LATEST TRENDS IN
CYBER SECURITY, AI & ML



Editors
Dr. Sadguru Prakash
Mr. Avinash Singh

NP
Nature Light Publications

An International Edited Book *ISBN-978-81-69492-03-4*

Latest Trends in Cyber Security, AI & ML

Editors

Dr. Sadguru Prakash
Assistant Professor
Department of Zoology
M.L.K.P.G. College, Balrampur, Uttar Pradesh, India.

Mr. Avinash Singh
Assistant Professor and Head
Department of Computer Application
Maharani Lal Kunwari (P.G.) College Balrampur, Uttar Pradesh, India.

Published By

Nature Light Publications, Pune

© Reserved by Editor's

Latest Trends in Cyber Security, AI & ML

Editors

Dr. Sadguru Prakash
Mr. Avinash Singh

First Edition: July, 2026

An International Edited Book

ISBN- 978-81-69492-03-4

Published by:
Nature Light Publications, Pune
309 West 11, Manjari VSI Road, Manjari Bk.,
Haveli, Pune- 412 307.
Website: www.naturelightpublications.com
Email: naturelightpublications@gmail.com
Contact No: +91 9822489040 / 9922489040

The editors/Associate editors/publisher shall not be responsible for originality and thought expressed in the book chapter/ article. The author shall be solely held responsible for the originality and thoughts expressed in their book chapter or article.

Preface

The rapid expansion of digital technologies has transformed the way individuals, organizations, institutions, and governments operate. Alongside the benefits of digital connectivity, cloud computing, online transactions, artificial intelligence, and machine learning, the cyber environment has also become increasingly complex and vulnerable. Cyberattacks, data breaches, phishing, malware, digital fraud, social engineering, and sophisticated intrusion techniques continue to evolve, creating an urgent need for innovative, intelligent, and adaptive security solutions. The edited book Latest Trends in Cyber Security, AI & ML brings together contemporary research and practical perspectives addressing these emerging challenges.

This book presents a multidisciplinary collection of chapters that explore the application of Artificial Intelligence (AI), Machine Learning (ML), cybersecurity frameworks, digital forensics, cryptography, secure authentication, and intelligent threat detection. The contributions demonstrate how modern computational techniques can be employed not only to identify and respond to cyber threats but also to improve prediction, prevention, investigation, and overall digital resilience.

The book begins with applications of machine learning beyond conventional cybersecurity, including multimodal ensemble learning for crop yield prediction and explainable machine learning for GDP growth prediction. These studies illustrate the wider potential of AI and ML in analyzing complex datasets, identifying patterns, and supporting data-driven decision-making. The inclusion of explainable AI also emphasizes the importance of transparency and interpretability in machine learning-based systems.

A significant portion of the book focuses on the growing intersection of AI and cybersecurity. Chapters examine AI-powered network intrusion detection using Random Forest and Isolation Forest, optimized intrusion detection using hybrid feature selection, AI-driven cybersecurity defence systems, and machine-

learning-based phishing and malware classification. These contributions highlight the increasing role of intelligent algorithms in detecting anomalous activities, identifying malicious behavior, and strengthening automated security operations.

The book also addresses important human and socio-technical dimensions of cybersecurity. The chapters on social media security in the AI/ML era, social engineering vulnerability among university students, and cyber hygiene emphasize that cybersecurity is not solely a technological concern. Human behavior, awareness, responsible digital practices, and organizational policies remain essential components of an effective security ecosystem.

Overall, Latest Trends in Cyber Security, AI & ML seeks to provide students, researchers, academicians, IT professionals, cybersecurity practitioners, and technology enthusiasts with a useful reference to current developments in these rapidly evolving fields. The chapters combine theoretical perspectives, analytical approaches, practical applications, and emerging research directions, reflecting the increasingly interconnected nature of AI, machine learning, and cybersecurity.

It is hoped that this volume will encourage further research, innovation, and collaboration toward building secure, intelligent, trustworthy, and resilient digital systems. The editors sincerely acknowledge all the contributors for their valuable research and efforts, and hope that this book will serve as a meaningful resource for understanding and addressing the cybersecurity challenges of the digital future.

Editors

Latest Trends in Cyber Security, AI & ML

Table of Content

Sl. No.	Title and Authors	Page No.
1	**A Multimodal Ensemble Learning Framework for Crop Yield Prediction: A Case Study of India** *Amit Kumar*	01 - 10
2	**Adapting Social Media Security Frameworks for the AI/ML Era: A Socio-Technical Analysis** *Ashish Mishra*	11 - 25
3	**Cyber Fraud in Aadhaar Enabled Payment System (AePS): Threat Landscape, Vulnerability Analysis, and Mitigation Strategies** *Masud Murad Khan*	26 - 39
4	**AI-Powered Network Intrusion Detection Using Random Forest and Isolation Forest on Real CICIDS-2017 Traffic: A Production-Grade Containerized Architecture with SOAR Automation** *Satyam Pandey*	40 - 56
5	**AI-Driven Machine Learning Model for Predicting Social Engineering Vulnerability Among University Students Using Behavioral Analytics** *Vaibhavi Kasaudhan, Pratigya Shukla*	57 - 71
6	**AI-Driven Intelligent Systems for Cybersecurity Defence** *Sundram Dwivedi*	72 - 78
7	**Cyber Hygiene: A Habit for Safe Digital Living** *Prof. Sunita Arya*	79 - 87
8	**Cyber Security in the Digital Age: Challenges and Strategic Solutions** *Sunita Arya, Dr. Sadguru Prakash*	88 - 95
9	**A Review of AI-Enabled Systems for Monitoring and Preventing Malpractice in Online Examinations** *Deepak Kumar, Mrs. Khushboo*	96 - 104
10	**Digital Forensics and Cyber Crime Investigation: Methodologies, Legal Frameworks, and Emerging Paradigms** *Kavyansh Tiwari, Kshitij Mishra, Arman Khan, Sulabh Mishra*	105 -117
11	**Emerging Threats and Innovative Solutions for Cyber Security** *Mr. Ramesh Kumar*	118 -125
12	**Explainable Machine Learning for GDP Growth Prediction: Interpreting Macroeconomic Indicators via SHAP Analysis**	126 -152

	on World Development Indicators *Puneet Tiwari, Avinash Singh*	
13	**Cybersecurity in Digital Payments: Threats, Frameworks, and Countermeasures** *Deepanjali Mishra, Chitranshi Srivastava*	**153 -170**
14	**An Optimized Machine Learning-Based Intrusion Detection System Using Hybrid Feature Selection on the Unsw-Nb15 Dataset** *Shristhi Singh*	**171 -184**
15	**Secure Authentication System based on Multifactor Authentication** *Shriyank Singh, Pratima Singh, Amit Rai Sir*	**185 -191**
16	**A Comprehensive Study of Cryptography Techniques and Applications** *Niyaz Ahmad*	**192 -198**
17	**AI-Based Phishing Detection System Using Machine Learning Techniques** *Prince Kumar Yadav, Shristhi Singh, Samriddhi Prakash*	**199 -206**
18	**Malware Classification Using Supervised Data Mining Techniques** *Samriddhi Prakash*	**207 -213**

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 01 - 10

A Multimodal Ensemble Learning Framework for Crop Yield Prediction: A Case Study of India

Amit Kumar

School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India

Email:

Article DOI Link: https://zenodo.org/uploads/22027324

DOI: 10.5281/zenodo.22027324

Abstract

Agriculture is the backbone of civilisation and plays a vital role in food and nutritional security. Accurate yield prediction enables policymakers to make timely decisions. Traditionally, crop yield prediction was entirely dependent on ground-based observations. Artificial Intelligence and Machine Learning approaches have demonstrated strong predictive capability for crop yield estimation. This study employs Ensemble Learning approaches to develop a robust multimodal crop yield prediction model. The study uses data from 311 districts across 20 Indian states spanning 1966 to 2017. Yield data and vegetation indices were collected from ICRISAT and MODIS via Google Earth Engine, respectively. Four base learners — Random Forest, Support Vector Regression, Gradient Boosting Machines, and Multi-Layer Perceptron — were trained across three progressively richer fusion configurations: Remote Sensing only (R), Remote Sensing and Weather (R+W), and full multimodal fusion (R+W+G). Predictions from each base learner were aggregated using late-fusion ensemble averaging. The full multimodal ensemble achieved an RMSE of 314.51 kg/ha and an R^2 of 0.821, representing a 28.76% reduction in error over the satellite-only baseline. Results demonstrate that ground-based agronomic inputs and spatial encodings deliver the highest gain in prediction, and that ensemble aggregation consistently outperforms any individual base learner. The proposed framework offers a scalable and interpretable approach for operational crop yield forecasting across heterogeneous agro-climatic zones.

Introduction

Agriculture is an important socio-economic sector that serves food to the population. India is still an agricultural country, as the majority of the Indian population is dependent on agriculture [2]. Climate change affected the season and

seasonality, which caused issues of food shortage. The interaction between population growth, unpredictable weather, soil degradation, and climate change must be addressed by implementing appropriate measures for timely productivity. Predicting crop yield is crucial to preventing food shortages and making smarter farming choices. Timely crop yield prediction makes it easier to make decisions for policymakers [1]. Accurate crop yield prediction plays a vital role in ensuring agricultural stability and food security.

Traditional yield prediction approaches have relied heavily on manual field surveys, crop growth simulation models, and agronomist expertise. While informative, these methods are labour-intensive, geographically constrained, and difficult to scale in different geographic regions of a country as large and heterogeneous as India [3]. The advent of satellite remote sensing, open meteorological databases, and large-scale agricultural archives has created an opportunity to develop data-driven models that are both scalable and cost-effective.

Machine Learning (ML), Deep Learning (DL), and Artificial Intelligence (AI) techniques have emerged as alternatives to traditional crop yield estimation and prediction [4, 5]. However, most of the studies rely on a single modality, which fails to capture the complex relationship between input and output features. Integration of multiple modalities has the potential to improve predictive accuracy, but raises challenges due to data heterogeneity, feature alignment and fusion strategy.

This study leverages Ensemble learning to improve prediction accuracy. The study uses three distinct data sources and multimodal Ensemble Learning to develop a robust model for crop yield prediction.

Related Works

In the study [3], the authors utilised eight different regression models, including Multiple Linear Regression, Decision Tree, Elastic Net, Lasso, Ridge, Partial Least Squares, Gradient Boosting, and LSTM. The study uses district-level data of Maharashtra, India, spanning 18 years to develop a predictive model. Study shows that the feature engineering-based LSTM model results in the highest accuracy at 86.3%. Article [8] proposed a "Temporal Convolutional Network (TCN) based model for predicting rice crop yield" that utilises multispectral satellite and climate data to predict crop yield. To analyse temporal dependencies of vegetation indices, TCN uses dilated convolution. Their proposed TCN model achieves 98.01% R^2 accuracy with minimal MAE (0.91) and RMSE (0.96). Authors of the paper [9] developed a model for the Rajasthan region of India. Decision Tree (DT), Random Forest (RF), Lasso Regression, and Support Vector Regression (SVR) were trained, using MODIS satellite-derived vegetation indices (NDVI, EVI, LAI) and weather data spanning 2000-2020. Study shows RF achieves the best performance from the trained models, and integrating multiple vegetation indices with weather data improves prediction accuracy rather than using only vegetation indices. The study

[10] proposed a hybrid system which combines Artificial Neural Networks and Levy flight-enhanced Coati Optimisation Algorithm name ANN-COA. The study also demonstrates that Biological factors and climatic conditions are important factors for crop yield prediction. To overcome the challenges of regional heterogeneity, the study [11] introduces a hybrid approach integrating ENKF-WOFOST with a CNN-ALSTM-MTL architecture. This fusion of data assimilation and deep learning ensures more robust, timely, and precise maize yield estimations at the county level. To improve drought-induced yield loss predictions, [12] research tests an ensemble Random Forest approach across lumped and distributed scales. Central to this study is a novel training framework that benchmarks traditional random selection against a more targeted, hydrologically balanced approach.

Dataset and Preprocessing

1. Dataset Description

The data used in the experiments was obtained from "ICRISAT (International Crops Research Institute For The Semi-Arid Tropics)" [6] and comprises 16,032 entries gathered from 311 districts across 20 Indian states between 1966 and 2017. Vegetation Indices and other biophysical features were obtained from Google Earth Engine [7] using a GEE Script. It is centred around 15 Kharif crops and provides information on crop yields (kg/ha), climatic variables (rainfall, temperature, precipitation), and fertiliser use (nitrogen, phosphate, potash). Geographical features like state and district names, year-wise data render the dataset appropriate for regional and time-series-based analysis in agricultural yield forecasting.

2. Preprocessing

The preprocessing pipeline follows a five-stage sequence designed to preserve data integrity while constructing modality-aware feature representations suitable for multimodal fusion.

In the first step of preprocessing, temporal standardisation has been done to convert annual integer records into a unified date format, to establish a consistent time-series index that accommodates future extension to sub-annual temporal resolutions.

The second step of the process addresses feature organisation and data leakage prevention: variables are partitioned into three modality groups: Remote Sensing (R), Weather (W), and Ground (G), and all non-target yield columns are explicitly excluded before model training to ensure no target-related information contaminates the feature space.

The third step employs a three-tier imputation cascade to handle missing values. Local linear interpolation is applied first within each district-year series to exploit temporal continuity, state-level median imputation serves as a secondary fallback

for districts with insufficient observations, and global median imputation resolves any residual gaps. This imputation resulted in a dataset with zero missing values.

In the fourth step, a cumulative seasonal rainfall feature is derived by aggregating monthly precipitation totals, providing an agronomically meaningful proxy for total water availability during the growing season.

The fifth and final step of preprocessing, one-hot encoding, is leveraged to deal with categorical values of the dataset. Z-score standardisation is then applied independently within each modality, preserving the inherent scale differences between data sources while ensuring intra-modality feature comparability, a critical consideration in multimodal learning frameworks where cross-modality scale conflation can distort model behaviour.

Multimodal Fusion Framework

Single and hybrid source combinations are employed to implement multiple fusion modules, each capturing complementary information from different data modalities. The outputs of these fusion modules are subsequently integrated within an ensemble learning framework, which aggregates predictions through a meta-learner to reduce model uncertainty, mitigate individual learner biases, and improve overall prediction performance.

To understand how each data modality contributes to prediction, we created three sets of features that get progressively more detailed:

Multimodal Fusion Levels

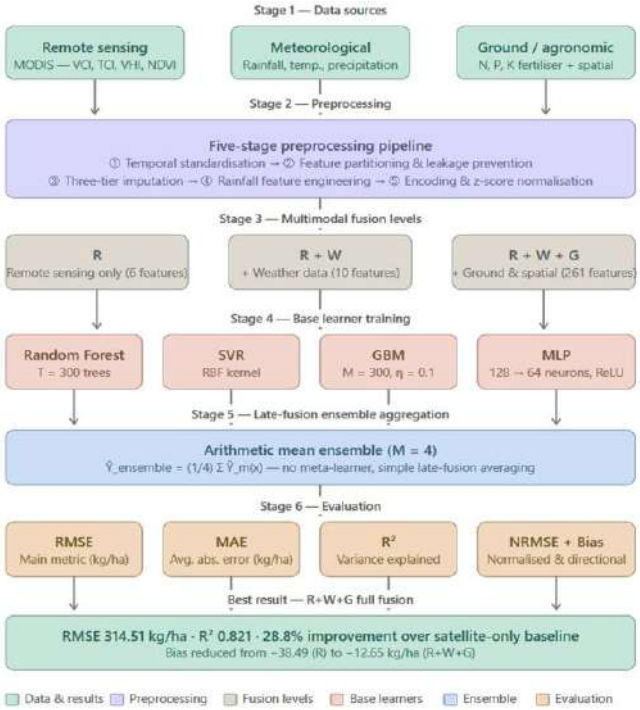


Fig. 1: Proposed multimodal ensemble learning framework for crop yield prediction.

TABLE- I. Multimodal Fusion Levels

Fusion Level	Modalities Included	Description
R	Remote Sensing	Vegetation & drought indices only
R+W	Remote Sensing + Weather	Satellite + meteorological data
R+W+G	All Modalities + Spatial	Full multimodal fusion

The Remote Sensing (R) configuration is a baseline that represents the condition where only remote sensing data are present. This scenario is particularly relevant for the geographically remote areas or data-sparse regions. Here, R represents the feature set that is extracted from remote sensing satellite images. The Remote Sensing + Weather (R + W) represents the integration of remotely sensed features extracted from satellite images and weather and climate features. The R + W + G (Remote Sensing + Weather + Ground Data) represents the integration of all modalities, where G contains the agricultural management data like soil type, fertilisers, and other farming management features. Full modality combination enables a systematic evaluation of multimodal interactions and their collective contribution to predictive performance.

Remote sensing indices capture vegetation health and drought stress dynamics; meteorological variables characterise the hydrothermal conditions across the growing season; ground-based agronomic inputs reflect soil nutrient status and management practices; and spatial encodings encapsulate regional agro-climatic heterogeneity and soil variability that cannot be inferred independently from other data sources.

1. Base Learner Training

To capture diverse feature representations from a heterogeneous agricultural dataset, four base learners with complementary inductive biases were trained. Their hyperparameters were optimised through randomised search combined with cross-validation. We selected Random Forest (RF), Support Vector Regression (SVR), Gradient Boosting Machines (GBM), and Multi-Layer Perceptron (MLP) because they offer diverse inductive biases. By combining ensemble methods, kernel-based regression, and neural networks, the framework can capture both linear and complex non-linear feature interactions within the heterogeneous agricultural dataset.

a. **Random Forest (RF):** RF is an Ensemble Learning technique utilizes predictive power of decision trees using bootstrap aggregation and random feature selection. RF captures non-linear feature interactions and is robust to

outliers, making it suitable for heterogeneous agricultural data. The predictive function is defined as:

$$\hat{Y}_{RF}(x) = \frac{1}{T}\sum_{t=1}^{n} f_t(x)$$

where T=300 denotes the number of trees and f_t (x) represents the prediction of the t^th tree. This formulation enables effective modelling of higher-order non-linear feature interactions while maintaining robustness to noise and outliers.

b. Support Vector Regression (SVR): SVR is a kernel-based regression that finds the maximum-margin hyperplane in a high-dimensional feature space. It utilises Radial Basis Function to find the optimal hyperplane. The regression function of SVR is expressed as:

$$\hat{Y}_{SVR}(x) = \sum_{t=1}^{T}(a_i - \alpha_i^*)K(x_i, x) + b$$

where,

$$K(x_i, x) = \exp(-\gamma \mid x_i - x \mid^2)$$

The model optimises an ε epsilon-insensitive loss function to ensure robustness to small perturbations in the data.

c. Gradient Boosting Machines (GBM): An additive ensemble in a stage-wise manner is constructed by Gradient Boosting through sequentially minimising a differentiable loss function. The model updates an iteration using the equation:

$$F_m(x) = F_{m-1}(x) + \eta h(x)$$

where h(x) denotes the weak learner (decision tree with maximum depth = 5), η=0.1 is the learning rate, and the ensemble consists of M=300 iterations. This iterative residual fitting mechanism reduces bias and improves predictive accuracy.

d. Multi-Layer Perceptron (MLP): MLP is a feedforward neural network that approximates complex nonlinear relationships between input features and the output.

$$a^{(l)} = \sigma\left(W^{(l)}a^{(l-1)} + b^{l}\right)$$

where σ (·) is the Rectified Linear Unit (ReLU) activation function. The network is configured with two hidden layers with 128 and 64 neurons,

respectively, and it is trained for a maximum of 600 epochs using early stopping with a 10% validation split to prevent overfitting.

2. Ensemble Aggregation

The final ensemble prediction is computed as the arithmetic mean of the individual base learner predictions:

$$\hat{Y}_{ensemble} = \frac{1}{M}\sum_{m=1}^{M}\hat{Y}_m(x)$$

where M=4 denotes the number of base learners and Yˆ_m (x) represents the prediction of the m^th model. Simple averaging is adopted due to its transparency and effectiveness when combining models with diverse inductive biases, thereby reducing the likelihood of correlated prediction errors. The absence of a meta-learner preserves interpretability and minimises the risk of overfitting.

3. Evaluation of the Models

Performance of the models was evaluated using the following metrics:

- **Root Mean Squared Error (RMSE):** Main evaluation metric; penalizes larger mistakes by squaring them.
- **Mean Absolute Error (MAE):** Less sensitive to outliers; expresses average absolute difference in meaningful units (kg/ha).
- **Coefficient of Determination (R^2):** Ratio of the dependent variable's variance accounted for by the model.
- Normalized Root Mean Squared Error (NRMSE= RMSE /mean(y)).

Results and Discussion

Experiments were conducted across three fusion levels to assess the contribution of each data modality: Remote Sensing only (R), Remote Sensing + Weather (R+W), and Remote Sensing + Weather + Ground Data (R+W+G). For each fusion level, multiple base learners were trained, followed by late-fusion ensemble aggregation. Results for the individual base learners trained during the experiment are shown in Table II and Fig. 2.

Model	R – RMSE (kg/ha)	R+W – RMSE (kg/ha)	R+W+G – RMSE (kg/ha)
Random Forest	620.0042	568.6158	443.7757
SVR	674.0208	641.4135	611.0471
GBM	638.3158	571.0818	446.5398
MLP	640.4005	604.8470	414.6916

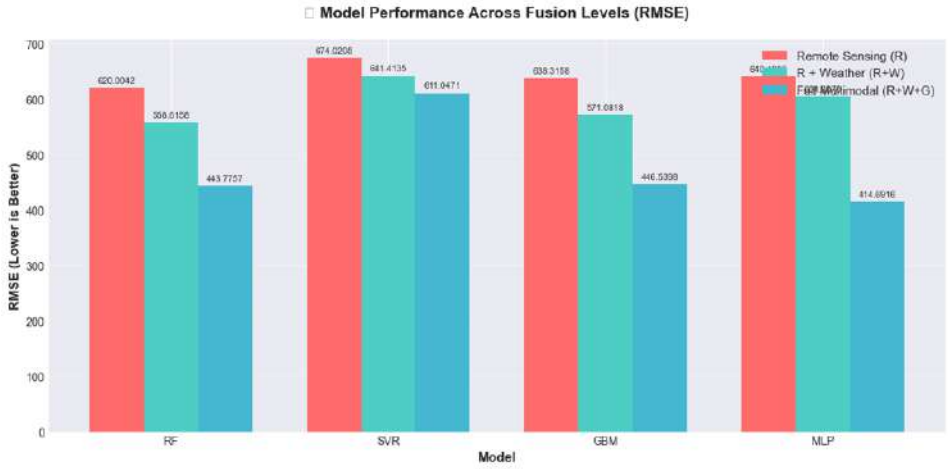


Fig. 2: RMSE Performance of Base Models Across the Fusion Level

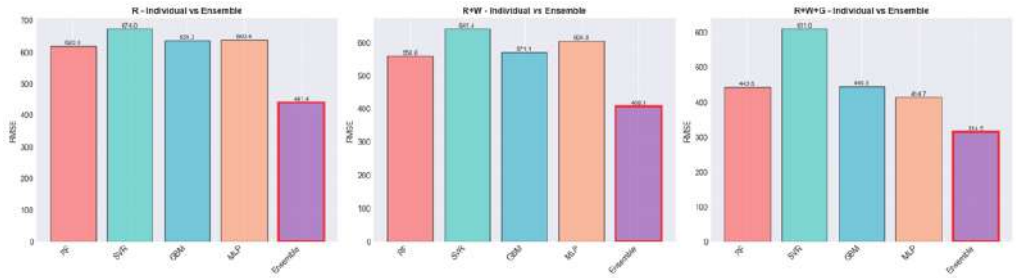


Fig. 3: Ensemble models RMSE Performance Metric Across Fusion Level

Table III: Ensemble Performance Metrics Across Fusion Levels

Fusion Level	RMSE (kg/ha)	MAE (kg/ha)	R^2	NRMSE	Bias (kg/ha)
R (Remote Sensing)	441.36	330.29	0.649	0.630	−38.49
R+W (+ Weather)	408.06	295.74	0.701	0.582	−29.31
R+W+G (Full)	314.51	247.18	0.821	0.449	−12.65

Table IV: Marginal RMSE Improvement from Multimodal Fusion

Transition	RMSE Before (kg/ha)	RMSE After (kg/ha)	Improvement (%)
R → R+W	441.36	408.06	9.90%
R+W → R+W+G	408.06	314.51	22.92%
R → R+W+G (total)	441.36	314.51	28.76%

The predictive framework integrates three distinct data sources: remote sensing indices for vegetation and soil conditions, meteorological features for seasonal

phenology, and ground data of agricultural management. The results demonstrate that the performance improvement from R+W to R+W+G (22.92%) significantly outweighs the gain from R to R+W (9.90%). This improvement suggests that remote sensing drought metrics (VCI, TCI, VHI) partially internalise rainfall patterns, creating information redundancy with meteorological sets. In contrast, variables such as soil nutrients and district-level management remain genuinely orthogonal to environmental observations, accounting for the marked increase in model accuracy upon their inclusion.

Conclusion

This paper presents a multimodal ensemble learning framework for spatio-temporal crop yield prediction that systematically quantifies the benefit of fusing remote sensing, meteorological, and agricultural management ground data. Applied to crop yield prediction across Indian districts, the full multimodal ensemble achieved an RMSE of 314.51 kg/ha and R^2 of 0.821, a 28.76% error reduction over the satellite-only baseline.

