T., Shulman, H., Waidner, M.: DNS-over-TCP is considered vulnerable. In: Proceedings of the Applied Networking Research Workshop, July 2021, pp. 76–81 (2021)
 29. Mohandoss, D.P., Shi, Y., Suo, K.: Outlier prediction using random forest classifier. In: 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), January 2021, pp. 0027–0033. IEEE (2021)
 30. Komazec, T., Gajin, S.: Analysis of flow-based anomaly detection using Shannon’s entropy. In: 2019 27th Telecommunications Forum (TELFOR), November, pp. 1–4. IEEE (2019)
 31. <https://www.kaggle.com/datasets/yashwanthkumbam/apaddos-dataset>
 32. Chai, T.U., Goh, H.G., Liew, S.Y., Ponnusamy, V.: Protection schemes for DDoS, ARP spoofing, and IP fragmentation attacks in smart factory. *Systems* **11**(4), 211 (2023)