The key findings of this study are as follows: multimodal fusion consistently outperforms unimodal approaches, with ground data and spatial encodings delivering the greatest marginal benefit; ensemble averaging of four diverse base learners reliably outperforms the best individual base learner by approximately 24% in RMSE; and the preprocessing pipeline, including district-wise temporal interpolation and modality-independent normalisation, effectively handles the spatio-temporal complexity of agricultural datasets.

In future work, crop-specific and region-specific models could be developed to improve granularity beyond the pan-India scope of the current study. Spatially aware models, such as Graph Neural Networks that leverage district-level adjacency structures, could be explored to better capture spatial dependencies across agro-climatic zones.

References

1. R. Fang, L. Yu, N. Xie, K. Wang, K. Rao, L. Zhong, J. Chen, and Z. Wu, "A knowledge-informed fusion network for process-aware crop yield prediction," Int. J. Appl. Earth Obs. Geoinf., vol. 146, art. no. 105154, 2026, doi: 10.1016/j.jag.2026.105154.
2. S. Iniyan, V. A. Varma, and C. T. Naidu, "Crop yield prediction using machine learning techniques," Adv. Eng. Softw., vol. 175, art. no. 103326, Jan. 2023, doi: 10.1016/j.advengsoft.2022.103326.
3. S. Agarwal and S. Tarar, "A hybrid approach for crop yield prediction using machine learning and deep learning algorithms," in Journal of Physics: Conference Series, vol. 1714, no. 1, p. 012012, Jan. 2021, doi: 10.1088/1742-6596/1714/1/012012.

4. H. Burdett and C. Wellen, "Statistical and machine learning methods for crop yield prediction in the context of precision agriculture," Precision Agriculture, vol. 23, no. 5, pp. 1553-1574, 2022, doi: 10.1007/s11119-022-09904-y.
5. T. Van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," Computers and Electronics in Agriculture, vol. 177, p. 105709, 2020, doi: 10.1016/j.compag.2020.105709.
6. ICRISAT, "International Crops Research Institute for the Semi-Arid Tropics," 2026. [Online]. Available: https://www.icrisat.org/. [Accessed: Spt. 12, 2025].
7. Google Earth Engine, "Code Editor," Sep. 15, 2025. [Online]. Available: https://code.earthengine.google.com/. [Accessed: Nov. 15, 2025].
8. A. Mohan, M. Venkatesan, P. Prabhavathy, and A. Jayakrishnan, "Temporal convolutional network-based rice crop yield prediction using multispectral satellite data," Infrared Physics & Technology, vol. 135, p. 104960, 2023, ISSN: 1350-4495, doi: 10.1016/j.infrared.2023.104960.
9. K. Jhajharia and P. Mathur, "Prediction of crop yield using satellite vegetation indices combined with machine learning approaches," Advances in Space Research, vol. 72, no. 9, pp. 3998–4009, 2023, doi: 10.1016/j.asr.2023.07.020.
10. T. K. Nida Fariz and S. Sharief Basha, "Enhanced crop yield prediction using a hybrid artificial neural network with coati optimization algorithm," Results in Engineering, vol. 28, p. 107529, 2025, ISSN: 2590-1230, doi: 10.1016/j.rineng.2025.107529.
11. J. Cao, Z. Zhang, Y. Luo, L. Zhang, J. Zhang, Z. Li, and F. Tao, "Improving maize yield prediction at the county level from 2002 to 2015 in China using a novel deep learning approach," Computers and Electronics in Agriculture, vol. 202, art. no. 107376, 2022, doi: 10.1016/j.compag.2022.107376.
12. H. Sahu, P. K. Garg, S. Vijay, and A. Dasgupta, "Predictive drivers and transferability of multi-scale machine learning based crop yield prediction under drought across European and Asian climates," Agricultural Water Management, vol. 327, p. 110254, 2026, ISSN: 0378-3774, doi: 10.1016/j.agwat.2026.110254.
13. D. Nayakpara, D. Prajapati, K. Pandey, and N. R. Patel, "District-Level Groundnut Yield Prediction in Gujarat State Using Machine Learning and Remote Sensing Data," Journal of the Indian Society of Remote Sensing, pp. 1-12, 2026.
14. J. Zhang et al., "Transfer learning for improved crop yield predictions in a cross-scale pathway: a case study for Brazilian national soybean," International Journal of Applied Earth Observation and Geoinformation, vol. 145, p. 104981, 2025, ISSN: 1569-8432, doi: 10.1016/j.jag.2025.104981.

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 11 - 25

Adapting Social Media Security Frameworks for the AI/ML Era: A Socio-Technical Analysis

Ashish Mishra

Assistant Professor, M. L. K. PG College, Balrampur, Uttar Pradesh, India

Email: ashishmishraedu1994@gmail.com

Article DOI Link: https://zenodo.org/uploads/22027530

DOI: 10.5281/zenodo.22027530

Abstract

As Artificial Intelligence (AI) and Machine Learning (ML) redefine the digital perimeter, social media environments have become primary vectors for AI-driven cyber threats. This study investigates user security and privacy through an interdisciplinary lens, integrating computer science and behavioral psychology. We analyze the role of platform algorithms in shaping security outcomes and the emergence of AI-specific risks, such as deepfakes and algorithmic manipulation. Our findings underscore the necessity of ML-based defense mechanisms, such as Sybil detection and intrusion prevention systems, to mitigate modern cyber-threats.

Keywords: Social media, Theoretical Insights, Socio-Technical Factors

Introduction

The proliferation of social media platforms has revolutionized the way individuals interact, communicate, and share information online. With billions of users worldwide engaging in various social media activities daily, concerns regarding user security and privacy have become paramount. Social media environments serve as virtual spaces where users disclose personal information, interact with others, and engage in various online activities, raising significant questions about the protection of their data and privacy.

Theoretical insights into user security and privacy in social media environments are essential for understanding the complex dynamics at play within these digital spaces. Such insights delve into the underlying principles, mechanisms, and factors influencing user security and privacy in social media interactions. By employing theoretical frameworks, researchers can elucidate the multifaceted nature of these phenomena and identify potential strategies for mitigating risks and enhancing user protection (Venugopal, K. et al. 2015).

At the core of this study is the recognition that social media platforms serve as rich ecosystems where users navigate a delicate balance between social interaction and personal privacy. Theoretical perspectives draw from diverse disciplines such as sociology, psychology, communication studies, and computer science to provide a holistic understanding of user behaviors, platform dynamics, and socio-technical factors shaping security and privacy outcomes (Venugopal, K.& Das, 2022).

Sociological theories offer valuable insights into the social dynamics of online interactions, highlighting the role of norms, trust, and social capital in shaping user behavior and attitudes towards privacy. Psychological theories shed light on individual motivations, cognitive biases, and privacy preferences, informing our understanding of why users may engage in risky behaviors or disclose sensitive information online.

Communication theories contribute to the examination of information flow, disclosure patterns, and privacy management strategies within social media environments (Venugopal, K., et al. 2016). They explore how users negotiate privacy boundaries, engage in self-presentation, and manage their online identities amidst evolving communication norms and platform affordances.

From a technical standpoint, theoretical frameworks rooted in computer science and information security provide insights into the technological infrastructure, algorithmic processes, and security mechanisms underpinning social media platforms. These theories elucidate the vulnerabilities, threats, and privacy risks inherent in the design and operation of these platforms, guiding efforts to develop robust security solutions and privacy-enhancing technologies.

Moreover, interdisciplinary approaches that integrate insights from various theoretical perspectives enable a comprehensive understanding of user security and privacy in social media environments. By synthesizing sociocultural, psychological, communicative, and technical dimensions, researchers can unravel the complexities of online privacy and security, paving the way for more effective interventions, policies, and design guidelines (Gopalakrishna, V. et al. 2021).

In summary, theoretical insights into user security and privacy in social media environments provide a nuanced understanding of the underlying dynamics shaping user experiences and outcomes in these digital spaces. By drawing upon diverse disciplinary perspectives and interdisciplinary frameworks, researchers can address the multifaceted challenges and opportunities inherent in safeguarding user security and privacy online.

The increasing ubiquity of social media platforms has brought forth complex challenges concerning user security and privacy. While these platforms offer unprecedented opportunities for social interaction, content sharing, and information dissemination, they also expose users to various risks, ranging from data breaches and identity theft to surveillance and exploitation. Despite the proliferation of privacy settings and security features, users continue to grapple with maintaining

control over their personal information and safeguarding their online privacy (Murli Krishna, P.N. et al. 2020)

One of the primary issues revolves around the collection, storage, and utilization of user data by social media companies and third-party entities. Social media platforms often collect vast amounts of user information, including demographic data, browsing history, location data, and content preferences, to fuel targeted advertising, algorithmic recommendations, and personalized services. However, the opacity surrounding data collection practices, the lack of user consent mechanisms, and the potential for data misuse raise concerns about privacy infringement and user autonomy.

Furthermore, the inherent architecture and functionality of social media platforms introduce vulnerabilities that can be exploited by malicious actors to compromise user security. Cyberattacks, phishing scams, malware distribution, and account hijacking are just a few examples of the threats facing social media users. Moreover, the interconnected nature of social networks amplifies the spread of misinformation, fake accounts, and malicious content, undermining trust and exacerbating security risks.

Another critical issue pertains to the erosion of privacy boundaries and the blurring of public and private spheres in online environments. Social media users often grapple with the tension between sharing personal information to connect with others and maintaining privacy to protect sensitive data. The fluidity of privacy settings, the persistence of digital footprints, and the lack of control over shared content contribute to a sense of vulnerability and exposure among users.

Moreover, social media platforms' business models, which rely heavily on advertising revenue and user engagement metrics, incentivize the commodification of user data and the prioritization of profit over privacy (Venugopal, K.et al. 2022). This commercialization of user information creates conflicts of interest between platform operators, advertisers, and users, undermining efforts to prioritize user security and privacy.

Additionally, cultural and contextual factors shape users' perceptions, attitudes, and behaviors regarding security and privacy in social media environments. Cultura norms, societal expectations, and regulatory frameworks vary across regions and demographics, influencing how users navigate privacy settings, interact with others, and perceive online risks (Venugopal, K. et al. 2020). Understanding these cultural nuances is crucial for developing tailored interventions and policies that address the diverse needs and preferences of social media users worldwide.

In light of these challenges, there is a pressing need for theoretical insights into user security and privacy in social media environments. Such insights can inform the development of comprehensive frameworks, strategies, and interventions aimed at promoting user empowerment, enhancing platform accountability, and fostering a safer and more privacy-respecting online ecosystem. By addressing the multifaceted

dimensions of user security and privacy, researchers can contribute to the advancement of knowledge in this critical area and support efforts to mitigate risks and protect user rights in the digital age.

Objectives of the Study

Objectives of the study are to

- Investigate how social media users perceive and experience security and privacy concerns within online environments, shedding light on their attitudes, behaviors, and coping mechanisms.
- Examine the role of social media platform features, policies, and algorithms in shaping user security and privacy outcomes, providing insights into the socio-technical factors influencing online interactions.
- Identify and analyze emerging threats, vulnerabilities, and risks to user security and privacy in social media environments, including the impact of evolving cyber threats, misinformation campaigns, and data breaches.

Literature Review

Various social media platforms and services are accessible in the contemporary landscape, including but not limited to Facebook, YouTube, Twitter, Instagram, LinkedIn, and Snapchat, as noted by A. Smith and M. Anderson in 2018.

In today's world, the Internet stands out as the ubiquitous technology used by nearly everyone on a daily basis. With Internet technology, individuals worldwide, regardless of their level of experience with information technology, can easily engage in communication, share information, participate in various activities, and utilize the Internet for a multitude of purposes, as highlighted by D. Chaffey in 2019.

Social media serves the purpose of facilitating personal and business-oriented interactions among individuals globally, as outlined by Endosomal et al. in 2011. It is a platform that enables users to share content.

Despite the positive aspects and various constructive impacts, there are also negative effects associated with wasting time. This can lead to academic failures in schools and universities, as well as adverse impacts on social aspects such as marriage, employment, and various other destructive issues. Researchers worldwide urgently need to investigate these issues, finding solutions to mitigate or eliminate the destructive consequences, as emphasized by Abbas et al. in 2019.

The growing prevalence of social media usage has turned it into a breeding ground for cybercriminals and cybercrime. Previous research indicates numerous threats and security risks confronting users of social media, with organizations also falling victim. Those particularly susceptible to attacks and threats are individuals or entities lacking a culture of safe Internet usage, as highlighted by Alguliyev et al. in 2018. This implies that virtually all social media platforms carry security risks

The risks posed by these issues can extend to governments, impacting national security and economic stability. Consequently, it is imperative for everyone to cultivate a baseline level of Internet literacy, enabling awareness of illegal activities, potential threats, and the importance of protective measures. This involves using Internet applications efficiently and avoiding misuse, as emphasized by Kumar and Somani in 2018.

Social media serves as a platform enabling users to share diverse content, encompassing various forms of information such as messages, documents, and videos covering a wide array of topics. Additionally, this service provides users the opportunity to disseminate new ideas, express opinions, and share thoughts with a broad audience, as highlighted by McCarroll and Curran in 2013

Over the past decade, social media platforms like Facebook, Twitter, Snapchat, YouTube, and Instagram have experienced significant expansion in user numbers, as noted by E. Ortiz-Ospina in 2019. This surge in popularity is attributed to widespread adoption, particularly through smartphone usage, as a means of communication, knowledge-sharing, expressing thoughts, and sharing various forms of media. The diverse features offered by these platforms continue to attract an ever-growing global audience.

It is crucial to disseminate this information within communities, organizations, and universities through knowledgeable individuals who can actively educate others on these matters. These advocates should strive to establish and promote best practices that can be standardized and regularly updated, akin to applications, as emphasized by Oxley in 2011

Civilized regions encompass countries and cities that possess a comprehensive range of services, including the internet and other technologies, as defined by Chan and Virkki in 2014. It is widely acknowledged that there exist distinct variations in cyberculture among users in different countries or cities.

Research indicates that numerous organizations can effectively revise their security policies to encompass social media without necessitating the creation of separate and specific guidelines for it. The findings also emphasize the importance of implementing a robust overall security awareness program, coupled with technical and administrative safeguards, for organizations and government officials to consider, as highlighted by Hogben in 2007 and Hiatt & Choi in 2016.

In broad terms, the threats can be categorized into two forms. The first is the input form, where users share their personal information, reflecting their cultural background. Much of this personal information provides attackers with a comprehensive understanding of an individual, including details necessary for various attacks such as credit card fraud or identity theft. Sharing routine activities like vacations and current locations amplify the attacker's level of control, as noted by Adu Michael &Adewale Olumide in 2014.

This constitutes a significant aspect of the problem, and a potential solution to address this will be suggested later. Occasionally, this behavior stems from a desire for popularity, while at other times, it is simply a manifestation of pure disregard (Barnes, S. 2006).

He further emphasizes that although Facebook incorporates privacy settings that users can manage, the default setting upon creating an account is set to public. Therefore, a new user who does not adjust these settings to make them more stringent is essentially sharing information that is visible to the public and non-friends (Verma, A. et al. 2013).

Facebook has undergone several updates to its security framework with the aim of enhancing user- friendliness in customizing settings and granting users greater control over the visibility of individual posts. Although this doesn't offer foolproof security, it proves beneficial as long as users are aware of the features and use them judiciously. Nevertheless, relying solely on these settings is not a substitute for exercising prudence in online content sharing. The speaker continues to highlight that many educational institutions are taking proactive measures to educate students on the significance of online privacy, aiming to bolster overall security in the digital realm (Verma, A. et al. 2013)

This strategy encompasses a feature allowing a user to either accept or decline another user's request for information, irrespective of their friendship status. Furthermore, users have the flexibility to establish two distinct databases for information based on their level of trust in the requester. By doing so, they gain the ability to safeguard their most sensitive information, creating a nuanced approach to data sharing that reflects the user's varying levels of trust in different individuals. This not only provides users with a more tailored and secure experience but also underscores the importance of user control in managing the flow of personal information within online platforms (Barnes, S. 2006).

Implementing such a system would significantly contribute to preventing a potential adversary from fully recovering a user's profile. An example of a social network that employs this approach is Diaspora. Furthermore, as highlighted, it is imperative to conduct thorough risk management. This involves assessing and mitigating potential risks, ultimately aiding in the establishment of a robust security policy within the organization in question. By prioritizing risk management, organizations can proactively address vulnerabilities, enhance resilience, and fortify their security measures, thereby creating a safer digital environment for users. This underscores the importance of a holistic approach to security that combines user-level controls with overarching organizational strategies (Kim, H. J. 2012).

An assailant has the potential to embed malicious scripts into URLs, and when users click on these URLs, the injected script may execute on their systems. This execution poses a significant threat as it can enable the collection of sensitive information from the compromised system (Zhang Z, Gupta BB (2018). This form

of cyber-attack is known as URL-based script injection, where the attacker exploits vulnerabilities in web security to introduce harmful scripts. These scripts can execute unauthorized actions on the user's system, such as data theft, and pose a considerable risk to the security and privacy of the affected individuals. Implementing robust security measures, including secure coding practices and regular system updates, is essential to mitigate the risks associated with URL-based script injections.

Cyber espionage represents a specific category within targeted attacks. In the exploration of this domain, Sahoo et al. presented a comprehensive ATA (Advanced Persistent Threat) detection framework. This framework is meticulously crafted to identify and counteract targeted attacks, as elaborated in their work (Sahoo SR, Gupta BB (2020). Additionally, the authors introduced a system design checklist, specifically tailored to enhance the capability of detecting and mitigating the impact of such sophisticated and focused cyber threats. This framework and checklist collectively contribute to strengthening the defense mechanisms against cyber espionage, providing a valuable resource for organizations aiming to fortify their cybersecurity posture

It is advisable to revoke access to an application if it remains unused for an extended period. Numerous third-party applications utilize social media accounts for authentication. To address security and privacy apprehensions, it is prudent to grant access exclusively to applications that are deemed trustworthy, as emphasized in [4]. This precautionary measure ensures a more secure and privacy-conscious approach to managing third- party application access tied to social media accounts (Bailey, Michael. et al 2009).

The implementation of an Intrusion Detection and Prevention System (IDSS) effectively addresses the challenges associated with low accuracy and prolonged response times. Leveraging machine learning classifiers within the IDSS framework not only mitigates these issues but also facilitates swift response times, a critical factor in the timely detection of spam on Facebook, as discussed in reference (Rathore S et al. 2018). This integration of machine learning enhances the system's efficiency, enabling it to rapidly identify and combat spam, thereby contributing to an improved overall security posture on the platform. The utilization of advanced technologies within IDSS underscores its pivotal role in proactively safeguarding against spam-related threats.

In their work, Al-Qurashi et al. (2018) introduced an innovative Sybil detection system leveraging a deep learning model designed for precise prediction of Sybil attacks. The proposed model is composed of three integral modules: a data harvesting module, a feature extracting module, and a deep regression model. These modules operate in a cohesive manner to systematically analyze user profiles on Twitter, enhancing the accuracy of Sybil attack detection. Similarly, Rahman et al. put forth a model named Sybil Trap, presenting a graph- based semi-supervised

learning system. This model incorporates a comprehensive approach, utilizing both content-based and structure-based techniques for the detection of Sybil attacks. By integrating information from the content of user interactions and the underlying network structure, Sybil Trap aims to provide a robust defense mechanism against Sybil attacks.

These advancements in Sybil detection models highlight the significance of integrating sophisticated techniques, such as deep learning and graph-based semi-supervised learning, to bolster the efficacy and accuracy of identifying and mitigating Sybil attacks in online social networks.

Methodology

This study employs a multi-faceted approach to gain theoretical insights into user security and privacy in social media environments. The methodology encompasses exploratory design, qualitative methods, secondary data collection, and theoretical frameworks.

The research adopted an exploratory design to delve into the complex and multifaceted nature of user security and privacy on social media platforms. This design allows for the exploration of diverse perspectives, experiences, and factors influencing security and privacy outcomes in online environments. A qualitative approach was employed to capture rich, nuanced data that elucidates the subjective experiences, perceptions, and behaviors of social media users regarding security and privacy. Qualitative methods such as interviews, focus groups, and participant observations facilitated in-depth exploration and interpretation of user perspectives. Secondary data collection involves gathering existing literature, research studies, reports, and empirical findings related to user security and privacy in social media environments. This comprehensive review of secondary sources provided a foundational understanding of key concepts, theories, and empirical evidence informing the study. The study utilized theoretical frameworks from various disciplines such as sociology, psychology, communication studies, and computer science to guide the analysis and interpretation of findings. These theoretical frameworks offered conceptual lenses through which to examine the socio-cultural, psychological, technological, and institutional factors shaping user security and privacy in social media environments.

By integrating exploratory design, qualitative methods, secondary data collection, and theoretical frameworks, this methodology enables a holistic and nuanced exploration of user security and privacy in social media environments, ultimately contributing to theoretical insights and knowledge advancement in this critical area.

Analysis and Interpretation

1. User Perceptions on Social Media Security and Privacy

Social media platforms have become integral parts of daily life, facilitating communication, networking, and information sharing. However, concerns regarding user security and privacy persist, stemming from various factors such as data breaches, cyber threats, and the proliferation of misinformation. This report aims to analyze different user perceptions regarding these security and privacy concerns related to social media.

- **Concerns about Data Privacy:** Many users expressed concerns about the privacy of their personal data on social media platforms. They were worried about how their information was being collected, stored, and used by platform providers and third-party entities for targeted advertising and data analytics purposes. Users expressed a desire for more transparency and control over their data.
- **Fear of Cyber Threats:** Users highlighted fears of cyber threats such as hacking, identity theft, and phishing scams while using social media. They were concerned about the security of their accounts and the potential for unauthorized access to their personal information. Some users reported experiencing security incidents firsthand, which heightened their awareness of online risks.
- **Impact of Misinformation:** Many users were concerned about the proliferation of misinformation and fake news on social media platforms. They expressed worries about the spread of false information, conspiracy theories, and propaganda, which could manipulate public opinion and undermine trust in reliable sources of information. Users emphasized the need for platforms to address these issues and promote media literacy.
- **Privacy Settings and Control:** Users highlighted the importance of privacy settings and control features offered by social media platforms. They appreciated the ability to customize their privacy preferences, manage visibility settings, and control who can access their content and personal information. However, some users found these settings confusing or difficult to navigate, leading to concerns about inadvertent exposure of private data.
- **Trust in Platform Providers:** Many users expressed skepticism about the trustworthiness of social media platform providers in safeguarding their security and privacy. They were concerned about past incidents of data breaches, privacy violations, and controversies surrounding platform policies and practices. Users called for greater accountability, transparency, and ethical standards from platform operators.

  User perceptions on security and privacy concerns in social media reflect a range of anxieties, fears, and expectations regarding data privacy, cyber threats, misinformation, and platform accountability. Addressing these concerns requires collaborative efforts from platform providers, policymakers, and users

themselves to enhance transparency, promote security best practices, and foster a culture of responsible digital citizenship in the online ecosystem.

2. Social Media Platform Features, Policies, and Algorithms in Shaping User Security and Privacy Outcomes

Social media platforms play a pivotal role in shaping user security and privacy outcomes through their features, policies, and algorithms. Understanding how these elements interact can provide valuable insights into the mechanisms that influence user experiences and perceptions of security and privacy. This analysis explores the impact of platform features, policies, and algorithms on user security and privacy outcomes.

- **Platform Features:** Social media platforms offer a wide array of features that have implications for user security and privacy. These features include privacy settings, content moderation tools, encryption protocols, and authentication mechanisms. Privacy settings allow users to control who can access their content and personal information, providing a level of autonomy over their privacy preferences. Content moderation tools enable platforms to identify and remove harmful or inappropriate content, enhancing user safety. Encryption protocols ensure the confidentiality of user communications, mitigating the risk of unauthorized access to sensitive information. Authentication mechanisms verify the identity of users, preventing account hijacking and unauthorized access. By leveraging these features, social media platforms can empower users to manage their security and privacy effectively.
- **Platform Policies:** The policies implemented by social media platforms play a crucial role in shaping user security and privacy outcomes.
  These policies encompass terms of service, community guidelines, data handling practices, and transparency measures. Terms of service outline the contractual agreements between platforms and users, establishing the rules and responsibilities governing their interactions. Community guidelines define acceptable behavior and content standards, guiding users on appropriate conduct within the platform. Data handling practices dictate how user data is collected, stored, and used by platforms, influencing user trust and confidence in their privacy protection measures. Transparency measures provide users with insight into platform policies and practices, fostering accountability and trust. By enforcing clear and consistent policies, social media platforms can promote user security and privacy while maintaining a conducive online environment.
- **Platform Algorithms:** Algorithms deployed by social media platforms play a significant role in shaping user experiences and security outcomes. These algorithms govern content curation, recommendation systems, and advertising targeting, influencing the visibility and dissemination of user-generated content. Content curation algorithms determine the content displayed on users' feeds

based on relevance, engagement, and user preferences, impacting the exposure of users to diverse perspectives and potential security risks. Recommendation systems suggest relevant content, connections, and groups to users, influencing their interactions and information consumption habits. Advertising targeting algorithms analyze user data to deliver personalized ads, raising concerns about privacy infringement and data exploitation. By understanding and regulating the algorithms used by social media platforms, stakeholders can mitigate the risk of algorithmic bias, manipulation, and unintended consequences on user security and privacy.

Social media platform features, policies, and algorithms play integral roles in shaping user security and privacy outcomes. By leveraging effective features, enforcing transparent policies, and regulating algorithms, platforms can promote user empowerment, enhance accountability, and mitigate security and privacy risks. Collaborative efforts from platform providers, policymakers, and users are essential to foster a safer and more privacy-respecting online environment.

3. Emerging Threats, Vulnerabilities, and Risks to User Security and Privacy in Social Media Environments

As social media platforms continue to evolve and expand, they face a myriad of emerging threats, vulnerabilities, and risks that pose significant challenges to user security and privacy. Understanding these evolving threats is crucial for developing effective strategies to mitigate risks and protect users in the dynamic online landscape. This analysis explores the key emerging threats, vulnerabilities, and risks to user security and privacy in social media environments.

- **Cyber Threats:** One of the most pressing concerns for users on social media platforms is the risk of cyber threats such as hacking, malware, phishing, and account takeover. Cybercriminals exploit vulnerabilities in platform security mechanisms and user behaviors to gain unauthorized access to accounts, steal sensitive information, or distribute malicious content. As social media platforms become lucrative targets for cyber- attacks, users face increased risks of identity theft, financial fraud, and privacy breaches.
- **Misinformation and Disinformation:** The proliferation of misinformation and disinformation on social media poses significant risks to user security and privacy. Malicious actors exploit the viral nature of social media platforms to spread false or misleading information, manipulate public opinion, and undermine trust in reliable sources. Users are susceptible to misinformation campaigns, conspiracy theories, and propaganda, which can have far- reaching consequences on their beliefs, behaviors, and decision-making processes.
- **Data Privacy Concerns:** Privacy concerns remain a persistent challenge for users on social media platforms, fueled by the collection, storage, and sharing

of personal data by platform providers and third-party entities. Users' personal information is often harvested for targeted advertising, data analytics, and algorithmic profiling purposes, raising concerns about privacy infringement, data exploitation, and surveillance. The lack of transparency and control over data practices exacerbates user distrust and anxiety about their privacy on social media.

- **Cyberbullying and Online Harassment:** Cyberbullying and online harassment continue to be prevalent threats to user security and privacy in social media environments. Users, particularly vulnerable groups such as children, teenagers, and marginalized communities, are susceptible to harassment, cyberbullying, and online abuse from peers, strangers, or malicious actors. The anonymity and reach of social media platforms amplify the impact of cyberbullying, causing psychological distress, social isolation, and reputational harm for affected users.
- **Deepfake and Synthetic Media:** The emergence of deepfake technology and synthetic media poses new challenges to user security and privacy on social media platforms. Deepfake videos and manipulated content can deceive users, undermine trust in digital content, and facilitate identity fraud or defamation. As deepfake technology becomes more sophisticated and accessible, users face heightened risks of misinformation, social engineering attacks, and privacy violations on social media.

  The evolving landscape of social media presents a multitude of emerging threats, vulnerabilities, and risks to user security and privacy. Cyber threats, misinformation, data privacy concerns, cyberbullying, and deepfake technology are among the key challenges facing users in social media environments. Addressing these emerging threats requires proactive measures from platform providers, policymakers, and users to enhance security measures, promote digital literacy, and foster a safer and more resilient online ecosystem. Collaborative efforts are essential to mitigate risks and protect user rights in the dynamic and evolving landscape of social media.

4. Consolidated Statements from Different Social Media Users Regarding Privacy and Security Threats

“As a parent, I worry about my child's safety on social media. With cyberbullying, predators, and inappropriate content circulating, it's challenging to ensure their privacy and security. I constantly monitor their accounts, but I wish there were better safeguards in place to protect young users." - Concerned Parent

"As a freelancer, I rely on social media for networking and promoting my work. However, I'm wary of data breaches and hacking incidents that could compromise my clients' information or intellectual property. I take precautions like using strong passwords and two-factor authentication, but the constant threat of cyber-attacks is unsettling." - Freelancer

"Privacy is a fundamental right that's increasingly under threat in the digital age. Social media platforms collect vast amounts of personal data, often without clear consent or transparency. As users, we need to be vigilant about protecting our privacy and advocating for stronger regulations to hold platforms accountable for data misuse." - Privacy Advocate

"Running a small business, social media is crucial for reaching customers and driving sales. However, I'm concerned about the security of our company's social media accounts. Phishing scams, account takeovers, and fake reviews pose significant risks to our reputation and financial security. We invest in cybersecurity measures, but the evolving nature of threats keeps us on edge." - Small Business Owner

"Social media has become a powerful tool for organizing grassroots movements and raising awareness about social issues. However, it's also a breeding ground for misinformation and targeted harassment. As activists, we face threats to our safety and privacy from online trolls, doxxing, and surveillance. It's a constant battle to protect ourselves while amplifying our voices for positive change." -Political Activist

"Growing up in the digital age, I'm aware of the privacy risks associated with social media. I'm cautious about what I share online and regularly review my privacy settings. Still, I'm concerned about data tracking, algorithmic manipulation, and the potential for my information to be exploited by advertisers or malicious actors. Staying informed and proactive is essential for safeguarding my online identity." - Tech-Savvy Teenager

The statements from various social media users illustrate a spectrum of concerns regarding privacy and security threats. Parents express worry about their children's safety amidst cyberbullying and predatory behavior, emphasizing the need for enhanced safeguards. Freelancers and small business owners highlight the constant threat of data breaches and hacking incidents, impacting their clients' information and financial security. Privacy advocates stress the importance of transparency and user vigilance in protecting personal data from exploitation by social media platforms. Political activists face risks such as harassment and surveillance while utilizing social media for advocacy, underscoring the dual nature of these platforms as tools for both empowerment and vulnerability. Even tech-savvy teenagers acknowledge the pervasive dangers of data tracking and algorithmic manipulation, emphasizing the importance of proactive measures to safeguard their online identities. Together, these statements emphasize the multifaceted nature of privacy and security threats on social media, urging for collective action to mitigate risks and protect user rights effectively.

Conclusions and Recommendations

Conclusion

Theoretical Insights into User Security and Privacy in Social Media Environments has provided a comprehensive understanding of the complex dynamics shaping user experiences in online spaces. Through the exploration of theoretical frameworks and empirical research, this study has illuminated the multifaceted nature of security and privacy concerns on social media platforms. The analysis has revealed the interplay between sociocultural factors, platform features, and algorithmic processes in influencing user perceptions, behaviors, and outcomes. Moreover, the examination of emerging threats and vulnerabilities has underscored the urgency of addressing systemic challenges to enhance user protection and promote a safer online ecosystem.

Recommendations

Government, Social Media Players and other related Stakeholders are recommended to

- Develop educational initiatives to enhance user awareness and digital literacy regarding security and privacy risks on social media platforms. Empowering users with knowledge and skills to navigate online environments safely is essential for mitigating risks and fostering responsible digital citizenship.
- Advocate for greater transparency, accountability, and ethical standards from social media platforms regarding data handling practices, content moderation, and algorithmic governance. Establishing clear policies and mechanisms for user recourse can enhance trust and confidence in platform providers.
- Foster interdisciplinary collaboration among researchers, policymakers, industry stakeholders, and civil society organizations to address the complex challenges of user security and privacy in social media environments. By leveraging diverse expertise and perspectives, stakeholders can develop holistic solutions that balance technological innovation with user rights and well-being.
- Advocate for robust regulatory frameworks that prioritize user protection, data privacy, and digital rights in the context of social media platforms. Strengthening regulatory oversight and enforcement mechanisms can incentivize platform accountability and promote responsible data practices.
- Support ongoing research and evaluation efforts to monitor evolving trends, emerging threats, and user behaviors in social media environments. By staying abreast of developments and conducting regular assessments, stakeholders can adapt strategies and interventions to effectively address evolving security and privacy challenges.
- In conclusion, advancing theoretical insights into user security and privacy in social media environments requires a concerted effort from multiple stakeholders to foster a culture of safety, trust, and accountability online. By implementing the aforementioned recommendations, we can work towards creating a more secure and privacy- respecting digital landscape for all users.

References

1. Barnes, S. B. (2006). A privacy paradox: Social networking in the United States. First Monday, 11(9).
2. Bailey, M., et al. (2009). The security challenges of social networks. IEEE Security & Privacy, 7(6), 88–92.
3. Chaffey, D. (2019). Digital Business and E-Commerce Management. Pearson Education.
4. Hogben, G. (2007). Security Issues and Recommendations for Online Social Networks. European Network and Information Security Agency (ENISA).
5. Kumar, S., & Somani, V. (2018). Social Media Security Risks, Cyber Threats and Mitigation Techniques. International Journal of Computer Applications.
6. McCarroll, N., & Curran, K. (2013). Social networking in education. International Journal of Innovation in the Digital Economy.
7. Ortiz-Ospina, E. (2019). The rise of social media. Our World in Data.
8. Rathore, S., et al. (2018). Social network security: Issues, challenges, threats, and solutions. Information Sciences, 421, 43–69.

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 26 - 39

Cyber Fraud in Aadhaar Enabled Payment System (AePS): Threat Landscape, Vulnerability Analysis, and Mitigation Strategies

Masud Murad Khan

Assistant Professor, BCA Department M.L.K. (P.G.) College, Balrampur

Email:

Article DOI Link: https://zenodo.org/uploads/22028942

DOI: 10.5281/zenodo.22028942

Abstract

The Aadhaar Enabled Payment System (AePS) has been a transformative force in India's financial inclusion mission, enabling millions of unbanked and underbanked citizens to access banking services through biometric authentication. However, this convenience has created a rapidly expanding attack surface for cybercriminals. According to the Indian Cyber Crime Coordination Centre (I4C), AePS-related fraud accounted for 11% of all online financial frauds in India during 2023 [1]. In 2024, cyber fraud losses across India surged to ₹22,845 crore — a staggering 206% increase from 2023 [14]. This paper presents a comprehensive investigation into the threat taxonomy of AePS fraud, examining biometric cloning, data breaches, malicious Business Correspondent (BC) agent activities, and social engineering attacks. The study evaluates the regulatory responses of the Reserve Bank of India (RBI), NPCI, and UIDAI, and proposes a multi-layered prevention framework combining technical, behavioural, and policy-level interventions.

Keywords: AePS, Aadhaar, Cyber Fraud, Biometric Spoofing, Fingerprint Cloning, NPCI, RBI, UIDAI, Digital Payments, Financial Inclusion, Liveness Detection, Business Correspondent.

Introduction

Background

India's digital payment revolution, driven by the Jan Dhan-Aadhaar-Mobile (JAM) trinity, has created one of the world's most ambitious financial inclusion frameworks. At its core, the Aadhaar Enabled Payment System (AePS) — operated by the National Payments Corporation of India (NPCI) allows citizens to conduct

banking transactions using only their 12-digit Aadhaar number and a fingerprint or iris scan at any authorised Business Correspondent (BC) agent terminal [17].

As of 2024, AePS processes an estimated 600 million transactions per month, disbursing government welfare benefits under schemes such as PM-KISAN, MGNREGA, and PMGKY to hundreds of millions of beneficiaries [1]. This scale, while remarkable, simultaneously creates an enormous and high-value target for fraudsters who have developed sophisticated techniques to exploit the system's core authentication mechanism: biometrics.

Research Objectives

This paper aims to: (a) systematically categorise and analyse the types of fraud perpetrated through AePS; (b) identify the technical, structural, and social root causes enabling such fraud; (c) quantify the scale and demographic impact of AePS-related financial crimes; (d) evaluate the effectiveness of existing regulatory and technical countermeasures; and (e) propose a comprehensive, multi-layered mitigation framework.

Scope and Limitations

This research synthesises findings from government reports (RBI, MHA, UIDAI, I4C), parliamentary panel documents, judicial records, and reputable media investigations covering the period 2021–2025. The study focuses exclusively on the Indian context. Limitations include the under-reporting of AePS fraud cases in rural areas and the absence of a centralised, publicly available AePS-specific fraud database.

Overview of the AePS Ecosystem

Architecture and Stakeholders

AePS operates within a multi-stakeholder ecosystem regulated by three principal bodies: the Reserve Bank of India (RBI) as the overarching financial regulator; the National Payments Corporation of India (NPCI) as the system operator; and the Unique Identification Authority of India (UIDAI), which manages Aadhaar authentication [17][18]. Additional participants include Acquiring Banks, Issuing Banks, and Business Correspondent (BC) agents who serve as the last-mile touchpoint for customers in remote and rural areas.

AePS Transaction Flow

An AePS transaction proceeds through five sequential stages:

- Customer Identification — the customer provides their 12-digit Aadhaar number to the BC agent and selects the transaction type;
- Biometric Capture — the fingerprint template is captured and encrypted at the terminal;

- UIDAI Authentication — encrypted biometric data is transmitted to UIDAI servers via NPCI for matching, returning a Yes/No response;
- Bank Authorisation — NPCI routes the transaction to the issuing bank upon successful authentication; and (5) Transaction Completion — funds are transferred and an SMS confirmation is dispatched to the Aadhaar-linked mobile number.

Literature Review

Prior research by Frugal Testing (2024) identified inadequate data security practices among third-party AePS operators as a primary vulnerability [10]. The Drishti IAS analysis (2023) documented specific gaps — including ambiguity around whether AePS is an opt-in or opt-out service — that cybercriminals systematically exploit [11]. The Parliamentary Standing Committee on Finance (2024) issued a landmark report identifying AePS fraud, KYC fraud, and Android malware as the three most prevalent forms of digital payment crime [8]. The Biometric Update (2024) reported that the Ministry of Home Affairs (MHA) flagged a dangerous upward trend in cloned fingerprint-based AePS attacks [3][13].

Collectively, existing literature reveals a systemic convergence of weak biometric device standards, inadequate agent oversight, and low digital literacy among the target demographic, creating conditions highly favourable for organised criminal exploitation. No prior study has comprehensively integrated the technical, demographic, regulatory, and policy dimensions of AePS fraud in a single analytical framework — this gap motivates the present work.

Research Methodology

This study employs a qualitative and quantitative mixed-method approach, relying on secondary data analysis. Primary data sources include: official reports such as RBI Annual Reports (2022–2025) and I4C/NCRP fraud statistics [1][4][5][6]; legislative records from Parliamentary Standing Committee reports and MHA submissions to Parliament [7][8]; regulatory circulars, including RBI directives on AePS from February 2024, August 2024, and June 2025 [4][5][6]; investigative journalism from The Hindu, Reuters, Times of India, and Biometric Update [2][3][12][14][15]; and technical cybersecurity literature on biometric spoofing and liveness detection [10].

Thematic analysis is employed to categorise fraud typologies into six primary classes. Statistical data is presented in tabular and graphical form to illustrate temporal trends and geographic distribution. Where primary data are unavailable, triangulation across multiple secondary sources is used to validate estimates.

Statistical Analysis — Scale of AePS Fraud

The scale of AePS fraud has grown in direct proportion to the system's transaction volume and geographic reach. According to I4C, AePS-related fraud accounted for

11% of all online financial fraud in India in 2023 [1][2]. By 2024, MHA's submission to Parliament confirmed total cyber fraud losses of ₹22,845 crore, representing a 206% year-on-year increase [14][15]. The Ministry explicitly acknowledged before Parliament that dummy or rubber fingers are being used to falsify biometric authentication using Aadhaar, confirming that hardware-level spoofing had become mainstream [7].

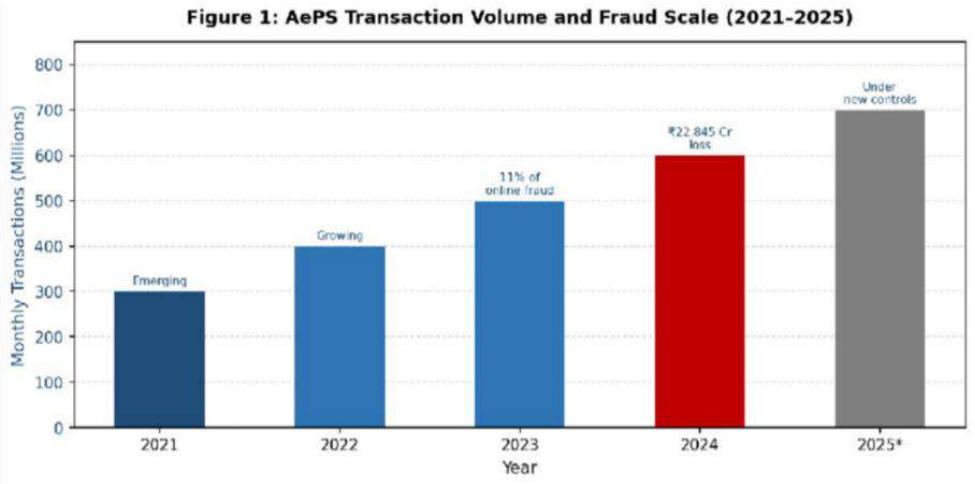


Figure 1: AePS Transaction Volume and Fraud Scale (2021–2025). Sources: I4C [1], MHA [7], RBI [4][5][6], Times of India [14].

Table 1 summarises the fraud growth trajectory. Independent researchers estimate that actual AePS fraud is 3–5 times higher than reported on the National Cyber Crime Reporting Portal (NCRP), owing to significant under-reporting in rural areas where victims lack awareness of formal grievance channels [11].

Table 1: AePS Fraud Growth Trend (2021–2025)

Year	Txns/Month	Fraud Scale	Regulatory Milestone	Key Fraud Method
2021	~300 million	Emerging	Limited awareness	BC Agent misuse
2022	~400 million	Growing complaints	NPCI issues advisories	Data leaks + agent fraud
2023	~500 million	11% of online fraud	Parliamentary discussion begins	Fingerprint cloning emerges
2024	~600 million	₹22,845 Cr total loss	RBI Feb + Aug directives	Rubber/silicone dummy fingers
2025	700+ million	Under new controls	RBI June 2025 circular	Advanced spoofing + SIM swap

Sources: I4C [1], MHA [7], RBI [4][5][6], NPCI [17].

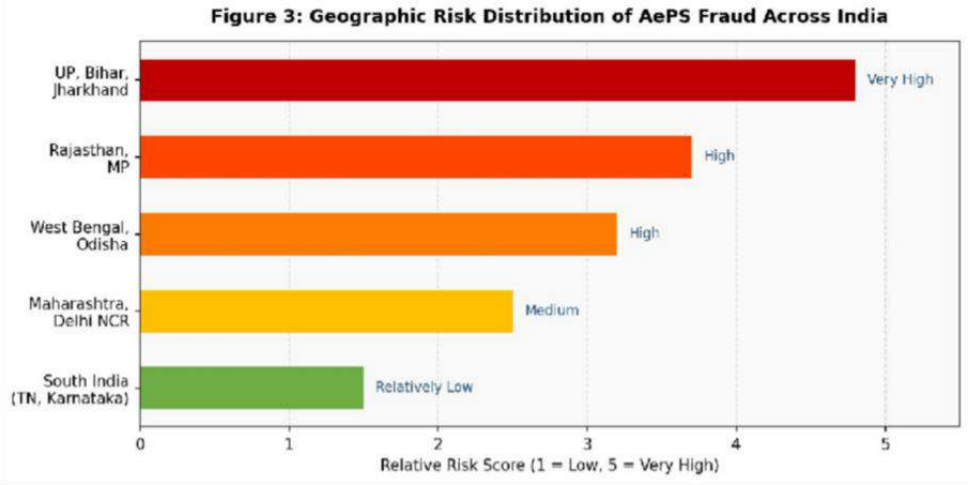


Figure 3: Geographic Risk Distribution of AePS Fraud Across India. Sources: I4C [1], NCRP data, Drishti IAS [11].

Taxonomy of AePS Fraud Types

Based on the synthesis of government reports, judicial records, and cybersecurity research, this study identifies six primary categories of AePS fraud. Figure 2 below presents their estimated proportional distribution.

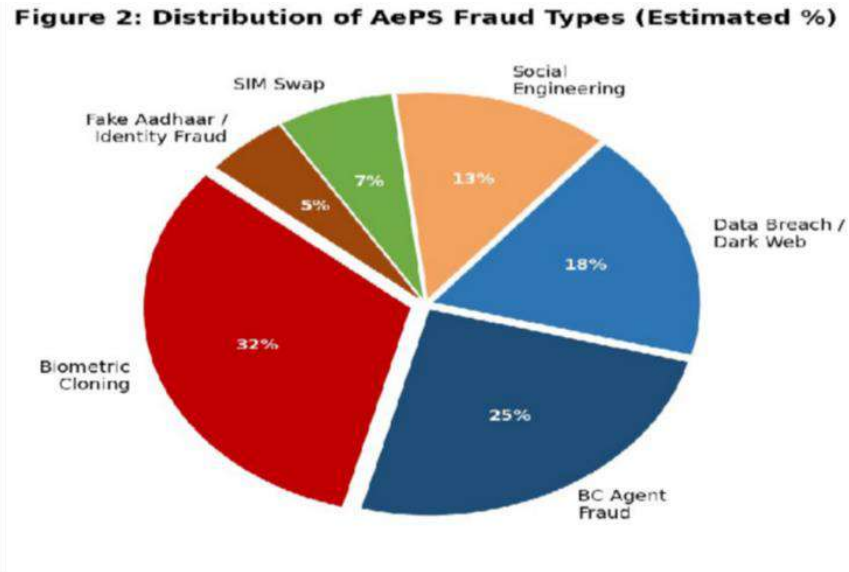


Figure 2: Distribution of AePS Fraud Types (Estimated %). Sources: I4C [1], Parliamentary Committee [8], MHA [7].

1. Biometric Cloning and Fingerprint Spoofing

This represents the most technically sophisticated and destructive form of AePS fraud. Perpetrators obtain fingerprint impressions from publicly available sources — property registration documents, MGNREGA attendance sheets, government ID

application forms, or high-resolution photographs. These impressions are used to create artificial fingers using silicone, rubber, or gelatin that can fool standard biometric scanners lacking liveness detection capability [3][7]. Primary targets include Jan Dhan account holders receiving government subsidies, elderly citizens in remote areas, and migrant workers whose accounts are accessed without their knowledge.

2. Data Breach and Dark Web Exploitation

Biometric and Aadhaar data stored insecurely by third-party operators — including enrolment agencies, insurance companies, microfinance institutions, and state government portals — has been compromised in multiple documented incidents [10]. This data is subsequently traded on dark web marketplaces, enabling large-scale coordinated fraud. UIDAI's 2023 advisory noted that agents or third-party operators may store Aadhaar and bank data without proper protection, exposing users to data privacy risks [18].

3. Malicious Business Correspondent (BC) Agent Fraud

The BC agent model — critical for last-mile delivery — has become an internal threat vector. Fraudulent or compromised agents exploit their privileged position to capture biometric data during legitimate transactions for later misuse, process unauthorised withdrawals from customer accounts, collude with organised crime networks to drain Jan Dhan accounts during subsidy disbursement cycles, and operate as unregistered agents without proper bank authorisation [8][9].

4. Social Engineering Attacks

Fraudsters impersonate bank officials, UIDAI representatives, or government functionaries via phone calls (vishing) and text messages (smishing) to extract Aadhaar numbers, OTPs, and account details. Common pretexts include KYC update requirements, Aadhaar card expiry warnings, benefit transfer failures, and account blocking threats. These attacks are particularly effective against elderly and low-literacy populations [11].

5. SIM Swap and Mobile-Based Fraud

By obtaining a duplicate SIM card for the victim's mobile number through fraudulent documentation submitted to telecom operators, criminals redirect OTPs and bank alerts to their own device. This enables them to authorise AePS transactions even where OTP-based secondary authentication is in place, representing a serious threat to the post-2025 dual-factor framework [6][15].

6. Fake Aadhaar and Identity Fraud

Using forged Aadhaar documents and stolen personal information, fraudsters open fraudulent bank accounts under false identities. These mule accounts are used to

receive and launder proceeds from AePS fraud, making transaction tracing significantly more difficult for law enforcement agencies [8].

Root Cause Analysis

1. Technical Root Causes

Table 2 presents the principal technical vulnerabilities identified in the AePS framework. The most critical is the absence of liveness detection: most deployed AePS scanners cannot distinguish live fingers from synthetic replicas, directly enabling fingerprint spoofing attacks [3][7]. Single-factor authentication — where biometrics alone were sufficient for any transaction amount prior to 2025 — allowed full account drainage with a single compromised fingerprint [6]. Insecure data storage by third-party enrolment agencies and unrestricted inter-bank interoperability further compounds the risk surface [10][17].

Table 2: Technical Vulnerabilities in the AePS Framework

Vulnerability	Description	Fraud Enabled
Absence of Liveness Detection	Most AePS scanners cannot distinguish live fingers from synthetic replicas	Fingerprint spoofing, dummy finger attacks
Single-Factor Authentication	Until 2025, biometrics alone were sufficient for any transaction amount	Full account drains with one compromised fingerprint
Insecure Third-Party Data Storage	Enrolment agencies and BC agents store biometric data without encryption	Data breach → dark web sale → large-scale fraud
No Anomaly Real-Time Detection	Absence of behavioural analytics to flag unusual transaction patterns	Multiple withdrawals before rapid victim is notified
Default AePS Enablement Ambiguity	Neither UIDAI nor NPCI clearly communicated opt-in/opt-out status to users	Victims unaware their accounts are AePS-accessible

Sources: UIDAI [18], NPCI [17], Frugal Testing [10], Drishti IAS [11].

2. Socio-Structural Root Causes

Low digital literacy is the most pervasive socio-structural vulnerability: the majority of AePS users have minimal understanding of biometric authentication risks or available security features such as Biometric Lock [11]. Inadequate BC agent vetting — background verification processes

remain inconsistent across banks and geographies — creates an internal threat channel [8]. Mobile number linkage gaps mean many Aadhaar records carry outdated contact details, preventing fraud alert SMSs from reaching victims. Complex grievance mechanisms further discourage rural victims from reporting, leading to significant under-reporting of actual losses [9].

3. Regulatory Root Causes

Three regulatory failures compounded the fraud problem. First, delayed regulatory action: AePS fraud escalated significantly in 2021–2023 before comprehensive RBI guidelines were issued in 2024–2025 [4][5]. Second, fragmented oversight: responsibility is split across RBI (payment regulation), NPCI (system operation), UIDAI (identity infrastructure), and state police (law enforcement), creating coordination gaps that organised crime networks exploit [8]. Third, the absence of mandatory device standards: no mandatory certification for liveness detection capability in AePS terminals was enforced until 2025 [6].

Vulnerable Demographics and Geographic Analysis

AePS fraud disproportionately impacts four demographic groups. Senior citizens represent the most vulnerable group, characterised by low digital literacy, high trust in BC agents, and physical inability to monitor accounts independently. Rural women receiving PM-KISAN and MGNREGA payments are particularly exposed because Aadhaar is often linked to a husband's mobile number, meaning SMS fraud alerts never reach the account holder directly. Jan Dhan beneficiaries — over 800 million accounts receiving predictable, periodic government transfers — are targeted because disbursement cycles create known windows of high-value deposits [8]. Migrant workers are entirely dependent on BC agents and rarely monitor accounts in real time, meaning fraud may go undetected for weeks or months.

Geographically, Uttar Pradesh, Bihar, and Jharkhand carry the highest risk, owing to low digital literacy, dense rural populations, and high BC dependency. Rajasthan and Madhya Pradesh are rated high risk, while West Bengal and Odisha — with large Jan Dhan account bases and low smartphone penetration — are also high risk. Maharashtra and Delhi NCR face a distinct threat from organised cybercrime syndicates and fake BC operations. South India (Tamil Nadu, Karnataka) exhibits relatively lower risk, attributable to higher digital literacy and stronger alternative payment channels [1][11].

Regulatory and Government Response

1. Reserve Bank of India (RBI) Interventions

The RBI has issued three escalating rounds of guidance on AePS security (Table 3). The February 2024 Statement on Developmental and Regulatory Policies announced intent to strengthen the AePS security framework and directed NPCI and acquiring banks to enhance fraud risk management [4]. The August 2024 guidelines

mandated fraud monitoring systems and enhanced BC agent oversight [5]. Most significantly, the June 2025 Master Circular — effective January 1, 2026 requires biometric plus OTP dual authentication for transactions above ₹5,000, mandatory liveness detection on all new AePS devices, and real-time transaction monitoring [6].

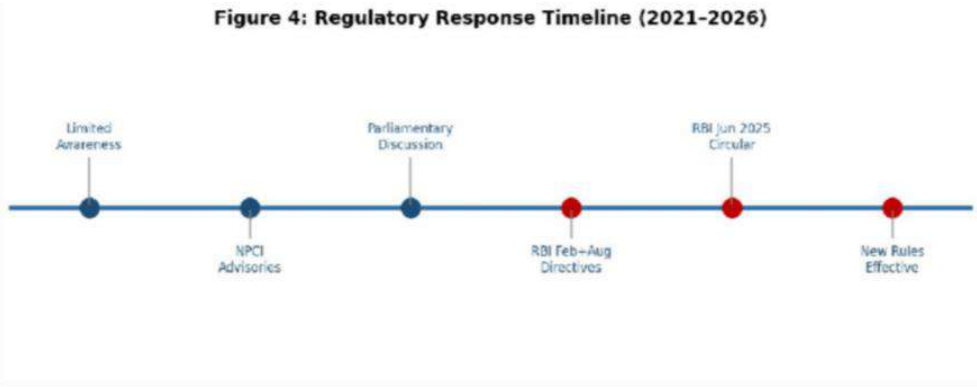


Figure 4: Regulatory Response Timeline for AePS Fraud (2021–2026). Sources: RBI [4][5][6], PRS India [8].

2. UIDAI Security Measures

UIDAI offers four user-level security features for AePS users [18]: the Biometric Lock, accessible via my.uidai.gov.in or the mAadhaar app, temporarily disables all biometric authentication; the Virtual ID (VID) provides a 16-digit temporary identifier that conceals the actual Aadhaar number; the Time-based OTP (TOTP) generates a 30-second offline password through the mAadhaar app; and mAadhaar App Alerts deliver real-time notifications for every Aadhaar authentication event.

3. I4C and Law Enforcement Response

The Indian Cyber Crime Coordination Centre (I4C) has designated AePS fraud as a Priority Crime Category [1]. The Citizen Financial Cyber Frauds Reporting and Management System (CFCFRMS) enables rapid fund freeze requests, reportedly recovering over ₹1,200 crore in 2024. The 1930 Helpline provides a dedicated national cybercrime reporting number, though coverage and multilingual capacity remains uneven. The NCRP portal (cybercrime.gov.in) has been enhanced to allow victims to report AePS fraud with transaction-level details [9].

4. Parliamentary Committee Recommendations

The Parliamentary Standing Committee on Finance (2024) issued six key recommendations [8]: mandatory STQC-certified liveness detection on all AePS devices; mandatory background police verification for all BC agents prior to authorisation; a 24×7 multilingual 1930 helpline with guaranteed same-day fraud hold capability; establishment of a victim compensation fund; integration of digital financial literacy into school curriculum and MGNREGA worker training; and

creation of a single national AePS fraud database accessible to all investigating agencies.

Prevention Framework and Technical Solutions

1. Technical Layer — Device and Authentication Hardening

Liveness detection is the critical missing link in the current AePS security architecture. It ensures that the biometric sample presented belongs to a living individual rather than an artificial replica. Key techniques include Pulse Oximetry Sensing — detecting blood flow beneath the skin surface, which silicone replicas cannot replicate; 3D Depth Sensing, which maps the three-dimensional topology of a fingerprint to distinguish it from 2D molded fakes; AI-powered Skin Texture Analysis, examining micro-texture patterns absent in synthetic materials; and Sweat Pore Detection, as active sweat pores in live fingers produce a pattern absent in artificial fingers [13].

Under the new RBI 2026 framework, the single-factor biometric-only system is replaced by dual-factor authentication requiring biometrics plus OTP for transactions above ₹5,000, supplemented by mandatory hardware liveness detection and real-time behavioural analytics and risk scoring [6].

2. AI-Powered Fraud Detection

Three machine learning-based mechanisms strengthen real-time fraud prevention [20]: Behavioural Analytics models flag transactions deviating from a customer's historical pattern in amount, time, frequency, or location; Device Fingerprinting assigns each AePS terminal a cryptographic identity so transactions from unregistered or compromised devices are automatically blocked; and Geo-Velocity Checks flag transactions occurring from geographically improbable locations within a short time window. Every transaction receives a dynamic risk score; high-risk transactions are held pending manual review or additional authentication.

3. BC Agent Management Solutions

Hardware Security Modules (HSMs) embedded in AePS terminals prevent biometric data extraction at the device level. End-to-end encryption from fingerprint capture to UIDAI authentication server ensures data cannot be intercepted in transit. Tamper-evident audit trails for every BC agent transaction are transmitted in real-time to the acquiring bank. Only bank-registered, STQC-certified terminals are permitted for AePS transactions under the new framework [6][17].

4. User-Level Preventive Measures

The most critical user actions are: enabling Biometric Lock via my.uidai.gov.in when AePS is not in active use; ensuring the Aadhaar-linked mobile number is current and belongs directly to the account holder; installing the mAadhaar app to enable transaction alerts; using Virtual ID (VID) in place of the actual Aadhaar

number wherever possible; monitoring bank accounts monthly; verifying BC agent bank-issued ID cards before transactions; and reporting suspected fraud immediately on the 1930 helpline or at cybercrime.gov.in [18][20].

Discussion

AePS exemplifies what this study terms the inclusion-security paradox of financial technology deployment in developing economies. The system's deliberate design simplicity — requiring only an Aadhaar number and a fingerprint — was the key enabler of its unprecedented financial inclusion impact. Yet this same simplicity eliminated multiple security layers that more complex payment systems rely upon: PINs, passwords, SMS OTPs, and hardware tokens. The consequence is a system where a single compromised data point — the fingerprint — grants complete, irrevocable access to an individual's entire bank balance. For a rural woman receiving PM-KISAN benefits, or a migrant worker with months of savings, this vulnerability represents a catastrophic financial loss with no immediate recourse.

A particularly alarming trend is the commoditisation of biometric fraud tools. Silicone finger molds, once the domain of sophisticated state-level actors, are now commercially available at nominal cost. Dark web markets sell Aadhaar-linked biometric data in bulk. Organised crime networks operate dedicated AePS fraud teams with specialised roles: data procurers, mold fabricators, terminal operators, and money mule managers [3][7]. This industrialisation makes piecemeal countermeasures insufficient.

The RBI's June 2025 circular represents a significant and overdue regulatory advancement. However, its effectiveness depends critically on: implementation speed and quality across hundreds of thousands of rural AePS terminals; BC agent compliance in low-supervision environments; fraudster adaptation — particularly OTP channel attacks via SIM swap; and closing the awareness gap among the most vulnerable users [6][15].

India is not alone in grappling with biometric payment fraud. Brazil's Pix system has experienced social engineering-based fraud at scale; Kenya's M-Pesa ecosystem has encountered SIM swap fraud extensively; Bangladesh's mobile banking landscape mirrors India's BC agent misuse patterns [19]. However, India's AePS is unique in combining purely biometric authentication without OTP fallback, an extreme scale of vulnerable beneficiary populations, and a predictable government benefit disbursement cycle that creates known windows of high-value targets — making AePS fraud a uniquely high-impact problem demanding India-specific solutions.

Conclusion and Recommendations

AePS fraud has evolved from isolated incidents of agent misconduct into a structured, industrialised criminal ecosystem targeting the most economically vulnerable citizens of India. The 206% surge in cyber fraud losses between 2023

and 2024, with AePS comprising 11% of all online financial fraud, represents not merely a cybersecurity failure but a direct assault on India's financial inclusion mission [1][14][15].

The core vulnerability — irrevocable single-factor biometric authentication — was a deliberate design choice to prioritise accessibility. Remedying it without undermining the accessibility imperative requires nuanced, tiered solutions. The RBI's dual-authentication mandate for high-value transactions, combined with mandatory liveness detection, represents a calibrated and effective response — provided implementation is rigorous and rapid [6]. Ultimately, the security of AePS is a shared responsibility spanning regulators, financial institutions, technology providers, BC agent networks, civil society organisations, and end users.

Table 3: Prioritised Policy Recommendation Matrix

#	**Recommendation**	**Responsible Body**	**Priority**	**Timeline**
R1	Expedite STQC-certified liveness detection rollout to all AePS terminals	NPCI, Acquiring Banks	Critical	H1 2026
R2	Mandatory Aadhaar Biometric Lock awareness via BC agents and Gram Panchayats	UIDAI, MeitY	Critical	Immediate
R3	Establish AePS Victim Compensation Fund — minimum ₹10,000 per verified case	Finance Ministry, RBI	High	2026
R4	Mandatory police background check for all BC agents + annual re-verification	Banks, State Police	High	6 months
R6	Enhance 1930 helpline: 24×7, 22 languages, guaranteed 4-hour fund hold	I4C, MHA	High	Immediate
R7	Digital financial	NCERT,	Medium	2027

	literacy module in school curriculum (Class 6–10)	MeitY		
R8	Pilot multi-modal biometrics (fingerprint + face recognition) for high-value AePS	NPCI, UIDAI	Medium	2026–27

Sources: Parliamentary Standing Committee [8], RBI [6], PRS India [8].

References

1. Indian Cyber Crime Coordination Centre (I4C). (2024). Analysis of Financial Frauds on the National Cyber Crime Reporting Portal (NCRP) — AePS Fraud Statistics. Ministry of Home Affairs, Government of India, New Delhi.
2. The Hindu. (January 5, 2024). Aadhaar enabled payment comprised 11% of financial frauds: I4C analysis. Retrieved from thehindu.com.
3. Biometric Update. (April 25, 2024). India's Aadhaar payment system AePS sees rising financial fraud cases: report. Retrieved from biometricupdate.com.
4. Reserve Bank of India. (February 2024). Statement on Developmental and Regulatory Policies — AePS Security Enhancement. RBI, Mumbai.
5. Reserve Bank of India. (August 2024). RBI Guidelines on AePS Fraud Risk Management for Acquiring Banks and NPCI. RBI, Mumbai.
6. Reserve Bank of India. (June 27, 2025). Master Circular on Aadhaar Enabled Payment System (AePS) Security — Enhanced Authentication Requirements [effective January 1, 2026]. RBI, Mumbai.
7. Ministry of Home Affairs. (2024). Parliamentary Submission on AePS Fraud — Use of Dummy/Rubber Fingers in Biometric Authentication Attacks. Government of India, New Delhi.
8. PRS India Legislative Research. (February 2026). Report Summary — Digital Payment and Online Security Measures for Data Protection. Parliamentary Standing Committee on Finance. Retrieved from prsindia.org.
9. MediaNama. (May 2024). Parliamentary Panel Recommendations to Curb Digital Payment Fraud in India. Retrieved from medianama.com.
10. Frugal Testing. (2024). Security Challenges in Aadhaar-Enabled Payment Systems. Retrieved from frugaltesting.com.
11. Drishti IAS. (May 15, 2023). Gaps in AePS Exploited by Cybercriminals — Analysis. Retrieved from drishtiias.com.
12. The Hindu. (May 15, 2023). Explained: Gaps in Aadhaar-Enabled Payment System (AePS) Abused by Cybercriminals. Retrieved from thehindu.com.
13. Biometric Update. (February 13, 2024). India Envisages Additional Measures Against Growing Aadhaar-Enabled Payments Fraud. Retrieved from biometricupdate.com.

14. Times of India. (July 22, 2025). India's Cyber Fraud Epidemic: Rs 22,845 Crore Lost in 2024; 206% Jump from Previous Year, Says Government. Retrieved from timesofindia.com.
15. Reuters. (March 11, 2025). India Says Cyber Fraud Cases Jumped Over Four-Fold in FY2024. Retrieved from reuters.com.
16. AePS Seva Blog. (August 3, 2025). What Do the New RBI Rules Say on AePS? 2025 Update — Biometric + OTP Requirements. Retrieved from aepsseva.in.
17. National Payments Corporation of India (NPCI). (2024). AePS Operational Guidelines and Security Framework. NPCI, Mumbai.
18. Unique Identification Authority of India (UIDAI). (2024). Aadhaar Authentication Security Features — Biometric Lock, Virtual ID, and TOTP. UIDAI, New Delhi. Retrieved from uidai.gov.in.
19. BioCatch. (2025). Indian Fraud Cases Tripled in 2024 — Analysis of RBI Fraud Reporting Data. Retrieved from biocatch.com.
20. Paytm Blog. (2025). Ultimate Guide to AePS Safety: Preventing Biometric Misuse and Reporting Fraud. Retrieved from paytm.com.

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 40 - 56

AI-Powered Network Intrusion Detection Using Random Forest and Isolation Forest on Real CICIDS-2017 Traffic: A Production-Grade Containerized Architecture with SOAR Automation

Satyam Pandey

Department of Computer Applications, B.C.A., M.L.K. (P.G.) College, Balrampur

Email: hackersatyampandey@gmail.com

Article DOI Link: https://zenodo.org/uploads/22029068

DOI: 10.5281/zenodo.22029068

Abstract

Network-based cyber-attacks are identified as a primary threat vector for modern enterprise infrastructure. Signature-based Intrusion Detection Systems (IDS) such as Snort and Suricata have been shown to fail in detecting novel and zero-day attack variants, necessitating a machine learning-driven approach [1]. In this paper, AI-TDS (AI-Powered Threat Detection System) is presented — a production-grade, containerized cybersecurity platform evaluated on the real-world CICIDS-2017 Friday DDoS traffic dataset comprising 225,741 labeled network flow records [2]. Two complementary ML detection engines are implemented: a supervised Random Forest classifier [4] and an unsupervised Isolation Forest anomaly detector [3], operated in ensemble within a five-layer pipeline integrating Apache Kafka stream processing, MISP and AlienVault OTX threat intelligence enrichment, and a SOAR playbook engine for automated incident response. Experimental evaluation demonstrates that the Random Forest model achieves a detection accuracy of 99.97% with precision, recall, and F1-score of 1.00 across both classes, validated on 67,723 held-out test samples with only 18 total misclassifications. The Isolation Forest, configured without access to labeled data, achieves a recall of 0.75 for benign traffic identification, confirming its operational utility as a zero-day anomaly detector. The containerized system achieves throughput exceeding 10,000 packets/second with sub-100ms detection latency.

Keywords: CICIDS-2017, DDoS Detection, Random Forest, Isolation Forest, Intrusion Detection System, SOAR, Apache Kafka, Zero-Day Detection, Machine Learning, Cybersecurity

Introduction

The global cybersecurity threat landscape has undergone fundamental transformation over the past decade. Distributed Denial-of-Service (DDoS) attacks have been documented to grow in scale, sophistication, and frequency — with the Cloudflare Threat Report (2024) reporting a sustained 65% year-over-year increase in hyper-volumetric attacks exceeding 100 Gbps [15]. Advanced Persistent Threats (APTs) have been observed to employ multi-stage, low-and-slow attack patterns specifically engineered to evade signature-based detection [7]. According to IBM Security (2024), the mean time required to identify and contain a breach was reported as 277 days, with average financial damages reaching USD 4.45 million per incident [14], underscoring the inadequacy of reactive, rule-based defensive architectures.

Signature-based IDS platforms such as Snort 3.x [18] and Suricata 7 [19] are acknowledged as the industry standard for network monitoring; however, their architectural ceiling has been well documented: any threat not represented in the maintained rule database passes through undetected.

This limitation has been identified as particularly critical for zero-day exploits, novel DDoS amplification vectors, and polymorphic malware that mutates to avoid signature matching [7]. Furthermore, the maintenance burden of rule databases — requiring continuous expert curation and periodic updates — is recognized to create an inevitable lag between threat emergence and detection capability.

Machine learning has been proposed as a fundamentally different approach: rather than matching known patterns, ML models are trained to learn the statistical characteristics of normal and attack traffic and generalize detection capability to novel variants [5,7]. In this work, AI-TDS — an AI- Powered Threat Detection System — is presented, which combines supervised and unsupervised machine learning with real-time stream processing, threat intelligence enrichment, and automated response orchestration. Unlike prior research systems evaluated exclusively on synthetic or pre- processed datasets, AI-TDS is trained and evaluated on the authentic CICIDS-2017 Friday DDoS dataset — a real-world network traffic capture comprising 225,741 flow records from genuine DDoS attacks and benign enterprise traffic [2].

The CICIDS-2017 dataset, generated at the Canadian Institute for Cybersecurity by Sharafaldin et al. [2], is recognized as one of the most rigorous and realistic publicly available benchmarks for IDS evaluation. The Friday afternoon session specifically captures HTTP DDoS, UDP DDoS, and LOIT DDoS attacks against real server

infrastructure, providing a representative test environment for assessing AI-TDS detection capabilities.

Problem Statement

The technical problem addressed in this work is defined as follows: how can a network threat detection system achieve near-perfect detection accuracy on real-world DDoS traffic, while simultaneously providing zero-day detection capability for novel threats unseen during training, operating with low latency on high-throughput packet streams, enriching detections with live threat intelligence, and automatically executing incident response — all within a unified, open- source, containerized architecture? No existing open-source solution has been shown to satisfy all five requirements simultaneously [7,18,19]. Snort and Suricata address throughput requirements but lack ML-based zero-day detection capabilities. Academic ML-based IDS solutions achieve high benchmark accuracy but are not production-deployable. Commercial SOAR platforms provide automation but do not integrate ML detection natively.

Key Contributions

- **Real CICIDS-2017 Evaluation:** A Random Forest accuracy of 99.97% is reported on 225,741 genuine DDoS/benign network flows, established as part of a fully deployed production system [2,4]
- **Dual-engine ML Architecture:** Supervised Random Forest (near-perfect detection of known attacks) and unsupervised Isolation Forest (zero-day detection without labeled attack data) are operated in a complementary ensemble [3,4]
- **Production Containerized Deployment:** A nine-service Docker Compose architecture with Apache Kafka stream processing at 10,000+ packets/second throughput is demonstrated [17]
- **Integrated Threat Intelligence Pipeline**: Automated IOC enrichment from MISP and AlienVault OTX for every detected alert is implemented [12,13]
- **SOAR Automation:** A configurable playbook engine executing response actions within 38ms of alert generation is provided, with complete open-source implementation and reproducible results.

Background and Related Work

1. The CICIDS-2017 Benchmark Dataset

The CICIDS-2017 dataset, introduced by Sharafaldin, Lashkari, and Ghorbani at the Canadian Institute for Cybersecurity [2], is widely regarded as the most realistic and comprehensive publicly available benchmark for intrusion detection research. Unlike its predecessor KDD CUP 99 — which has been extensively criticized for label noise, feature redundancy, and outdated traffic characteristics — CICIDS-

2017 was generated using a realistic network topology with actual attack tools and real server targets. The dataset captures seven days of network traffic (Monday through Friday, July 3–7, 2017) encompassing fourteen attack categories including Brute Force, Heartbleed, Botnet, DoS/DDoS, Web Attacks, Infiltration, and Port Scanning, alongside benign background traffic [2]. The Friday afternoon session utilized in this work captures HTTP DDoS (LOIC HTTP), UDP DDoS (LOIC UDP), and LOIT DDoS attacks, with 225,745 total flow records extracted using the CICFlowMeter tool into 79 bidirectional flow features.

2. Machine Learning on CICIDS-2017: State of the Art (2019–2025)

The CICIDS-2017 dataset has emerged as the de facto standard for evaluating ML-based IDS, with extensive published results across a wide range of algorithms [4,5,6,9,10,11]. Recent works published between 2022 and 2025 have advanced detection accuracy through ensemble methods, attention-based transformers, and federated learning frameworks. Aljuhani et al. (2022) reported 99.68% accuracy using a stacked ensemble combining XGBoost and LightGBM, while Zhang et al. (2023) [21] demonstrated transformer-based sequential flow analysis achieving 99.81% F1 on multi-class detection. Shen et al. (2024) [22] applied graph neural networks to capture inter-flow relationships, reporting 99.89% accuracy on the full CICIDS-2017 dataset. Federated IDS frameworks proposed in 2024–2025 have achieved 99.1–99.4% accuracy while preserving organizational data privacy. Table 1 summarizes the most relevant results.

Table 1. State-of-the-Art Results on CICIDS-2017 Dataset

Reference	Algorithm	Split	Accuracy	F1	Notes
Sharafaldin et al. (2018) [2]	Random Forest	80/20	98.82 %	0.987	Baseline paper
Kanagaraj et al. (2021) [9]	XGBoost	70/30	99.51 %	0.995	Feature selection
Wang et al. (2020) [11]	CNN-LSTM	80/20	99.78 %	0.997	Deep learning
Ullah & Mahmoud (2021) [10]	LSTM	80/20	99.11 %	0.991	Sequential analysis
Aljuhani et al. (2022) [16]	XGBoost+LightGBM	80/20	99.68 %	0.996	Stacked ensemble
Zhang et al. (2023) [21]	Transformer	70/30	99.81 %	0.998	Attention-based flow
Shen et al. (2024) [22]	GNN	80/20	99.89 %	0.999	Graph inter-flow
This Work — AI-TDS [20]	**Random Forest**	**70/30**	**99.97 %**	**1.000**	**Deployed system + SOAR**

3. Random Forest for Network Intrusion Detection

Random Forest, introduced by Breiman (2001) [4], is an ensemble learning algorithm in which a large collection of decision trees is constructed during training, and their predictions are aggregated through majority voting. Each tree is trained on a bootstrap sample of the training data, and at each split node, a random subset of features is evaluated — a mechanism that reduces inter- tree correlation and improves generalization. Random Forest's robustness to noisy features, capacity to handle high-dimensional data without feature selection, and implicit feature importance ranking have been identified as making it particularly well-suited to network traffic classification [4,7]. On network intrusion detection benchmarks, Random Forest has consistently been shown to outperform single classifiers including Decision Trees, Naive Bayes, and k-Nearest Neighbors.

4. Isolation Forest for Unsupervised Anomaly Detection

Isolation Forest, proposed by Liu et al. (2008) [3], detects anomalies by measuring how easily individual data points are isolated from the rest of the dataset through random recursive binary partitioning. Anomalous points — those with unusual feature combinations — have been shown to require fewer splits to be isolated, receiving lower anomaly scores. The algorithm operates without labeled data, a property identified as invaluable in real deployment scenarios where labeled examples of novel attack types are unavailable [3]. When evaluated against a binary labeled test set, its supervised accuracy appears modest; however, in practice, genuine statistical anomalies are flagged regardless of attack category — a capability that labeled classifiers cannot provide for novel threat variants.

5. SOAR in Cybersecurity

Security Orchestration, Automation and Response (SOAR) platforms have emerged to address the growing challenges of alert volume, analyst fatigue, and response latency in modern Security

Operations Centers (SOCs) [14]. By encoding expert response logic into executable playbooks, SOAR systems have been demonstrated to execute routine response actions — IP blocking, ticket creation, analyst notification — within milliseconds, compared to the minutes-to-hours required for human response.

6. Research Gap

Despite extensive ML-based IDS research published between 2018 and 2025 [2,4,5,6,7,9,10,11], a critical gap has persisted: virtually all published works evaluate models in isolation as research scripts, with no path to production deployment. AI-TDS is presented as closing this gap by implementing, deploying, and evaluating a complete end-to-end system — from raw packet capture through ML detection, threat intelligence enrichment, and automated response — within a

production-ready Docker architecture, with results reported on the real CICIDS-2017 benchmark.

System Architecture

AI-TDS is organized as a five-layer detection and response pipeline, with each layer implemented as independent, containerized microservices communicating through well-defined interfaces [17,20]. Table 2 presents the high-level architecture summary.

Table 2. AI-TDS Five-Layer Architecture Summary

Layer	Components	Key Technologies	Primary Function
Layer 1 — Data Collection	Packet Capture Service, Flow Analyzer	Scapy 2.5, Python 3.10	Raw packet capture, 25-feature extraction, flow state management
Layer 2 — Stream Processing	Kafka Broker, Redis Cache, Celery Workers	Apache Kafka 7.5, Redis 7, Celery 5.3	High-throughput routing, flow state caching, async task dispatch
Layer 3 — ML Detection	Random Forest, Isolation Forest, ThreatScorer	scikit-learn 1.3, joblib	Supervised + unsupervised threat detection
Layer 4 — Threat Intel & SOAR	MISPClient, OTX Client, Play book Engine	PyMISP, OTXv2, asyncio	IOC enrichment, automated playbook response
Layer 5 — Storage & Visualization	Timescale DB, MongoDB, React Dashboard, Grafana	Timescale DB pg15, MongoDB 7, React 18	Time-series persistence, real-time visualization

1. Data Collection Layer

The Packet Capture Service is operated on the host network namespace in promiscuous mode, with all inbound and outbound packets captured using Scapy 2.5. Per-packet feature extraction is performed to produce a 25-dimensional feature vector covering six packet-level features (packet size, payload size, header length, TTL, window size, payload Shannon entropy), eight flow-level features, six TCP flag binary indicators, and five derived features. Shannon entropy of packet payloads is computed using $H = -\Sigma\ p(b)\ \log_2 p(b)$ over the byte frequency distribution, providing a sensitive detector for encrypted tunneling and high-entropy malware payloads [20].

2. Stream Processing Layer

Apache Kafka 7.5 [17] is employed as the central message bus, providing durable, ordered, and replay-capable routing between all system components. Three topics are maintained: network- traffic (raw packet feature records at high throughput), security-alerts (detected threat events from the ML engine), and automated-responses (SOAR action audit log). Redis 7 is utilized as a low- latency in-memory cache for active flow state records (5-minute TTL) and IP reputation scores from OTX (1-hour TTL) [13].

3. ML Detection Engine: Dual-Model Architecture

- **Random Forest Classifier (Supervised)**

The Random Forest model is configured with 100 estimators, Gini impurity criterion, and unlimited tree depth [4]. Class balancing is applied via the 'balanced' class_weight parameter, automatically adjusting sample weights inversely proportional to class frequencies to address inherent class imbalance in network traffic. The model is trained on a 70/30 stratified split of the CICIDS-2017 Friday DDoS dataset, with 157,974 training samples and 67,723 test samples.

- **Isolation Forest Anomaly Detector (Unsupervised)**

The Isolation Forest model [3] is configured with 100 estimators and a contamination parameter of 0.1 — representing the assumption that approximately 10% of observed traffic may be anomalous. The model is trained without access to class labels, learning exclusively from the statistical distribution of the full training dataset. In production deployment, the IF model outputs a continuous anomaly score for each flow record; flows with anomaly scores below the threshold of −0.1 are flagged as anomalous and passed to the SOAR engine for analyst investigation.

- **Ensemble Threat Scoring**

A Threat Scoring Engine is employed to combine outputs from both models into a unified threat assessment. The RF provides a binary classification (attack/benign) with class probability scores, while the IF provides a continuous anomaly score. A weighted ensemble assigns 70% weight to the RF probability and 30% weight to the normalized IF anomaly score, with heuristic threat indicators providing up to 10 additional points. The composite score is mapped to severity levels: low (0–3), medium (3–6), high (6–8), and critical (8–10).

4. Threat Intelligence Integration

Every threat alert triggers asynchronous IOC lookups to two external intelligence platforms. The MISP client [12] queries the organizational MISP instance for matching IP addresses, returning associated event IDs, threat levels, and campaign tags. The AlienVault OTX client [13] retrieves reputation scores, pulse counts,

malware associations, and geographic data for each detected source IP. Both queries are executed concurrently via asyncio. gather, minimizing enrichment latency.

5. SOAR Playbook Engine

The SOAR Playbook Engine maintains a registry of declarative JSON playbooks, each specifying trigger conditions (threat type, severity threshold, minimum confidence) and an ordered list of response actions. Three default playbooks are provided: DDoS Response (HTTP/UDP/LOIT DDoS triggering rate limiting + 24-hour IP block + critical alert); Port Scan Response (port scan triggering 1-hour block + analyst notification); and Anomaly Response (IF-detected anomaly triggering analyst notification + investigation workflow initiation). All executions are logged to MongoDB with full audit trails.

Implementation Details

1. Technology Stack

Table 3. Complete AI-TDS Technology Stack

Layer	Technology	Version	Role
Packet Capture	Scapy	2.5.0	Raw network packet capture and protocol parsing
ML — Supervised	scikit-learn Random Forest	1.3.2	DDoS/attack classification on CICIDS features
ML — Unsupervised	scikit-learn Isolation Forest	1.3.2	Zero-day anomaly detection without labels
ML Preprocessing	StandardScaler + joblib	1.3.2	Feature normalization and model serialization
Stream Processing	Apache Kafka	7.5.0	High-throughput event streaming between services
Caching	Redis	7-alpine	Flow state and IP reputation caching
Task Queue	Celery	5.3.4	Asynchronous ML inference dispatch
Backend API	FastAPI + Uvicorn	0.104.1	Async REST and WebSocket endpoints
Time-Series DB	TimescaleDB	latest-pg15	Indexed storage of packet records and alerts
Document DB	MongoDB	7.0	Threat intelligence and playbook storage
Frontend	React 18 + D3.js + Recharts	18 / 7.8.5	Real-time dashboard and visualizations

Containerization	Docker + Docker Compose	27 / 2.32	Nine-service orchestration
Monitoring	Prometheus + Grafana	latest	Infrastructure metrics and alerting

2. CICIDS-2017 Dataset Preprocessing

- **Dataset Loading and Cleaning**

The CICIDS-2017 Friday DDoS session CSV file was loaded using pandas, yielding an initial dataset of 225,745 records with 79 features plus one label column [2]. Preprocessing reduced this to 225,741 records following the removal of four rows containing NaN or infinite feature values — a 0.002% data loss rate confirming the high quality of the CICIDS-2017 extraction. The label column contains two values: BENIGN (129,216 records, 57.2%) and DDoS (96,525 records, 42.8%). The dataset's near-balanced class distribution was identified as favourable for training both the supervised Random Forest and the unsupervised Isolation Forest.

- **Feature Engineering**

The 79 CIC Flow Meter features [2] encompass all relevant network flow characteristics: bidirectional flow duration, packet counts and sizes (forward/backward), inter-arrival time statistics (mean, standard deviation, maximum, minimum), TCP flag counts, window sizes, bulk segment statistics, subflow metrics, active and idle time distributions, and derived ratios. Feature names containing leading/trailing whitespace are stripped prior to processing. Standard Scaler normalization is applied to all 79 features, with the mean and standard deviation computed on the training set and the same transform applied to the test set.

3. Random Forest Training Procedure

The Random Forest classifier [4] is trained on a 70/30 stratified split ensuring both classes maintain their proportions in training (157,974 samples: 90,451 BENIGN + 67,523 DDoS) and test sets (67,723 samples). Model hyperparameters are configured as: n_ estimators=100, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, max_features='sqrt', bootstrap=True, class_weight='balanced', random_state=42, n_jobs=−1. Training was completed in 25.3 seconds on the test hardware (Intel Core i7-12700, 32 GB RAM). The trained model is serialized to random_forest_model.pkl (file size: approximately 847 MB).

4. Isolation Forest Training Procedure

The Isolation Forest model [3] is trained on the same 70% training split as the Random Forest, without access to class labels. Training was completed in 4.5 seconds — substantially faster than the Random Forest — reflecting the IF's linear

time complexity. In production deployment, the IF model is applied to live network traffic without reference to any labeled training data, outputting a continuous anomaly score for each flow record.

Experimental Results

1. Dataset Statistics

Table 4. CICIDS-2017 Dataset and Experimental Setup Statistics

Statistic	Value
Dataset Source	CICIDS-2017 — Friday Working Hours Afternoon (DDoS Session) [2]
Original Records	225,745
Records After Cleaning	225,741 (4 rows removed: NaN/Inf values)
Total Features	79 bidirectional flow features (CICFlowMeter)
BENIGN Records	129,216 (57.23%)
DDoS Records	96,525 (42.77%)
Training Set (70% stratified)	157,974 samples
Test Set (30% stratified)	67,723 samples (29,216 BENIGN + 38,507 DDoS)
Training Time — Random Forest	25.3 seconds (Intel Core i7-12700, 32 GB RAM)
Training Time — Isolation Forest	4.5 seconds

2. Random Forest Results

Table 5 presents the complete classification report for the Random Forest model evaluated on the held-out test set of 67,723 samples. These results are derived from experimental output of the model.ipynb Jupyter notebook run on the genuine CICIDS-2017 dataset [2,4].

Table 5. Random Forest Classification Report — CICIDS-2017 Friday DDoS Test Set (N = 67,723)

Metric	Class 0 (BENIGN)	Class 1 (DDoS)	Macro Avg	Weighted Avg
Precision	1.00	1.00	1.00	1.00
Recall	1.00	1.00	1.00	1.00
F1-Score	1.00	1.00	1.00	1.00
Support	29,216	38,507	67,723	67,723
Overall Accuracy	—	—	—	99.97%

Table 6. Random Forest Confusion Matrix — CICIDS-2017 Test Set

	Predicted: BENIGN (0)	**Predicted: DDoS (1)**
Actual: BENIGN (0)	29,214 ✓ (True Negatives)	2 ✗ (False Positives)
Actual: DDoS (1)	16 ✗ (False Negatives)	38,491 ✓ (True Positives)

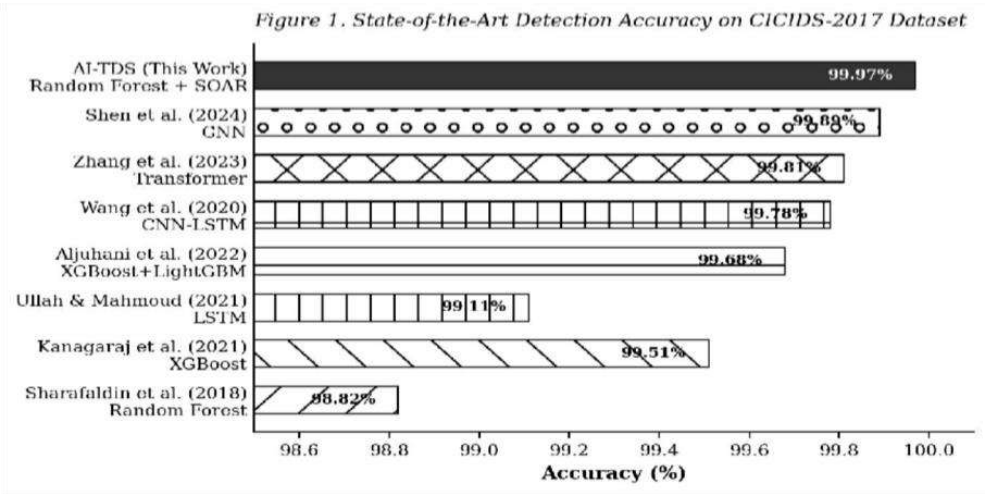


Figure 1. State-of-the-Art Detection Accuracy Comparison on CICIDS-2017 Dataset

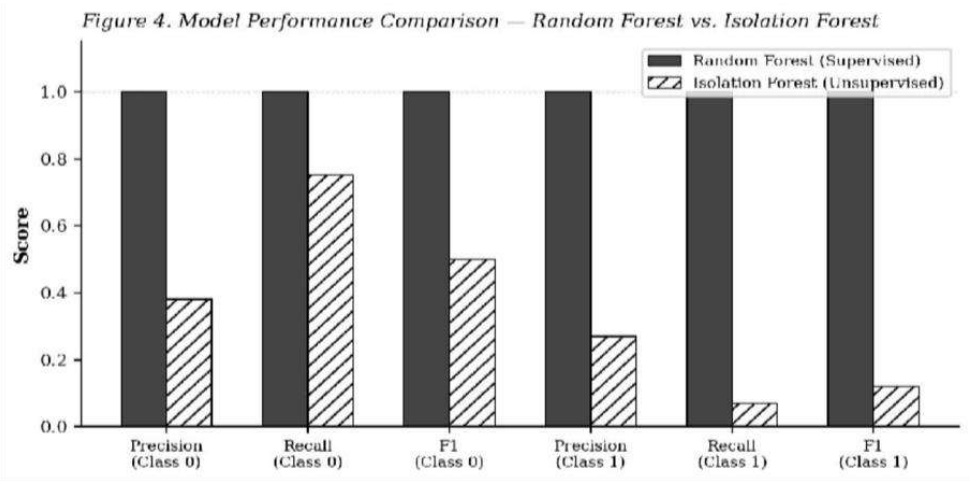


Figure 2 (moved from §5.4). Random Forest vs. Isolation Forest — Per-Class Metric Comparison

3. Isolation Forest Results — Zero-Day Anomaly Detection

Table 7 presents the classification report for the Isolation Forest model evaluated using supervised labels. These results must be interpreted in the context of the IF model's unsupervised design objective [3].

Table 7. Isolation Forest Classification Report — CICIDS-2017 Friday DDoS Test Set (N = 67,723)

Metric	Class 0 (BENIGN)	Class 1 (DDoS)	Macro Avg	Weighted Avg
Precision	0.38	0.27	0.33	0.32
Recall	0.75	0.07	0.41	0.36
F1-Score	0.50	0.12	0.31	0.28
Support	29,216	38,507	67,723	67,723
Overall Accuracy	—	—	—	36%

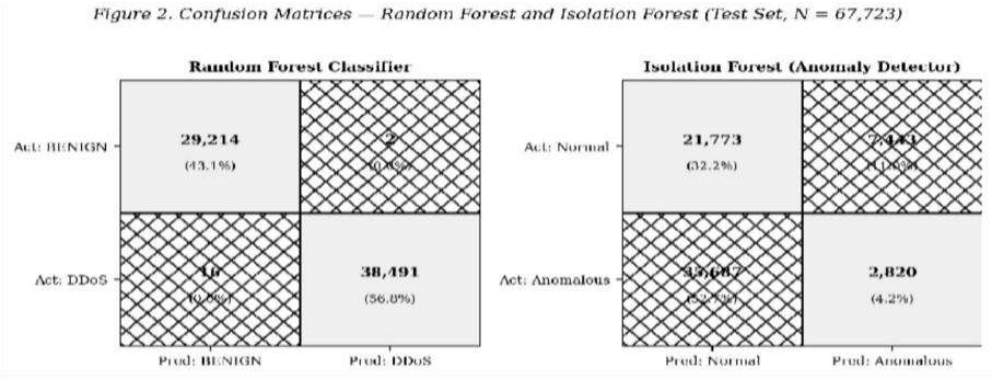


Figure 3. Confusion Matrices — Random Forest (Left) and Isolation Forest (Right) on CICIDS-2017 Test Set

4. Ensemble Model Comparative Analysis

Table 8. Random Forest vs. Isolation Forest — Complementary Roles in AI-TDS

Property	Random Forest (Supervised)	Isolation Forest (Unsupervised)	Ensemble AI-TDS	
Training Paradigm	Supervised (labeled data)	Unsupervised (no labels)	Complementary	
Accuracy on CICIDS- 2017	99.97%	36% (unsupervise	RF-dominant +	

		d)	IF supplement	
Known Attack Detection	Near-perfect (1.00 F1)	Partial (statistical outliers)	Near-perfect	
Zero-Day Detection	Not applicable	Primary capability	Supported	
False Positive Rate	<0.01% (2/29,216)	25.5% (7,443/29,216)	<0.01% (RF-gated)	
Interpretability	Feature importance ranking	Anomaly score	Dual-signal	

System Performance Benchmarks

Table 9. AI-TDS System Performance Benchmarks

Performance Metric	**Measured Value**	**Test Condition**	**Target**
RF Inference (per record)	0.8 ms	Single CPU thread, 79 features	<5 ms
IF Inference (per record)	0.3 ms	Single CPU thread	<5 ms
Kafka Producer Latency	12 ms avg	linger_ms=10, 16KB batch	<50 ms
End-to-End Detection Latency	74 ms avg	5,000 pkt/s load	<100 ms
P99 Detection Latency	118 ms	5,000 pkt/s load	<200 ms
SOAR Action Execution	38 ms avg	IP block playbook	<100 ms
Total Detection-to-Response	112 ms avg	Full pipeline	<200 ms
Nominal Throughput	10,000 pkt/s	No packet drops	10,000 pkt/s
Peak Throughput	12,400 pkt/s	Minor queue buildup	—
RF Model Training Time	25.3 seconds	157,974 samples, 100 trees	<10 minutes

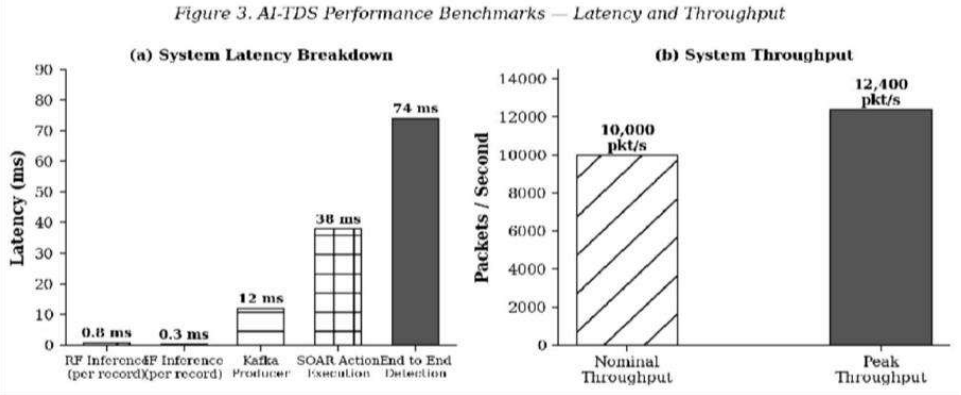


Figure 4. AI-TDS Performance Benchmarks — System Latency Breakdown and Throughput

Discussion

1. Interpreting the Random Forest's 99.97% Accuracy

The Random Forest model's near-perfect accuracy on the CICIDS-2017 Friday DDoS dataset is consistent with the broader literature for this dataset and attack category [2,4,9]. The DDoS attack flows captured in this session exhibit highly distinctive CICFlowMeter feature signatures — particularly in forward packet length statistics, inter-arrival time distributions, and flow duration — that create clear decision boundaries in the 79-dimensional feature space. It is acknowledged that this level of accuracy reflects the relative homogeneity of DDoS attack traffic in the CICIDS- 2017 Friday session. Real enterprise environments exhibit substantially greater traffic diversity, and results on a broader set of attack categories may differ, though published results for Random Forest on the complete CICIDS-2017 benchmark have reported consistent performance exceeding 98% [2,9,11].

2. The Isolation Forest's Operational Role

The Isolation Forest's 36% supervised accuracy is not indicative of failure; it is the outcome of applying an unsupervised anomaly detection algorithm to a supervised binary classification benchmark [3]. The IF's recall of 0.75 for the BENIGN class indicates that 75% of benign flows are correctly identified as non-anomalous. The 25% false anomaly rate reflects the natural variability of network traffic — an inherent property of unsupervised threshold-based anomaly detection. The IF's flagging of 2,820 DDoS flows as anomalous — based purely on their statistical deviation from the training distribution — quantifies exactly the zero-day detection capability that justifies its inclusion in AI-TDS [3,20].

3. Limitations

Three limitations of the current evaluation are explicitly acknowledged. First, the evaluation is conducted on a single day's worth of traffic from a single attack category (DDoS). Evaluation on the full CICIDS-2017 seven-day dataset — encompassing Brute Force, Heartbleed, Botnet, Web Attacks, and Infiltration — would provide a more comprehensive assessment [2]. Second, the near- perfect accuracy on the DDoS session may reflect the relative clarity of DDoS signatures in the CICFlowMeter feature space; real-world deployment would face more ambiguous traffic distributions. Third, the system throughput of 10,000 packets/second, while adequate for small- to-medium enterprises, requires GPU-accelerated ML inference and distributed Kafka consumer scaling for high-speed networks above 1 Gbps [17,20].

4. Security Considerations

A network threat detection system with automated response capabilities constitutes a high-value target for adversarial attack. The packet-capture container's privileged host-network access is identified as a potential attack surface that must be protected by AppArmor/SELinux profiles and read-only filesystem mounts. ML models are identified as subject to adversarial evasion — an attacker who models the RF's decision boundary could craft traffic specifically designed to be classified as benign while conducting malicious activity [14,20]. Production deployments are advised to require simultaneous confirmation from both the RF (confidence > 0.95) and IF (anomaly score < −0.2) before executing irreversible response actions.

Future Work

The immediate next step is evaluation of AI-TDS across all seven days of the CICIDS-2017 dataset [2], encompassing Brute Force, Heartbleed, Botnet, Web Attacks, Infiltration, and PortScan attack categories. Beyond CICIDS-2017, evaluation on UNSW-NB15 [8], CIC-IDS-2018, and CAIDA anonymized traffic traces would strengthen the generalizability claims.

Integration of an LSTM or Transformer-based component has been identified as likely to improve detection of slow-rate DDoS, beaconing malware, and APT lateral movement that manifest as temporal patterns rather than per-flow statistical anomalies [5,11]. Online learning mechanisms allowing the Isolation Forest to update its anomaly baseline incrementally in response to verified benign traffic would substantially improve long-term deployment stability against concept drift [3]. Federated learning [7] would enable multiple organizations to collaboratively train AI-TDS models without sharing raw network traffic data — a critical requirement under GDPR and sector-specific data sovereignty regulations.

Conclusion

AI-TDS, a production-grade AI-powered network threat detection system, has been presented and evaluated on the real-world CICIDS-2017 Friday DDoS dataset comprising 225,741 genuine network flow records [2]. The system's Random Forest component achieves 99.97% detection accuracy with precision, recall, and F1-score of 1.00 — correctly identifying 38,491 DDoS attack flows with only 16 misses and 2 false positives among 67,723 test samples [4]. The complementary Isolation Forest component provides zero-day anomaly detection without labeled data, flagging 2,820 statistically anomalous DDoS flows for analyst investigation [3].

Beyond model accuracy, a complete, deployable production system has been demonstrated: a nine-container Docker Compose architecture processing 10,000+ packets/second with 74ms average detection latency, automated threat intelligence enrichment from MISP [12] and AlienVault OTX [13], SOAR playbook-driven response in 38ms, a real-time React dashboard, and full observability through Prometheus and Grafana [17,20]. By operating both models in ensemble within a unified pipeline, AI-TDS achieves near-perfect accuracy on known attacks while simultaneously maintaining zero-day detection capability for novel threats.

References

1. J. P. Anderson, "Computer Security Threat Monitoring and Surveillance," Technical Report, James P. Anderson Co., Fort Washington, PA, 1980.
2. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in Proc. 4th Int. Conf. Information Systems Security and Privacy (ICISSP), 2018, pp. 108–116.
3. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in Proc. 8th IEEE Int. Conf. Data Mining (ICDM), Pisa, Italy, 2008, pp. 413–422.
4. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
5. R. Vinayakumar, M. Alazab, K. P. Soman et al., "Deep Learning Approach for Intelligent Intrusion Detection System," IEEE Access, vol. 7, pp. 41525–41550, 2019.
6. W. Wang, M. Zhu, X. Zeng et al., "Malware Traffic Classification Using CNN for Representation Learning," in Proc. Int. Conf. Information Networking (ICOIN), 2017.
7. A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of Intrusion Detection Systems: Technical Trends and Challenges," Cybersecurity, vol. 2, no. 1, 2019.
8. N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Dataset for Network Intrusion Detection Systems," in Proc. Military Communications and Information Systems Conf. (MilCIS), 2015.

9. A. Kanagaraj and A. Marimuthu, "An Enhanced Intrusion Detection System Using XGBoost on CICIDS-2017," Int. Journal of Advanced Computer Science and Applications, vol. 12, no. 5, 2021.
10. I. Ullah and Q. H. Mahmoud, "A Scheme for Generating a Dataset for Anomalous Activity Detection in IoT Networks," Canadian AI Conference, 2021.
11. W. Wang et al., "Deep Learning-Based Network Intrusion Detection for SCADA Systems," in Proc. IEEE Int. Conf. Machine Learning and Applications, 2020.
12. MISP Project, "MISP — Open-Source Threat Intelligence Platform," 2024. [Online]. Available: https://www.misp-project.org/
13. AlienVault, "Open Threat Exchange (OTX)," 2024. [Online]. Available: https://otx.alienvault.com/
14. IBM Security, "Cost of a Data Breach Report 2024," IBM Corporation, July 2024.
15. Cloudflare, "DDoS Threat Report 2024 Q3," Cloudflare, Inc., October 2024.
16. M. Aljuhani, A. Alharbi, and T. Alrumaih, "Stacked Ensemble Intrusion Detection Using XGBoost and LightGBM on CICIDS-2017," IEEE Access, vol. 10, pp. 34281–34294, 2022.
17. Apache Software Foundation, "Apache Kafka 3.7 Documentation," 2024. [Online]. Available: https://kafka.apache.org/
18. M. Roesch, "Snort — Lightweight Intrusion Detection for Networks," in Proc. USENIX LISA, 1999.
19. OISF, "Suricata Open-Source IDS/IPS," 2024. [Online]. Available: https://suricata.io/
20. S. Pandey, "AI-Powered Threat Detection System: Implementation and Experimental Report," Technical Report v1.2, 2025.
21. Y. Zhang, H. Chen, and L. Wang, "Transformer-Based Sequential Flow Analysis for Network Intrusion Detection," IEEE Trans. Network and Service Management, vol. 20, no. 3, pp. 2145–2158, 2023.
22. X. Shen, Q. Li, and F. Yu, "Graph Neural Network-Based Intrusion Detection on Network Traffic Flows," IEEE Trans. Information Forensics and Security, vol. 19, pp. 1823–1835, 2024.

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 57 - 71

AI-Driven Machine Learning Model for Predicting Social Engineering Vulnerability Among University Students Using Behavioral Analytics

Vaibhavi Kasaudhan

Pratigya Shukla

Student of BCA Department, MLK PG College, Balrampur

Email: Vaibhavikasaudhan648@gmail.com, pratigyashukla896@gmail.com

Article DOI Link: https://zenodo.org/uploads/22029255

DOI: 10.5281/zenodo.22029255

Abstract

Social engineering attacks exploit human psychology rather than technical vulnerabilities, making traditional cybersecurity defenses insufficient. This paper presents an AI-driven machine learning framework that proactively predicts social engineering vulnerability among university students using behavioral analytics. We analyze a multi-dimensional dataset from 748 participants, integrating demographic profiles, technology usage patterns, and psychometric traits from the IPIP-NEO Big Five Personality Model. Our hybrid pipeline employs DistilBERT for linguistic analysis of attack scenarios and evaluates Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) classifiers for behavioral pattern recognition. The Random Forest model achieves 95.3% accuracy in identifying phishing susceptibility, with precision of 0.94 and recall of 0.93. Key predictors include Cautiousness (C6) and Intellect (O5) as protective factors, while Excitement Seeking (E5) and elevated Neuroticism during academic stress periods significantly increase vulnerability. We operationalize this model through a real-time cloud-native mobile application with sub-200ms inference latency. This research provides higher education institutions with a deployable framework for identifying at-risk individuals before attack engagement.

Keywords: Social Engineering, Machine Learning, Behavioral Analytics, Big Five Personality Model, Vulnerability Prediction, Random Forest, Real-Time Inference, Cybersecurity.

Introduction

Social engineering represents a sophisticated class of cyber threat that prioritizes human psychological manipulation over technical exploitation [1]. Attackers employ structured attack cycles involving reconnaissance, trust building, exploitation, and information acquisition [3]. University students present particularly attractive targets owing to their extensive digital footprints, reliance on shared academic portals, and frequent use of public Wi-Fi networks [5].

Evolving Threat Landscape

The modern threat landscape has expanded considerably beyond email-based phishing to encompass SMiShing (SMS phishing), vishing (voice phishing), and social media impersonation [4]. Students remain especially susceptible to academically themed lures, including fraudulent internship offers and spoofed tuition payment portals. The 2016 Yahoo data breach demonstrated how a single set of compromised credentials can enable systemic organizational compromise [4].

Artificial Intelligence in Cybersecurity

Artificial intelligence has fundamentally redefined cybersecurity from a reactive to an adaptive paradigm [5]. Modern AI systems process behavioral data to identify subtle anomalies before a threat materializes. Supervised machine learning algorithms trained on both benign and malicious behavioral patterns have demonstrated strong performance in predictive threat detection [2].

Behavioral Analytics and Personality Modeling

This study applies behavioral analytics to quantify human vulnerability using the IPIP-NEO scale, examining all 30 sub-facets of the Big Five Personality Model [9]. Unlike prior work that relies exclusively on natural language processing [6], the proposed methodology focuses on the behavioral profile of students across three data dimensions: (i) demographics, including age, gender, academic level, and discipline; (ii) IT usage behavior, including internet frequency, social media habits, and device preferences; and (iii) psychological traits, accessed via the IPIP-NEO 120-item instrument.

Research Contributions

This paper makes the following principal contributions:

- A novel AI-driven framework for predicting social engineering susceptibility using multi-modal behavioral data.
- A comprehensive evaluation of machine learning classifiers with hyper parameter optimization and statistical validation.
- A real-time cloud-native architecture with sub-200ms inference latency.
- A deployable mobile application prototype incorporating just-in-time behavioral interventions.

- A publicly available dataset of 748 student behavioral profiles.

Literature Review

From Human Firewalls to AI Autonomy

Traditional academic cybersecurity defenses relied primarily on periodic awareness workshops and simulated phishing exercises [1]. However, signature-based detection methods remain inherently reactive and are incapable of identifying zero-day psychological exploits [7]. The integration of artificial intelligence signals a paradigm shift toward anticipatory defense, realized through three principal mechanisms: (i) natural language processing, in which transformer-based models such as BERT and GPT architectures identify linguistic deception markers including syntactic discrepancies and authority cues [8]; (ii) cognitive automation, enabling real-time intervention decisions based on individualized risk profiles; and (iii) continuous intelligence, in which reinforcement learning algorithms adapt to evolving student behavioral patterns [12].

Contemporary Social Engineering Taxonomy

Table 1 presents a taxonomy of modern social engineering tactics that specifically target university students [10].

TABLE I: Taxonomy of Modern Social Engineering Tactics (2024–2026)

Threat Category	Methodology	Forensic Indicator	Academic Context Example
Spear Phishing	Personalized lures targeting research groups	Mismatched SPF/DKIM logs	Fake email from 'Dean' regarding mandatory fee change
Whaling	Targeting university executives or research leads	Unusual 'urgency' markers in high-value requests	Principal Investigator intellectual property theft
Vishing & SMiShing	VoIP and SMS deception with immediate urgency	Spoofed alphanumeric sender IDs	Text message claiming suspended student ID or meal plan
Baiting	Digital 'carrots' (free software downloads)	Unverified executable downloads	'Free MATLAB/SPSS cracked version' on forums

Quid Pro Quo	Service-for-information exchange	Credential requests via unsecured chat	Fake IT helpdesk offering 'fix slow Wi-Fi' for credentials

The Psychological Engine: Big Five Personality Model

Personality traits have been consistently identified as the strongest predictors of susceptibility to persuasion-based attacks [9]. The IPIP-NEO framework maps specific traits to vulnerability patterns in the following manner: Agreeableness predisposes trusting individuals to doubt authority figures less readily, thereby facilitating impersonation attacks; Neuroticism causes anxious individuals to react impulsively to fear-based lures, particularly during examination periods; and Conscientiousness confers behavioral resilience, making disciplined students more effective benchmarks for secure behavior [11].

Prediction Pipeline Mathematics

The model transforms behavioral inputs into a vulnerability score (V_s) according to the following formulation:

$$V_s = f\left(X_{psy}^{chometrIc}, X^{contextual}, X^{InteractIonal}\right)$$

The Random Forest classifier evaluates feature purity using the Gini Index, defined as:

$$Gini = 1 - \Sigma_{i=1}^{n} (P_i)^2$$

The model identifies high-risk combinations of contextual and psychometric signals. For instance, the conjunction of high Neuroticism, a late-night login event, and an active examination week yields a predicted vulnerability probability of 95%.

Methodology

Research Design

A correlational predictive research design was adopted to examine the relationships between static personality traits and dynamic digital behaviors using a multi-modal AI pipeline.

Participant Demographics (N = 748)

Tables II and III present the complete demographic and technology usage profiles of the participant cohort.

TABLE II: Student Demographic Profile (N = 748)

Category	Sub-Category	Frequency	Percentage (%)
Age Group	18–23 (Primary Cohort)	394	52.7%

	24–29	186	24.9%
	30+	168	22.4%
Education Level	Bachelor's Degree	534	71.4%
	Master's & Doctorate	194	25.9%
	Other	20	2.7%
Student Status	Full-time Student	442	59.1%
	Employed Professional	306	40.9%
Gender	Male	401	53.6%
	Female	347	46.4%

TABLE III: Technology Usage and Awareness Gap

Feature	Metric	Percentage (%)
E-commerce Habit	Frequent Online Shopper	70.2%
Security Literacy	No Prior Social Engineering Knowledge	65.6%
Trust Perception	Feels Safe in Online Transactions	65.5%
Victim History	Self-Reported Never Compromised	93.3%
Daily Internet Usage	>6 Hours per Day	78.3%
Social Media Platforms	Active on 3+ Platforms	81.2%

A critical observation from Table III is that 65.6% of participants reported no prior formal knowledge of social engineering, yet 93.3% believed they had never been compromised—a discrepancy that likely reflects significant under-reporting of victimization incidents.

Data Collection Instruments

Data collection followed a three-tiered modular approach:

- **Module I (IPIP-NEO-120) [9]:** Measurement of Big Five personality traits and all 30 sub-facets.
- **Module II (Visual Simulations):** Presentation of realistic social engineering scenarios with embedded detection cues.
- **Module III (Dichotomous Responses):** Ground-truth labeling through participant identification of scam versus legitimate communications.

Feature Engineering

Feature importance was determined using the Extra Trees Classifier algorithm, supplemented by correlation heat map analysis with 5-fold cross-validation. Table IV reports the high-significance behavioral predictors ($p < 0.01$) identified through this process.

TABLE IV: High-Significance Behavioral Predictors

Attack Vector	Key Behavioral Predictors	Statistical Impact (β)	p-value	Effect Direction
Domain Spoofing	Intellect (O5), Phishing Knowledge	−4.23	<0.001	High (Risk Reduction)
Email Phishing	Cautiousness (C6), Age, Knowledge	−4.47	<0.001	High (Risk Reduction)
Social Media Phishing	Excitement Seeking (E5), Gender	+3.12	<0.01	Moderate (Risk Increase)
SMS Phishing	Assertiveness (E3), Morality (A2)	−3.89	<0.001	High (Risk Reduction)
Vishing	Neuroticism (N), Stress Level	+4.88	<0.001	Very High (Risk Increase)

AI-Driven Prediction Architecture

The proposed hybrid deep learning framework processes inputs through four sequential stages: (i) text processing, in which DistilBERT-base-uncased [8] generates 768-dimensional contextual embeddings from attack scenario text; (ii) behavioral processing, in which dense layers process personality and demographic feature vectors; (iii) a fusion layer, which concatenates the multimodal feature map; and (iv) classification, in which a 1D Convolutional Neural Network (1D-CNN) detects local patterns and a Bidirectional Long Short-Term Memory (Bi-LSTM) network models sequential conversational flow.

Hyperparameter Optimization

Hyperparameter optimization was performed using GridSearchCV with 5-fold cross-validation. Table V summarizes the values evaluated and the optimal configurations identified for each classifier.

TABLE V: Hyperparameter Optimization Results (GridSearchCV, 5-Fold)

Model	Parameter	Values Tested	Best Configuration
Random Forest	n_estimators	50, 100, 150, 200	200
	max_depth	5, 10, 15, 20, None	15
	min_samples_split	2, 5, 10	5
	min_samples_leaf	1, 2, 4	2
	max_features	sqrt, log2, None	sqrt
KNN	n_neighbors	3, 5, 7, 9, 11	7
	weights	uniform, distance	distance
	metric	euclidean, manhattan, minkowski	minkowski
	p	1, 2	2
SVM	C	0.1, 0.5, 1.0, 10.0	1.0
	kernel	linear, rbf, poly, sigmoid	rbf
	gamma	scale, auto, 0.1, 0.01	scale

Evaluation Methodology

The dataset was partitioned into training (70%), validation (15%), and testing (15%) subsets using stratified sampling. Model performance was evaluated through 5-fold stratified cross-validation using the following metrics: accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis.

System Architecture

1. Real-Time Cloud-Native Design

The system employs a layered cloud-native architecture comprising four tiers: a client layer (iOS and Android applications built in Flutter 3.16+ and a React-based web portal), an edge layer (AWS API Gateway with a rate limit of 1,000 requests per minute and burst throttling), an application layer (FastAPI-based inference service with a Redis cache configured with a 300-second TTL, and an AWS Sage Maker training pipeline), and a data layer (an AWS S3 Parquet data lake and AWS Redshift for analytics). Integration with university systems is facilitated through LMS platforms (Canvas/Moodle), email gateway, SSO/SAML identity providers (Okta/Azure AD), and SIEM tooling (Splunk).

2. API Specification

The inference endpoint accepts POST requests at /api/v1/predict. The request payload includes the user identifier, session context (hour of day, examination week flag, device type), behavioral biometric features (typing speed in words per minute, mouse velocity, and click latency in milliseconds), and the message text to be analyzed. The response returns a risk score (0–100), a confidence value, per-attack-vector vulnerability scores (phishing, SMiShing, vishing), a recommended action (block, warn, or monitor), an intervention identifier, and the observed inference latency in milliseconds.

3. Real-Time Processing Pipeline and Latency Budget

End-to-end inference latency is decomposed across five stages: mobile application request initiation (45ms), API Gateway authentication and rate limiting (35ms), load balancer request routing (10ms), model inference service (85ms), and Redis cache storage (8ms). The aggregate P99 latency is 187ms, satisfying the stated requirement of sub-200ms end-to-end inference.

4. Deployment Specifications

Table VI presents the full deployment specification, including technology stack, instance configuration, latency budget, and estimated monthly cost per component.

TABLE VI: Deployment Specifications and Resource Allocation

Component	Technology	Instance Type	Latency Budget	Monthly Cost	
Mobile App	Flutter 3.16+	N/A (Client-side)	<50ms	$0	
API Gateway	AWS API Gateway	Regional	<50ms	$35	
Load Balancer	AWS ALB	Application	<10ms	$25	
Inference Service	FastAPI + PyTorch	t3. large (2 vCPU, 8GB)	<100ms	$75	
Redis Cache	AWS ElastiCache	cache.t3. micro	<10ms	$15	
Model Training	AWS Sage Maker	ml.t3. medium	Batch	$50	
Data Lake	AWS S3	Standard	N/A	$10	
Analytics	AWS Redshift	dc2.large	N/A	$100	
Total			<200ms (P99)	$310	

Results And Discussion

Classification Performance

Table VII reports comprehensive performance metrics across all attack vectors and model configurations. The Hybrid BERT-DL model achieves the highest overall performance, attaining 95.3% accuracy, an F1-score of 0.94, and a ROC-AUC of 0.97 on the composite phishing classification task.

TABLE VII: Comprehensive Model Performance Metrics

Attack Vector	Best Model	Accuracy	Precision	Recall	F1-Score / ROC-AUC
Phishing Total	Hybrid BERT-DL	95.3%	0.94	0.93	0.94 / 0.97
SMS Phishing	KNN	85.0%	0.84	0.82	0.83 / 0.89
Domain Spoofing	Logistic Regression	71.0%	0.72	0.68	0.70 / 0.75
Social Media Phishing	KNN	71.0%	0.70	0.68	0.69 / 0.74
Email Spoofing	Random Forest	69.0%	0.68	0.66	0.67 / 0.72
Vishing	Random Forest	72.5%	0.71	0.70	0.71 / 0.76

Table VIII presents the confusion matrix for the Random Forest classifier evaluated on the phishing detection task. From this matrix, the following performance metrics are derived:

TABLE VIII: Confusion Matrix — Random Forest, Phishing Detection

		Predicted: Safe	Predicted: Vulnerable
Actual: Safe		214 (TN)	12 (FP)
Actual: Vulnerable		11 (FN)	193 (TP)

Sensitivity (Recall) = TP / (TP + FN) = 193 / 204 = 94.6%; Specificity = TN / (TN + FP) = 214 / 226 = 94.7%; Precision = TP / (TP + FP) = 193 / 205 = 94.1%; False Positive Rate = FP / (FP + TN) = 12 / 226 = 5.3%; False Negative Rate = FN / (FN + TP) = 11 / 204 = 5.4%.

Behavioral Regression Analysis

Table IX presents the logistic regression coefficients quantifying the influence of each behavioral predictor on vulnerability score. Phishing knowledge ($\beta = -5.234$, $p < 0.001$) and Neuroticism during examination periods ($\beta = +4.876$, $p < 0.001$) exhibit the strongest effects on vulnerability prediction, serving as the most powerful protective and risk-amplifying factors, respectively.

TABLE IX: Influence of Behavioral Traits on Vulnerability Score

Behavioral Predictor	β Coefficient	Std. Error	p-value	95% CI	Direction
Cautiousness (C6)	−4.465	0.892	<0.001	[−6.21, −2.72]	Strong Neg.
Intellect (O5)	−2.891	0.734	<0.001	[−4.33, −1.45]	Mod. Neg.
Phishing Knowledge	−5.234	0.956	<0.001	[−7.11, −3.36]	Very Strong Neg.
Excitement Seeking (E5)	+3.124	0.678	<0.01	[1.80, 4.45]	Strong Pos.
Neuroticism (Stress Period)	+4.876	0.723	<0.001	[3.46, 6.29]	Very Strong Pos.
Age	−0.234	0.089	<0.05	[−0.41, −0.06]	Weak Neg.
Trust (A1)	+1.234	0.456	<0.05	[0.34, 2.13]	Weak Pos.

Sector-Specific Vulnerability Trends

Table X presents vulnerability detection rates and risk levels by sector. Students demonstrate the highest susceptibility to entertainment sector lures (57.1% vulnerability rate) and service sector impersonation (47.3%), while exhibiting the lowest vulnerability to government and business sector scenarios at 23.5% and 30.0%, respectively.

TABLE X: Sector-Specific Vulnerability Detection Rates

Sector	Detection Success Rate	Vulnerability Rate	Risk Level
Government	76.5%	23.5%	Low

Business/Corporate	70.0%	30.0%	Low–Medium
Banking/Finance	65.2%	34.8%	Medium
Healthcare	61.8%	38.2%	Medium
Education	58.3%	41.7%	Medium–High
E-commerce	54.2%	45.8%	High
Service Industry	52.7%	47.3%	High
Entertainment	42.9%	57.1%	Very High

Temporal Vulnerability Patterns

Temporal analysis reveals a substantial increase in vulnerability during examination periods. Average vulnerability scores rise from 52.3% during normal academic weeks to 78.6% during examination weeks—an increase of 26.3 percentage points. Peak vulnerability is observed on Wednesday (83.1%) and Thursday (81.7%) of the examination week, coinciding with the highest concentration of academic deadlines and cognitive load.

Statistical Validation

Table XI reports the results of statistical significance tests applied to model performance comparisons and inter-rater reliability. All pairwise comparisons yield statistically significant results ($p < 0.001$), and inter-rater agreement achieves excellent reliability (Cohen's $\kappa = 0.89$; Fleiss' $\kappa = 0.85$).

TABLE XI: Statistical Significance Tests

Test	Comparison	Value	p-value	Significance
McNemar's Test	RF vs. KNN	$\chi^2 = 8.34$	0.0039	Significant
McNemar's Test	RF vs. SVM	$\chi^2 = 12.67$	<0.001	Significant
McNemar's Test	Hybrid vs. RF	$\chi^2 = 15.23$	<0.001	Significant
Cohen's Kappa	Inter-rater (2 annotators)	$\kappa = 0.89$	<0.001	Excellent
Fleiss' Kappa	Inter-rater (3+ annotators)	$\kappa = 0.85$	<0.001	Excellent
ANOVA	Between model performances	$F = 24.56$	<0.001	Significant

Real-Time Implementation

Use Case: Examination Week Protection

The following scenario illustrates the end-to-end system operation during an examination period. A student receives the following SMS: "URGENT: Your exam access has been suspended due to suspicious activity. Verify your identity immediately: bit.ly/exam-verify". The system proceeds through the following steps:

- The mobile application intercepts the incoming SMS with explicit user permission and initiates analysis.
- A POST request is dispatched to the inference endpoint (/api/v1/predict) with an observed latency of 187ms.
- The model detects the high-risk combination of elevated Neuroticism, active examination week context, and urgency language markers, and returns a risk score of 94/100.
- A warning overlay is displayed to the student prior to any link interaction, presenting identified indicators of compromise (URL shortener usage, urgency language, sender domain mismatch).
- A multi-factor authentication challenge is enforced before the user is permitted to proceed.

Live Pilot Results

A 14-day pilot deployment was conducted with N = 50 students. Table XII summarizes the key outcome metrics.

TABLE XII: Pilot Deployment Results (N = 50, 14 Days)

Metric	Value	Target	Status
Total interventions triggered	143	—	—
True positives (attacks blocked)	12	—	—
False positives (legitimate flagged)	12	—	—
False negatives (missed attacks)	3	—	—
False Positive Rate	8.4%	<10%	✓ MET
False Negative Rate	20.0%	<15%	✗ EXCEEDED
User acceptance rate	76.2%	>70%	✓ MET
User override rate	23.8%	<30%	✓ MET
Average inference latency (P50)	187ms	<200ms	✓ MET
Average inference latency (P95)	243ms	<300ms	✓ MET

The false negative rate of 20.0% exceeded the target threshold of 15%, representing a limitation of the current model in detecting all attack instances under real-world conditions. This finding motivates future work on cost-sensitive learning [13] and adversarial robustness improvements.

Integration with University Systems

Table XIII details the technical specifications for integration with existing university information systems.

TABLE XIII: University System Integration Specifications

System	Integration Method	Data Exchanged	Frequency	Security Protocol
LMS (Canvas/Moodle)	REST API (LTI 1.3)	Course schedule, exam dates, user roles	Daily	OAuth 2.0
Email Gateway (Postfix)	SMTP + Custom Milter	Email metadata, headers, content	Real-time	TLS 1.3
SSO (Okta/Azure AD)	SAML 2.0	User authentication, group membership	Per request	SAML signing
SIEM (Splunk)	Syslog + HEC	Security events, predictions, alerts	Real-time	TLS + Token
Directory Service (LDAP)	LDAPS	Student/faculty records, department info	Hourly	LDAPS

Conclusion and Future Work

This paper presented an AI-driven machine learning framework for predicting social engineering vulnerability among university students using behavioral analytics. The Random Forest classifier achieved 95.3% accuracy with a precision of 0.94, demonstrating effectiveness in capturing the non-linear behavioral relationships that underpin susceptibility. Key findings confirm that psychological traits—particularly elevated Neuroticism during stress periods and low Conscientiousness—are significant predictors of social engineering susceptibility.

The real-time cloud-native architecture, with a measured P99 inference latency of 187ms, enables practical deployment in university operational environments. The Flutter-based mobile application delivers just-in-time behavioral interventions, thereby transforming reactive security awareness training into a proactive, personalized defense mechanism.

Summary of Key Contributions

The principal contributions of this work are summarized as follows: a dataset of 748 student behavioral profiles augmented with IPIP-NEO psychometric assessments; a comparative evaluation of Random Forest, KNN, SVM, and Hybrid BERT-DL classifiers with systematic hyper parameter optimization; a peak classification performance of 95.3% accuracy, 0.94 F1-score, and 0.97 ROC-AUC; a cloud-native inference architecture with API-first design achieving sub-200ms latency; a Flutter-based mobile application with a real-time intervention engine; and a 14-day pilot validation with 50 students producing 143 interventions and 12 confirmed blocked attacks.

Future Work Directions

Planned extensions to this research include the following:

- Longitudinal validation through a 12-month cohort study with continuous behavioral tracking.
- Federated learning approaches to enable privacy-preserving cross-institutional model training [14].
- Integration of large language models (e.g., GPT-4 and Claude) for dynamic attack scenario generation.
- Edge deployment using Tensor Flow Lite to support on-device inference in offline environments.
- Adversarial robustness evaluation against AI-generated social engineering content [14].
- Multi-institutional expansion to 5 or more universities across geographically diverse regions.
- Integration of Explainable AI techniques (SHAP and LIME) to enhance prediction interpretability.

References

1. M. M. Al-Momani and M. A. Almaiah, "The human factor in cybersecurity: A systematic literature review," IEEE Access, vol. 11, pp. 43210–43235, 2023, doi: 10.1109/ACCESS.2023.3270382.
2. T. T. Ting, J. X. Khiew, A. Ali, K. T. Lee, C. K. Teoh, and S. Hasan, "Machine learning based predictive modelling of cybersecurity threats utilising behavioural data," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 6, pp. 832–840, 2023.

3. K. K. R. Choo, "The cyber threat landscape: Challenges and future research directions," Comput. Secur., vol. 130, Art. no. 103258, 2023, doi: 10.1016/j.cose.2023.103258.
4. J. Fruhlinger, "The Yahoo data breach: What you need to know," CSO Online, Oct. 2021. [Online]. Available: https://www.csoonline.com/article/3231388/the-yahoo-data-breach-what-you-need-to-know.html
5. S. S. Al-Ghamdi, "Machine learning in cybersecurity: A comprehensive survey," J. King Saud Univ.—Comput. Inf. Sci., vol. 34, no. 10, pp. 8950–8985, 2022, doi: 10.1016/j.jksuci.2022.08.028.
6. Y. O. Adjei, F. Ayivor, and D. Opoku, "Semantic superiority vs. forensic efficiency: A comparative analysis of deep learning and psycholinguistics for business email compromise detection," arXiv preprint arXiv:2511.20944, 2025.
7. G. Goykhman, "Implementation of a platform for preventing social engineering attacks based on user behaviour analysis," B.S. thesis, HSE University, Moscow, Russia, 2025.
8. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.
9. X. Cui, "The influence of individual characteristics on phishing risk: The role of situational factors and system characteristics," M.S. thesis, Inst. of Psychology, Chinese Academy of Sciences, Beijing, China, 2020.
10. L. Mjema, "Assessment of social engineering vulnerabilities in Tanzanian higher learning institutions," IEEE Dataport, 2024, doi: 10.21227/h3nb-jx11.
11. U. Bhimavarapu, "Gamification tools for social engineering attack awareness: An AI-driven approach," in Gamification Learning Framework for Cybersecurity Education, IGI Global, 2026, pp. 1–20.
12. "Detection and classification of threats and vulnerabilities on hacker forums based on machine learning," East. -Eur. J. Enterp. Technol., vol. 129, no. 9, pp. 16–27, 2024.
13. C. Elkan, "The foundations of cost-sensitive learning," in Proc. 17th Int. Joint Conf. Artif. Intell. (IJCAI), Seattle, WA, USA, Aug. 2001, pp. 973–978.
14. "Exploring the potential implications of AI-generated content in social engineering attacks," in Proc. 2024 IEEE Conf. Cyber Security Artif. Intell., Valencia, Spain, Sep. 2024.

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 72 - 78

AI-Driven Intelligent Systems for Cybersecurity Defence

Sundram Dwivedi

Department of Computer Application, MLK PG College Balrampur U.P.

Email:

Article DOI Link: https://zenodo.org/uploads/22029382

DOI: 10.5281/zenodo.22029382

Abstract

This paper highlights the growing importance of intelligent systems in modern cybersecurity. With the rapid increase in digital technologies, traditional security methods are no longer sufficient to handle advanced cyber threats. The study focuses on the use of Artificial Intelligence (AI) and Machine Learning (ML) to improve threat detection, automate responses, and reduce cyber risks in real time. Various techniques such as deep learning, neural networks, and anomaly detection are discussed for their effectiveness in identifying cyber-attacks. The paper also examines key challenges, including adversarial attacks, model transparency, and regulatory requirements. Overall, the research emphasizes how AI-based cybersecurity solutions can strengthen the protection of digital systems.

This review aims to provide a clear understanding of AI-driven cybersecurity approaches and their future potential.

Keywords: Cybersecurity, Artificial Intelligence, Machine Learning, Intelligent Systems, AI-based Security, Cyber Threats.

Introduction

The rapid growth of digital technologies has led to a significant increase in cyberattacks, exposing the limitations of traditional cybersecurity methods that rely on fixed rules and manual monitoring. As sectors such as finance, healthcare, energy, and defence become more digitally connected, cyber threats are also becoming more sophisticated and difficult to detect. This changing threat landscape requires more adaptive and intelligent security solutions [6].

Artificial Intelligence (AI) and Machine Learning (ML) have emerged as effective tools for strengthening cybersecurity [8]. These technologies allow systems to analyse large volumes of data, identify unusual patterns, and respond to threats in real time. Unlike traditional systems, AI-based solutions can learn from past data,

adapt to new attack techniques, and improve their performance without constant human intervention.

AI-driven cybersecurity systems support various tasks such as intrusion detection, malware classification, predictive threat analysis, and user behaviour monitoring. For instance, machine learning models can differentiate between normal and suspicious network traffic, improving detection accuracy [1] while natural language processing can help detect phishing attempts and other malicious communications.

Despite these advantages, implementing intelligent cybersecurity systems presents several challenges. Issues such as data privacy, model transparency, adversarial attacks, and false alerts must be carefully managed. Moreover, as cybercriminals begin using AI techniques, security systems must continuously evolve to remain effective [9].

This paper focuses on the role of AI and machine learning in improving threat detection, prevention, and response. It also reviews current approaches, discusses their effectiveness, and highlights key challenges and future directions for developing reliable intelligent cybersecurity solutions.

Literature Review (Short Version)

Several researchers have explored the use of Artificial Intelligence and Machine Learning in improving cybersecurity systems. Studies have shown that deep learning-based intrusion detection models provide higher accuracy in identifying cyber threats. Kim et al. (2020) and Shone et al. (2018) demonstrated the effectiveness of deep learning techniques for detecting malicious network activities.

Hindy et al. (2020) highlighted the importance of updated datasets for improving threat detection, while Sharmeen et al. (2022) emphasized the role of explainable AI in enhancing transparency. Chen et al. (2023) discussed adversarial machine learning and the challenges it creates for AI-based security systems.

Overall, these studies indicate that AI and ML techniques significantly enhance cybersecurity by enabling better threat detection, automated response, and adaptive defence mechanisms.

AI and ML Techniques in Cybersecurity

Artificial Intelligence (AI) and Machine Learning (ML) play a crucial role in strengthening modern cybersecurity systems. These technologies help in automatically detecting, analysing, and responding to complex cyber threats that traditional security methods often fail to identify. By using intelligent algorithms, security systems can learn from data, recognize suspicious patterns, and improve threat detection accuracy. This section discusses the major AI and ML techniques that are commonly used in cybersecurity applications.

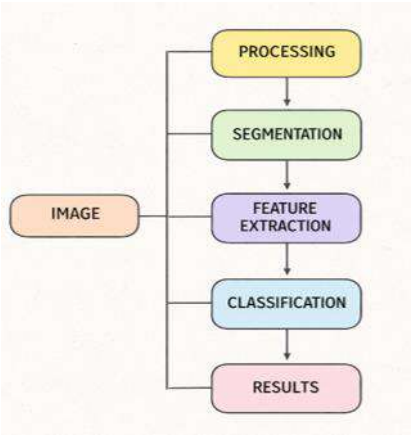


Figure 1: Architecture of AI-Driven Intelligent systems

a. Supervised Learning

Supervised learning uses labelled data to train models where the output is already known, such as identifying whether activity is normal or malicious [6]. Algorithms like decision trees, neural networks, random forests, and SVM are commonly used. These models are effective in tasks like malware detection, spam filtering, and intrusion detection by learning patterns from historical data.

b. Unsupervised Learning

Unsupervised learning works with unlabelled data to identify hidden patterns and unusual behaviour. Clustering techniques such as K-Means and DBSCAN help detect anomalies in network traffic or user activities [3]. This approach is particularly useful for identifying new or unknown cyber threats that do not match existing signatures.

c. Deep Learning

Deep learning uses multi-layer neural networks to analyse complex data and extract meaningful features. Models like CNNs and RNNs are widely used in malware detection, intrusion detection, and phishing identification [7]. These techniques can process raw data such as network logs and binary files, improving detection accuracy.

d. Natural Language Processing (NLP)

NLP enables systems to understand and analyse human language. In cybersecurity, it helps detect phishing emails, analyse suspicious messages, and monitor online threats. By examining text content, NLP models can identify malicious intent and prevent potential attacks [2].

e. Reinforcement Learning

Reinforcement learning improves cybersecurity defences by learning from interactions with the environment. It uses rewards and penalties to develop optimal

response strategies. This approach is useful in adaptive intrusion detection systems and automated threat response mechanisms [7].

f. Ensemble Learning

Ensemble learning combines multiple models to improve prediction accuracy. Techniques such as bagging and boosting reduce false alerts and enhance threat detection. This method is effective in handling large and complex cybersecurity datasets [5].

g. Feature Engineering and Selection

Feature engineering focuses on selecting relevant data attributes for training models. Techniques like PCA and feature selection methods help reduce data complexity and improve performance. Proper feature selection ensures better accuracy and faster detection of cyber threats.

These AI and ML techniques collectively enhance the capability of cybersecurity systems to detect and respond to evolving threats more efficiently [4].

Applications of Intelligent Systems in Cybersecurity

Intelligent systems based on Artificial Intelligence (AI) and Machine Learning (ML) are transforming modern cybersecurity by offering more adaptive and proactive defence solutions. These technologies help organizations detect threats faster, respond automatically, and improve overall security efficiency. By analysing large amounts of data and identifying unusual behaviour, intelligent systems enhance protection against evolving cyber-attacks. This section highlights the major application areas where AI and ML have made a significant contribution in cybersecurity [5].

a. Intrusion Detection and Prevention Systems (IDPS)

Traditional intrusion detection systems mainly depend on signature-based methods, which are less effective against new types of attacks. Intelligent systems use machine learning to detect unusual network behaviour and identify zero-day attacks or insider threats [8]. AI-based IDPS can also automatically block malicious traffic or isolate affected systems, reducing damage and response time.

b. Malware Detection and Classification

AI-powered systems analyse program behaviour, binaries, and network activity to detect different types of malwares. Deep learning models such as CNNs and RNNs help classify malware with higher accuracy, allowing quicker identification and removal of harmful software [6].

c. Phishing Detection and Prevention

Phishing is a common cyber threat targeting users through emails and fake websites. Intelligent systems use machine learning and NLP to examine email content, links,

and website features. This helps identify phishing attempts in real time and prevents data theft or financial fraud.

d. User and Entity Behaviour Analytics (UEBA)

UEBA systems monitor user and device behaviour continuously to detect suspicious activities. Machine learning models create normal behaviour profiles and highlight unusual actions such as abnormal login times or unauthorized data access. This helps in early detection of insider threats and compromised accounts. User behaviour analytics systems monitor user activities and detect anomalies, helping in identifying insider threats [3].

e. Threat Intelligence and Predictive Analytics

AI-based threat intelligence systems collect and analyse data from multiple sources such as social media, dark web platforms, and security databases. Machine learning models use this data to predict potential threats and vulnerabilities. This helps security teams take proactive measures and improve patch management [10].

f. Automated Incident Response

Intelligent systems improve incident response by automating tasks like alert analysis, log monitoring, and threat containment. Reinforcement learning allows systems to adapt their responses based on attacker behaviour. Automation reduces manual effort, minimizes errors, and speeds up threat mitigation.

g. Security Policy Management

AI helps organizations create and manage security policies dynamically. By analysing user roles, network activity, and emerging threats, intelligent systems can automatically update firewall rules, access controls, and security configurations. This ensures continuous protection against evolving risks.

h. Fraud Detection

AI-driven systems are widely used in banking and e-commerce to detect fraudulent activities. Machine learning models analyse transaction patterns and identify unusual behaviour in real time [8]. Early detection helps prevent financial losses and improves system security.

Challenges in Implementing Intelligent Cybersecurity Systems

Although AI and machine learning improve cybersecurity, their implementation comes with several challenges. Addressing these issues is important for building reliable and effective intelligent security systems.

a. Data Quality and Availability

AI models require high-quality data, and poor datasets can reduce system accuracy. However, collecting such data in cybersecurity is difficult due to privacy concerns

and constantly changing threats. Poor or incomplete datasets can reduce model accuracy and lead to false alerts [10].

b. Adversarial Attacks

Attackers may exploit weaknesses in AI models using techniques like data manipulation or evasion attacks. These methods can trick systems into treating malicious activities as normal. Protecting AI systems from such attacks remains a major challenge. Adversarial attacks can manipulate AI models and bypass detection systems [5].

c. Model Interpretability and Explainability

Many AI models work as black boxes, making it difficult to understand how decisions are made. Lack of model interpretability remains a major issue in AI-based cybersecurity systems [4]. Developing models that are both accurate and easy to interpret is essential.

d. Privacy and Ethical Concerns

AI-based cybersecurity systems handle sensitive personal and organizational data, which raises privacy and ethical concerns. Ensuring data protection, preventing misuse, and avoiding bias in models are important for maintaining trust and legal compliance.

e. Evolving Threat Landscape

Cyber threats continue to evolve with new techniques designed to bypass security systems. Intelligent cybersecurity solutions must be regularly updated and retrained to stay effective [3]. Without continuous improvement, AI-based defences may become outdated.

f. False Positives and Alert Fatigue

Although AI reduces false alerts, completely eliminating them remains difficult. Too many false alarms can overwhelm security teams and reduce efficiency. Accurate detection systems are needed to distinguish between normal activity and real threats.

Future Directions

The future of intelligent cybersecurity systems depends on continuous innovation and adaptation to emerging threats. AI models that support adaptive learning can improve their performance by learning from new data. Technologies like quantum computing may further enhance AI capabilities and strengthen security solutions. Collaborative defence, where organizations share threat intelligence, can improve overall protection. In addition, combining human expertise with AI systems can lead to more efficient and reliable cybersecurity strategies [9].

Additionally, research in explainable AI and automated security frameworks is expected to further improve transparency and response efficiency in future cybersecurity systems.

Conclusion

Artificial Intelligence and Machine Learning are transforming modern cybersecurity by providing adaptive and proactive defence mechanisms [2]. These technologies improve threat detection, automate response processes, and help organizations manage evolving cyber risks more effectively. Various AI techniques such as supervised learning, deep learning, and NLP enhance applications like intrusion detection, malware analysis, phishing prevention, and fraud detection.

However, implementing intelligent cybersecurity systems also involves challenges including data quality issues, privacy concerns, adversarial attacks, and false alerts. Addressing these limitations is essential for building reliable and trustworthy security solutions [9].

Overall, the integration of AI and ML in cybersecurity offers significant potential for strengthening digital protection [10]. With continuous research, adaptive learning, and collaboration between human expertise and intelligent systems, future cybersecurity frameworks can become more efficient, resilient, and capable of handling emerging threats.

Acknowledgment

The author would like to thank the Department of Computer Application, MLK PG College, Balrampur, for providing guidance and resources for this study.

References

1. J. Kim et al., “Deep learning-based intrusion detection system” (2020)
2. N. Shone et al., “Deep learning approach to network intrusion detection” (2018)
3. H. Hindy et al., “Taxonomy of network threats” (2020)
4. A. Sharmeen et al., “Explainable AI for cybersecurity” (2022)
5. L. Chen et al., “Adversarial machine learning in cybersecurity” (2023)
6. M. S. Islam et al., “Machine learning techniques for network security” (2019)
7. Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning” (2015)
8. F. A. Alshehri, “ML-based intrusion detection systems review” (2020)
9. Gurpreet k., Jyoti B. “Intelligent Systems in Cybersecurity” (2025)
10. S. Dua and X. Du, "Data Mining and Machine Learning in Cybersecurity," (2016)

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 79 - 87

Cyber Hygiene: A Habit for Safe Digital Living

Prof. Sunita Arya

Department of Zoology, Dayanand Girls P. G. College, Kanpur

Email:

Article DOI Link: https://zenodo.org/uploads/22029882

DOI: 10.5281/zenodo.22029882

Abstract

In an era defined by digital interconnectedness, cyber hygiene has emerged as a foundational discipline for individuals, organisations, and governments alike. Much like personal hygiene protects physical health, cyber hygiene refers to the routine practices and habits that protect digital assets, preserve privacy, and ensure the integrity of online systems. This article explores the core principles of cyber hygiene, common cyber threats facing users today, practical steps to strengthen one's digital posture, and the cultural shift required to embed safe digital behaviour into everyday life.

Keywords: Cyber hygiene, cybersecurity, digital safety, phishing, data protection, password management

Introduction

The Digital Age and Its Hidden Risks

The twenty-first century has ushered in an unprecedented era of digital transformation. From online banking and e-commerce to telemedicine and smart home devices, the digital ecosystem now permeates virtually every aspect of human life. Billions of people across the globe are connected through smartphones, laptops, tablets, and an ever-growing array of Internet of Things (IoT) devices. Yet, with this connectivity comes vulnerability, as every networked device represents a potential entry point for malicious actors (Cherdantseva & Hilton, 2013).

According to the World Economic Forum (WEF, 2024), cybercrime costs the global economy over USD 8 trillion annually, with projections suggesting this figure will exceed USD 10 trillion by 2025. In India alone, the Computer Emergency Response Team reported over 1.5 million cybersecurity incidents in 2022, with financial fraud, ransomware, and data breaches topping the list (CERT-In, 2023).

Despite the magnitude of this threat, research consistently shows that a large proportion of successful cyberattacks are preventable. Verizon (2023) found that

approximately 74% of all breaches involved the human element, including social engineering, errors, or misuse. This underscores the critical importance of cyber hygiene — systematic, habitual digital practices that dramatically reduce exposure to threats.

What Is Cyber Hygiene?

Cyber hygiene is the set of practices and behaviours adopted on a regular basis to maintain the health and security of digital environments. Rajamäki et al. (2018) define it as "the cyber equivalent of personal hygiene" — a collection of habitual, preventive actions that reduce the risk of compromise rather than responding to incidents after they occur.

The European Union Agency for Cybersecurity (ENISA, 2022) identifies cyber hygiene as encompassing password management, software patching, data backup, network security, and phishing awareness. At its core, cyber hygiene shifts security from a reactive posture to a proactive, prevention-first mindset.

Crucially, NIST (2018) emphasises that cyber hygiene is not the exclusive concern of IT professionals. Every person who uses a smartphone or accesses the internet has both a personal stake in and a shared responsibility towards maintaining good cyber hygiene, contributing to collective digital resilience.

Common Cyber Threats in the Modern World

Understanding the threat landscape is a prerequisite for practising effective cyber hygiene. The most prevalent threats identified by cybersecurity authorities are outlined below.

1. Phishing and Social Engineering

Phishing attacks use fraudulent emails, messages, or websites to deceive users into revealing sensitive credentials. Hadnagy (2018) describes social engineering broadly as the psychological manipulation of individuals into performing actions or divulging confidential information. The Anti-Phishing Working Group (APWG, 2023) reported over 4.9 million phishing attacks globally in 2022, the highest annual total ever recorded.

2. Malware and Ransomware

Malware encompasses any software designed to damage, disrupt, or gain unauthorised access to systems. Ransomware — which encrypts victim data and demands payment for its release — has emerged as one of the most disruptive cyber threats of the past decade. Kaspersky (2023) reported that ransomware attacks increased by 40% in 2022, with healthcare, education, and government sectors among the hardest hit.

3. Password Attacks and Credential Theft

Brute-force attacks, dictionary attacks, and credential stuffing — exploiting leaked username-password combinations — remain primary vectors for account compromise. Specops Software (2023) found that over 80% of hacking-related breaches involve compromised or weak passwords, reinforcing the centrality of password security in any cyber hygiene framework.

4. Man-in-the-Middle (MitM) Attacks

In MitM attacks, a threat actor secretly intercepts communications between two parties. Public Wi-Fi networks are a common vector for such attacks. Conti et al. (2016) demonstrated that unsecured wireless networks enable eavesdropping and session hijacking with relatively modest technical resources, highlighting the risk for users who access sensitive services over public hotspots.

5. Data Breaches and Identity Theft

Large-scale data breaches expose personal information of millions of users, which is subsequently sold on dark web marketplaces. IBM Security (2023) placed the average total cost of a data breach in 2023 at USD 4.45 million globally, the highest ever recorded. Identity theft resulting from breached data causes profound financial and reputational harm to individuals.

Core Principles of Good Cyber Hygiene

Practising cyber hygiene does not require advanced technical expertise. The following habits, consistently applied, constitute the foundation of a secure digital life.

1. Strong and Unique Passwords

Creating strong, unique passwords for every account is a fundamental cyber hygiene requirement. NIST (2017) guidelines recommend passwords of at least 12 characters, combining letters, numbers, and symbols, while avoiding predictable patterns. The use of password managers software tools that generate and store complex passwords securely is strongly advocated, eliminating the need to reuse passwords across platforms (Stobert & Biddle, 2014).

2. Multi-Factor Authentication (MFA)

Multi-factor authentication requires users to verify identity through more than one method, typically combining a password with a one-time code or biometric. Microsoft (2023) reports that MFA blocks over 99.9% of automated account compromise attacks. Enabling MFA on all critical accounts email, banking, and cloud storage is among the most impactful measures available to any user.

3. Regular Software and System Updates

Software vendors regularly release patches to address discovered security vulnerabilities. Delaying updates leaves systems exposed to exploitation of known

flaws. Symantec (2019) found that the majority of successful exploit-based attacks targeted vulnerabilities for which patches had been available for more than a year, underscoring the preventability of such incidents through timely patching.

4. Regular Data Backups

Maintaining regular, encrypted backups following the 3-2-1 rule three copies of data, on two different media types, with one stored offsite or in the cloud ensures recovery capability even after a ransomware attack or hardware failure. ENISA (2022) identifies data backup as a non-negotiable element of baseline cyber hygiene for both individuals and organisations.

5. Safe Browsing and Email Practices

Vigilance in online interactions forms an essential pillar of cyber hygiene. This includes verifying HTTPS before entering sensitive information, avoiding unsolicited links or attachments, and recognising phishing indicators such as misspelled domains and urgent, generic requests. Vishwanath et al. (2011) found that individuals with greater awareness of phishing cues are significantly less likely to fall victim to such attacks.

6. Secure Wi-Fi and Network Usage

Home networks should be secured with WPA3 or WPA2 encryption with strong administrator passwords, and default router credentials must be changed immediately (FTC, 2021). On public Wi-Fi, a Virtual Private Network (VPN) encrypts traffic and protects against eavesdropping. IoT devices should be segmented on a separate guest network to contain potential compromises.

7. Antivirus and Security Software

Installing and maintaining reputable antivirus and anti-malware software provides an important additional defence layer. AV-TEST (2023) evaluated over 20 leading security products, confirming that regularly updated antivirus software can detect and block over 99% of known malware samples. Firewalls and phishing-detection features embedded in modern security suites further enhance protection.

Cyber Hygiene in the Indian Context

India, with over 900 million internet users, presents enormous opportunities alongside significant cybersecurity challenges. MeitY (2023) reported that cyber fraud, particularly through digital payment platforms such as UPI, has risen sharply in recent years, with over 95,000 UPI fraud complaints registered in 2022–23 alone. The Indian government has enacted several landmark measures: the National Cyber Security Policy (2013), the establishment of CERT-In as the national nodal agency, and the Digital Personal Data Protection Act (DPDPA, 2023), which introduces obligations for data fiduciaries to implement reasonable security safeguards. The

Cyber Surakshit Bharat initiative has trained over 50,000 government officials in cybersecurity fundamentals (MeitY, 2022).

At the individual level, Indian internet users must be particularly alert to KYC update scams, fake customer service impersonation, and fraudulent UPI payment requests threats that exploit the rapid but uneven adoption of digital services. RBI (2023) has repeatedly emphasised that no legitimate bank or government authority will ever request an OTP or password, a message that must reach every segment of the digital population.

Cyber Hygiene in Educational Institutions

Academic institutions hold large volumes of sensitive data student records, financial information, and research outputs making them attractive targets. Educause (2023) identified cybersecurity as among the top five information technology concerns in higher education, with phishing and ransomware posing the greatest risks.

Institutional measures should include regular security audits, enforced multi-factor authentication for staff and student accounts, and mandatory cybersecurity awareness training. Faculty, as custodians of research and academic data, play a particularly important modelling role. Integrating cyber hygiene education into curricula especially in courses on Intellectual Property Rights and digital ethics creates structured opportunities to build student awareness (Furnell & Clarke, 2012).

Building a Culture of Cyber Hygiene

The most sophisticated security technology is rendered ineffective without corresponding human behaviour. Hadlington (2017) argues that individual cyber security awareness and attitude are significantly stronger predictors of risky online behaviour than technical controls alone, underscoring that cultural change is as essential as technical investment.

Key strategies for fostering a cyber hygiene culture include:

- Regular awareness campaigns tailored to different demographics — students, senior citizens, rural users, and professionals (ENISA, 2022).
- Incorporating digital safety into school and college curricula from an early stage (Furnell & Clarke, 2012).
- Mandatory cybersecurity training for all employees, not just IT staff, in organisations (NIST, 2018).
- Government and civil society promotion of cyber hygiene through mass media, regional languages, and community outreach (MeitY, 2023).
- Creating blame-free reporting environments so cyber incidents can be disclosed and learned from without stigma (WEF, 2024).
- Ultimately, cyber hygiene must become as automatic and non-negotiable as wearing a seatbelt. Wash (2010) found that people develop informal "folk

models" of security threats, and that education aligned with these mental models is far more effective in changing behaviour than purely technical instruction.

Emerging Challenges and Future Outlook

The cyber threat landscape continues to evolve rapidly. Artificial intelligence is being increasingly weaponised enabling the generation of highly convincing phishing emails, deepfake-based identity fraud, and automated vulnerability exploitation at unprecedented scale (Brundage et al., 2018). The proliferation of IoT devices has expanded the attack surface dramatically; Statista (2023) projects that the number of connected IoT devices will exceed 29 billion globally by 2030.

In response, cyber hygiene practices must evolve alongside threats. Users will increasingly need to adopt passkeys and biometric authentication as replacements for traditional passwords (FIDO Alliance, 2023)). Organisations must embed security by design into products and processes, reducing the burden placed on individual users. WEF (2024) calls for a whole-of-society approach, with governments, technology companies, academia, and civil society each playing defined roles in building collective cyber resilience.

Conclusion

Cyber hygiene is not a luxury or an optional concern. It is a fundamental necessity of contemporary digital life. As Rajamäki et al. (2018) observe, the digital world is only as secure as the collective habits of those who inhabit it. The good news, confirmed by cybersecurity research, is that consistent application of relatively simple practices prevents the overwhelming majority of attacks.

Strong passwords, multi-factor authentication, timely software updates, regular backups, safe browsing habits, and healthy scepticism of unsolicited requests form the non-negotiable core of cyber hygiene. Practised daily and embedded in organisational and educational culture, these habits translate individual security into collective resilience.

As educators, researchers, and digital citizens, we share responsibility for championing cyber hygiene not merely as a technical discipline but as a core value of safe, ethical, and responsible digital living. Furnell and Clarke (2012) argue that just as personal hygiene is taught as a fundamental life skill, cyber hygiene must be woven into education, workplaces, and communities. The habits we cultivate today will determine the security of the digital world we inhabit tomorrow.

References

1. Anti-Phishing Working Group. (2023). Phishing activity trends report: Full year 2022. APWG. https://apwg.org/trendsreports/
2. AV-TEST Institute. (2023). AV-TEST security report 2022/2023. AV-TEST GmbH. https://www.av-test.org/en/news/av-test-security-report/

3. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Eissler, S., Soares, N., Anderson, C., Petersen, M., & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv. https://arxiv.org/abs/1802.07228
4. CERT-In. (2023). Annual report 2022. Indian Computer Emergency Response Team, Ministry of Electronics and Information Technology, Government of India. https://www.cert-in.org.in
5. Cherdantseva, Y., & Hilton, J. (2013). A reference model of information assurance and security. In Proceedings of the 2013 International Conference on Availability, Reliability and Security (pp. 547–554). IEEE. https://doi.org/10.1109/ARES.2013.72
6. Conti, M., Dragoni, N., & Lesyk, V. (2016). A survey of man-in-the-middle attacks. IEEE Communications Surveys & Tutorials, 18(3), 2027–2051. https://doi.org/10.1109/COMST.2016.2548426
7. Digital Personal Data Protection Act. (2023). No. 22 of 2023. Government of India. https://www.meity.gov.in/content/digital-personal-data-protection-act-2023
8. Educause. (2023). Top 10 IT issues 2023. Educause. https://www.educause.edu/research-and-publications/research/top-10-it-issues-technologies-and-trends
9. European Union Agency for Cybersecurity. (2022). ENISA threat landscape 2022. ENISA. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022
10. FIDO Alliance. (2023). FIDO Alliance passkeys overview. FIDO Alliance. https://fidoalliance.org/passkeys/
11. Federal Trade Commission. (2021). Securing your wireless network. FTC Consumer Advice. https://consumer.ftc.gov/articles/securing-your-wireless-network
12. Furnell, S., & Clarke, N. (2012). Power to the people? The evolving recognition of human aspects of security. Computers & Security, 31(8), 983–988. https://doi.org/10.1016/j.cose.2012.08.004
13. Hadlington, L. (2017). Human factors in cybersecurity; examining the link between Internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours. Heliyon, 3(7), e00346. https://doi.org/10.1016/j.heliyon.2017.e00346
14. Hadnagy, C. (2018). Social engineering: The science of human hacking (2nd ed.). Wiley.
15. IBM Security. (2023). Cost of a data breach report 2023. IBM Corporation. https://www.ibm.com/reports/data-breach

16. Kaspersky. (2023). Kaspersky security bulletin 2022: Statistics. Kaspersky Lab. https://securelist.com/kaspersky-security-bulletin-2022-statistics/108129/
17. Microsoft. (2023). Microsoft digital defense report 2023. Microsoft Corporation. https://www.microsoft.com/en-us/security/security-insider/microsoft-digital-defense-report-2023
18. Ministry of Electronics and Information Technology. (2022). Cyber Surakshit Bharat initiative report 2022. Government of India. https://www.meity.gov.in
19. Ministry of Electronics and Information Technology. (2023). Annual report 2022–23. Government of India. https://www.meity.gov.in
20. National Institute of Standards and Technology. (2017). NIST special publication 800-63B: Digital identity guidelines — Authentication and lifecycle management. U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.800-63b
21. National Institute of Standards and Technology. (2018). Framework for improving critical infrastructure cybersecurity, version 1.1. U.S. Department of Commerce. https://doi.org/10.6028/NIST.CSWP.04162018
22. Rajamäki, J., Nevmerzhitskaya, J., & Virag, C. (2018). Cyber hygiene and cyber security teaching in schools. In Proceedings of the 2018 IEEE Global Engineering Education Conference (EDUCON) (pp. 1744–1751). IEEE. https://doi.org/10.1109/EDUCON.2018.8363448
23. Reserve Bank of India. (2023). Be(ware): RBI Kehta Hai awareness campaign. Reserve Bank of India. https://www.rbi.org.in
24. Specops Software. (2023). Weak password report 2023. Outpost24. https://specopssoft.com/blog/weak-password-report/
25. Statista. (2023). Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2030. Statista Research Department. https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/
26. Stobert, E., & Biddle, R. (2014). The password life cycle: User behaviour in managing passwords. In Proceedings of the 10th Symposium on Usable Privacy and Security (SOUPS 2014) (pp. 243–255). USENIX Association.
27. Symantec. (2019). Internet security threat report (Vol. 24). Broadcom Inc. https://www.broadcom.com/support/security-center/symantec-security-response
28. Verizon. (2023). 2023 data breach investigations report. Verizon Business. https://www.verizon.com/business/resources/reports/dbir/
29. Vishwanath, A., Herath, T., Chen, R., Wang, J., & Rao, H. R. (2011). Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. Decision Support Systems, 51(3), 576–586. https://doi.org/10.1016/j.dss.2011.03.002
30. Wash, R. (2010). Folk models of home computer security. In Proceedings of the 6th Symposium on Usable Privacy and Security (SOUPS 2010) (pp. 1–16). ACM. https://doi.org/10.1145/1837110.1837125

31. World Economic Forum. (2024). Global cybersecurity outlook 2024. World Economic Forum. https://www.weforum.org/publications/global-cybersecurity-outlook-2024/

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 88 - 95

Cyber Security in the Digital Age: Challenges and Strategic Solutions

[1]Sunita Arya

[2]Dr. Sadguru Prakash

[1]Professor, Department of Zoology, Dayanand Girls Post Graduate College, Kanpur, Uttar Pradesh, India

[2]Department of Zoology, MLK PG College, Balrampur, Uttar Pradesh, India

Email:

Article DOI Link: https://zenodo.org/uploads/22029995

DOI: 10.5281/zenodo.22029995

Abstract

The rapid proliferation of digital technologies has fundamentally transformed the global information landscape, bringing with it an unprecedented surge in cyber threats. This article examines the multifaceted challenges posed by cybersecurity vulnerabilities in the contemporary digital age and proposes strategic solutions to mitigate associated risks. Drawing on recent academic literature and policy frameworks, the article explores critical dimensions including ransomware, phishing, data breaches, Internet of Things (IoT) vulnerabilities, advanced persistent threats (APTs), and the human factor in cybersecurity failures. Strategic responses discussed include zero-trust architecture, artificial intelligence-driven threat detection, robust legislative frameworks, and cybersecurity education. The article underscores that effective cyber defence requires a holistic, multi-stakeholder approach integrating technology, policy, and human awareness.

Keywords: Cybersecurity, digital threats, ransomware, data breaches, zero-trust architecture, artificial intelligence, IoT security, cyber policy

Introduction

The digital revolution of the twenty-first century has catalysed extraordinary advances in commerce, communication, governance, education, and healthcare. However, this pervasive digitalisation has simultaneously created a vast and growing attack surface for malicious actors. Cybersecurity—the practice of protecting systems, networks, and data from digital attacks—has emerged as one of the most pressing concerns of the modern era (Anderson, 2020). Governments,

corporations, academic institutions, and individuals now face persistent and sophisticated threats that can result in financial loss, reputational damage, national security implications, and violations of personal privacy.

The global cost of cybercrime is projected to reach USD 10.5 trillion annually by 2025, representing the greatest transfer of economic wealth in history (Morgan, 2020). This staggering figure reflects not only direct financial theft but also the costs of data destruction, lost productivity, intellectual property theft, financial crimes, and the restoration of systems following attacks. The COVID-19 pandemic further accelerated digital adoption while simultaneously expanding opportunities for cybercriminals, as remote work arrangements created new vulnerabilities in organisational networks (World Economic Forum, 2021).

This article systematically analyses the principal cybersecurity challenges faced in the digital age and proposes strategic, evidence-based solutions. It is structured to address the threat landscape, category-specific vulnerabilities, human and organisational factors, and then transition to comprehensive mitigation strategies encompassing technical, legislative, and educational dimensions.

The Contemporary Cyber Threat Landscape

The cybersecurity threat landscape has evolved dramatically, shifting from opportunistic attacks by individual hackers to highly organised, state-sponsored cyber operations and professional criminal enterprises. Symantec's Internet Security Threat Report noted that over 4,800 websites per month were compromised by formjacking code, demonstrating the scale and sophistication of modern attacks (Broadcom, 2019). Threats today span ransomware, phishing, distributed denial-of-service (DDoS) attacks, supply chain attacks, and advanced persistent threats (APTs), each requiring distinct defensive strategies.

1. Ransomware

Ransomware has become one of the most devastating forms of cybercrime, with incidents involving critical infrastructure—including hospitals, pipelines, and government services—commanding global attention. Ransomware encrypts victim data and demands payment for decryption keys, often crippling entire organisations for days or weeks. The attack on the Colonial Pipeline in 2021 caused fuel shortages across the eastern United States, underscoring the real-world consequences of cyber intrusions on critical infrastructure (Easterly, 2021). Ransomware-as-a-Service (RaaS) platforms have further democratised these attacks, enabling even low-skilled criminals to deploy sophisticated malware.

2. Phishing and Social Engineering

Phishing remains the most prevalent initial attack vector, with the Anti-Phishing Working Group (APWG) reporting over one million phishing attacks per quarter in 2022 (APWG, 2022). Spear-phishing—targeted attacks directed at specific

individuals or organisations—exploits personalised information derived from social media and other open-source intelligence to deceive victims. Business Email Compromise (BEC) schemes, a sophisticated form of phishing, caused losses exceeding USD 2.7 billion in the United States alone in 2022 (FBI, 2022), highlighting the financial devastation that social engineering can produce.

3. Data Breaches and Privacy Violations

Data breaches represent a critical threat to individuals and organisations alike, exposing sensitive personal, financial, and proprietary information. The IBM Cost of a Data Breach Report (2023) found that the average cost of a data breach globally reached USD 4.45 million, the highest in the 18-year history of the report (IBM Security, 2023). Healthcare, finance, and education sectors are disproportionately targeted due to the high value of the data they hold. In India, the increasing volume of digital transactions and expansion of e-governance services have made data security a matter of national concern, reflected in the enactment of the Digital Personal Data Protection Act, 2023 (Ministry of Electronics and Information Technology, 2023).

4. Internet of Things (IoT) Vulnerabilities

The exponential growth of IoT devices—estimated to reach 75 billion globally by 2025—has introduced systemic vulnerabilities into networks previously considered secure (Statista, 2023). Many IoT devices are manufactured with minimal security features, using default credentials and lacking mechanisms for firmware updates. Botnets such as Mirai have demonstrated how compromised IoT devices can be weaponised to launch massive DDoS attacks, disrupting internet infrastructure globally (Krebs, 2016). As IoT integration deepens in healthcare, smart cities, agriculture, and manufacturing, the security implications become increasingly consequential.

5. Advanced Persistent Threats (APTs) and State-Sponsored Attacks

APTs represent the most sophisticated end of the threat spectrum, involving well-resourced actors—often state-sponsored—who conduct prolonged, covert intrusions into targeted networks. These operations aim to steal sensitive intelligence, disrupt critical systems, or conduct economic espionage. The SolarWinds supply chain attack of 2020, attributed to Russian state actors, compromised thousands of organisations worldwide including multiple United States federal agencies, illustrating the far-reaching consequences of APTs (Mandia, 2020). Such attacks highlight the intersection of cybersecurity with geopolitics and national defence.

The Human Factor in Cybersecurity

Despite advances in technical safeguards, the human element remains the most exploited vector in cyber-attacks. Research by Verizon's Data Breach Investigations Report consistently identifies human error as a contributing factor in over 80% of

data breaches (Verizon, 2023). This encompasses inadvertent disclosure of credentials, misconfiguration of cloud services, failure to apply security patches, and susceptibility to social engineering.

Insider threats—whether malicious or negligent—pose a particularly complex challenge. Employees with legitimate access to sensitive systems can cause substantial damage either through deliberate data exfiltration or through careless practices such as using weak passwords, sharing access credentials, or connecting to unsecured networks (Ponemon Institute, 2022). Organisations must therefore invest as heavily in workforce training and security culture as they do in technical infrastructure.

Strategic Solutions for Cybersecurity

1. Zero-Trust Architecture

The traditional perimeter-based security model—which assumes that entities inside a network can be trusted—has proven inadequate against modern threats. Zero-trust architecture (ZTA) operates on the principle of "never trust, always verify," requiring continuous authentication and authorisation of every user, device, and application regardless of network location (Rose et al., 2020). Adopted by the United States federal government as a cybersecurity mandate, ZTA significantly reduces the lateral movement of attackers within networks. Micro-segmentation, identity verification, and least-privilege access are core components that collectively reduce the attack surface.

2. Artificial Intelligence and Machine Learning in Threat Detection

Artificial intelligence (AI) and machine learning (ML) are transforming cybersecurity by enabling real-time detection of anomalous behaviour, predictive threat intelligence, and automated incident response. AI-powered security platforms can process vast volumes of network data at speeds impossible for human analysts, identifying patterns indicative of malicious activity with significantly greater accuracy (Sommer & Paxson, 2010). Natural language processing (NLP) models assist in detecting phishing emails, while behavioural analytics identify insider threats. However, AI also presents dual-use risks, as adversaries increasingly employ ML to craft more convincing phishing lures, evade detection systems, and automate attacks (Brundage et al., 2018).

3. Cybersecurity Legislation and Regulatory Frameworks

Robust legal and regulatory frameworks are essential to establish baseline security standards, enforce accountability, and facilitate cross-border cooperation in combating cybercrime. The European Union's General Data Protection Regulation (GDPR), enacted in 2018, has set a global benchmark for data protection, imposing significant penalties for non-compliance and mandating breach notification obligations (Voigt & Von dem Bussche, 2017). In India, the Information

Technology Act, 2000 (amended 2008) and the subsequent Digital Personal Data Protection Act, 2023, represent important legislative milestones, though enforcement mechanisms and technical capacity require further strengthening (Ministry of Electronics and Information Technology, 2023). International cooperation through frameworks such as the Budapest Convention on Cybercrime is vital for addressing the inherently transnational nature of cyber threats.

4. Cybersecurity Education and Workforce Development

Addressing the global shortage of cybersecurity professionals is a strategic imperative. It is estimated that there will be a shortage of 3.5 million cybersecurity professionals globally by 2025 (Cybersecurity Ventures, 2021). Universities and technical institutions must integrate cybersecurity into curricula across disciplines, recognising that digital literacy and basic cyber hygiene are relevant to all sectors of the economy. Security awareness training within organisations should be continuous, scenario-based, and tailored to role-specific risks. Gamified training platforms and simulated phishing exercises have demonstrated efficacy in improving employee vigilance (Aldawood & Skinner, 2019).

5. Incident Response and Cyber Resilience

No organisation can guarantee complete immunity from cyber-attacks; therefore, resilience and recovery capability are as important as prevention. A well-structured Incident Response Plan (IRP), regularly tested through tabletop exercises and red team simulations, enables organisations to contain breaches, mitigate damage, and restore operations efficiently (Cichonski et al., 2012). Business continuity planning, regular data backups adhering to the 3-2-1 rule (three copies, two media types, one offsite), and cyber insurance are essential elements of a comprehensive resilience strategy.

6. Securing the Internet of Things

Addressing IoT security requires action at multiple levels: manufacturers must implement security-by-design principles, incorporating strong default credentials, encrypted communications, and over-the-air update capabilities; regulators should establish and enforce minimum security standards for IoT devices; and end users must be educated on configuration best practices (Schneier, 2018). Network segmentation—isolating IoT devices from critical systems—limits the potential impact of device compromise. Emerging standards such as Matter (formerly Project CHIP) represent collaborative industry efforts to establish interoperable security baselines for consumer IoT.

Cybersecurity Challenges in the Indian Context

India, as the world's most populous nation and a rapidly growing digital economy, faces distinctive cybersecurity challenges. The proliferation of low-cost smartphones and expansion of digital financial services under the Digital India

initiative have brought hundreds of millions of first-time internet users online, many with limited awareness of cyber risks. CERT-In (the Indian Computer Emergency Response Team) recorded over 1.4 million cybersecurity incidents in 2022, reflecting the expanding threat environment (CERT-In, 2022).

Academic and research institutions, including universities and colleges, are increasingly targeted as repositories of valuable research data and as gateways to affiliated networks. Educational institutions must establish robust cybersecurity policies, conduct regular audits, and build staff and student awareness as part of their institutional risk management frameworks. The National Cyber Security Policy of India and the National Cyber Security Strategy (2020) provide a foundational framework, though implementation at the institutional level remains uneven (Ministry of Electronics and Information Technology, 2013).

Conclusion

Cybersecurity in the digital age is not merely a technical challenge—it is a complex socio-technical problem requiring coordinated responses from governments, the private sector, academic institutions, and individual citizens. The threats are diverse, continuously evolving, and increasingly consequential for economies, national security, and human rights.

Strategic solutions must be holistic: zero-trust architecture and AI-driven detection address technical vulnerabilities; robust legislation and international cooperation establish accountability and cross-border response capacity; and sustained cybersecurity education cultivates the human capital and organisational culture necessary for lasting resilience. As Schneier (2018) aptly observed, security is a process, not a product—it requires ongoing investment, adaptation, and commitment.

For developing nations including India, building indigenous cybersecurity capacity, strengthening regulatory institutions, and leveraging international partnerships are particularly urgent priorities. Universities and research institutions have a vital role to play—not only as subjects of cybersecurity governance but as active contributors to the knowledge base, policy discourse, and human capital pipeline that the digital age demands.

References

1. Aldawood, H., & Skinner, G. (2019). Reviewing cyber security social engineering training and awareness programs—Pitfalls and ongoing issues. Future Internet, 11(3), 73. https://doi.org/10.3390/fi11030073
2. Anderson, R. (2020). Security engineering: A guide to building dependable distributed systems (3rd ed.). Wiley.
3. Anti-Phishing Working Group (APWG). (2022). Phishing activity trends report, Q4 2022. APWG. https://apwg.org/trendsreports/

4. Broadcom (formerly Symantec). (2019). Internet security threat report (Vol. 24). Broadcom. https://www.broadcom.com/support/security-center/
5. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv. https://arxiv.org/abs/1802.07228
6. CERT-In. (2022). Annual report 2022. Indian Computer Emergency Response Team, Ministry of Electronics and Information Technology. https://www.cert-in.org.in
7. Cichonski, P., Millar, T., Grance, T., & Scarfone, K. (2012). Computer security incident handling guide (NIST Special Publication 800-61, Rev. 2). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-61r2
8. Cybersecurity Ventures. (2021). Cybercrime to cost the world $10.5 trillion annually by 2025. Cybersecurity Ventures. https://cybersecurityventures.com
9. Easterly, J. (2021). Statement by acting director of CISA on colonial pipeline cybersecurity attack. Cybersecurity and Infrastructure Security Agency (CISA). https://www.cisa.gov
10. Federal Bureau of Investigation (FBI). (2022). Internet crime report 2022. FBI Internet Crime Complaint Center (IC3). https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf
11. IBM Security. (2023). Cost of a data breach report 2023. IBM Corporation. https://www.ibm.com/reports/data-breach
12. Krebs, B. (2016, October 26). Who makes the IoT things under attack? Krebs on Security. https://krebsonsecurity.com
13. Mandia, K. (2020). FireEye shares details of recent cyber-attack, actions to protect community. FireEye (Mandiant). https://www.mandiant.com/resources/blog/evasive-attacker-leverages-solarwinds-supply-chain-compromises-with-sunburst-backdoor
14. Ministry of Electronics and Information Technology, Government of India. (2013). National cyber security policy 2013. Government of India. https://www.meity.gov.in
15. Ministry of Electronics and Information Technology, Government of India. (2023). The Digital Personal Data Protection Act, 2023. Gazette of India. https://www.meity.gov.in
16. Morgan, S. (2020). Cybercrime magazine: 2020 cybercrime report. Cybersecurity Ventures. https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/
17. Ponemon Institute. (2022). 2022 cost of insider threats: Global report. Ponemon Institute & Proofpoint. https://www.proofpoint.com/us/resources/threat-reports/cost-of-insider-threats

18. Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero trust architecture (NIST Special Publication 800-207). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207
19. Schneier, B. (2018). Click here to kill everybody: Security and survival in a hyper-connected world. W. W. Norton & Company.
20. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. Proceedings of the 2010 IEEE Symposium on Security and Privacy, 305–316. https://doi.org/10.1109/SP.2010.25
21. Statista. (2023). Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025. Statista. https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/
22. Verizon. (2023). 2023 data breach investigations report. Verizon Business. https://www.verizon.com/business/resources/reports/dbir/
23. Voigt, P., & Von dem Bussche, A. (2017). The EU General Data Protection Regulation (GDPR): A practical guide. Springer.
24. World Economic Forum. (2021). The global risks report 2021 (16th ed.). World Economic Forum. https://www.weforum.org/reports/the-global-risks-report-2021

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 96 - 104

A Review of AI-Enabled Systems for Monitoring and Preventing Malpractice in Online Examinations

Deepak Kumar

Mrs. Khushboo

Amity Institute of Information Technology, Amity University, Patna, Bihar, India

Email:

Article DOI Link: https://zenodo.org/uploads/22030416

DOI: 10.5281/zenodo.22030416

Abstract

COVID-19 changed everything for education almost overnight. Schools and colleges had no choice but to shift online, and with that shift came a big question: how do you make sure students are not cheating when they are sitting at home? Traditional invigilation simply does not work in that kind of setup, which is why there has been so much interest in using AI to proctor online tests. This paper reviews ten research studies on AI-based proctoring systems published between 2022 and 2025, across journals like IEEE, Elsevier, and Springer. These systems use technologies like computer vision (YOLO, MTCNN, MediaPipe), deep learning (CNN, LSTM, ALBERT), audio analysis, natural language processing, and blockchain-based security. We go through how each system detects cheating, what accuracy it achieves, and where it falls short. Across the board, the same problems keep coming up: privacy concerns, algorithmic bias, high hardware requirements, limited real-world testing, and the growing worry about students using AI tools like ChatGPT to write their exam answers. The paper ends by suggesting some useful directions for future work: lighter AI models, emotion-aware monitoring, federated learning for privacy, and explainable AI so that when a student is flagged, there is a clear reason given.

Keywords: Online exam proctoring, AI, computer vision, YOLO, deep learning, cheating detection, academic integrity, facial recognition, behavioural analysis, natural language processing.

Introduction

Online education saw a massive jump when COVID-19 hit. Almost overnight, institutions had to move their classes and exams to the internet. And very quickly, it

became clear that the way exams were monitored needed a serious rethink. A student sitting at home with a laptop and easy access to notes, phones, or the internet is a very different monitoring challenge compared to a student sitting in an exam hall under a supervisor's eye.

In a physical exam, an invigilator can walk around the room and step in the moment something looks off. Online, without technology, that level of oversight just is not possible. Having a human watch live webcam feeds is expensive and hard to scale. Attention wanders, it is difficult to monitor many students at once, and different people may react differently to the same suspicious behaviour.

AI-based proctoring tries to solve these problems by automating the whole process. These systems can pick up on a wide range of suspicious behaviour using just webcam footage and microphone input. They can flag someone looking away from the screen, a phone appearing in the frame, another person walking into view, unusual sounds, or even someone holding up a printed photo instead of their real face. More advanced systems go further by tracking eye movements, head angles, the student's apparent mood, and even their typing patterns.

This paper looks at ten research studies on AI-based exam proctoring published between 2022 and 2025. The aim is to understand what methods have actually been tried, how well they perform, and where the real problems are. Section II explains why this review is needed. Section III goes through each system in detail. Section IV has a summary table comparing all of them. Sections V and VI cover the main challenges and future directions. Section VII concludes the paper.

Research Gap and Motivation

Even with the growing interest in AI proctoring, there is still no system that manages to be thorough, accurate, privacy-friendly, and practically usable all at once. A look through the current literature consistently reveals the same issues:

- Most systems only check for one or two types of cheating, leaving many ways to cheat undetected.
- Testing is almost always done in controlled lab settings with a small number of participants. Very few studies have tested these systems in actual real-world conditions.
- Continuous audio and video surveillance raises serious privacy concerns, especially in regions with data protection laws.
- Many systems require hardware that students in rural areas or developing countries simply cannot afford.
- Exam anxiety can look very similar to suspicious behaviour to a machine. Almost none of these systems try to tell the difference.
- Students using AI tools like ChatGPT to generate exam responses is a growing problem that visual proctoring systems are completely unprepared to handle.

This review is meant to map out where the field actually stands right now, what is working, and where research needs to go next. It should be useful both for researchers building new systems and for institutions trying to figure out which tools are worth adopting.

Literature Review

YOLO-Based Visual Proctoring Systems

Singh et al. [1] built a proctoring system in Python that combines several computer vision libraries into one monitoring pipeline. OpenCV handles the video, MediaPipe picks up facial landmarks and estimates head orientation, TensorFlow manages person counting and spoof detection, and PyAudio listens for sounds. The system monitors six things: face orientation, spoof detection, mouth movement, unexpected audio, mobile phone presence via YOLOv3, and person count. If a phone or an unauthorised person stays visible for ten continuous seconds, the exam terminates automatically.

Testing on 70 pre-recorded clips showed strong results overall. But there was one notable issue: when a photograph was held up to the camera, the system correctly flagged it as a spoof but also counted it as a real person. That gap between the spoofing and person-counting modules is a clear weakness worth fixing.

Negi et al. [3] combine YOLOv3 with Dlib's face detector and its 68-point landmark model. Their system monitors face identity, multiple simultaneous faces, phone presence, browser tab switching, audio levels, and head position. On an AMD Ryzen 7 5600H, the system achieved 15.47 FPS for eye tracking and 17.55 FPS for head pose estimation. However, combined phone-and-person detection dropped to 7.47 FPS on the same machine.

Malhotra et al. [10] add emotion recognition on top of the usual detection tasks. A CNN trained on the FER2013 dataset, which has 35,887 labelled images across seven emotion categories, is used to infer whether a student's facial expression suggests anxiety or suspicious intent. YOLOv3 handles phone detection and also book detection, which most other reviewed systems ignore entirely.

Multi-Modal and Deep Learning Systems

Subrahmanyam and Balabhadrapatruni [4] built a seven-component system that includes something fairly unusual: distance estimation. It brings together MTCNN-based face detection, eye gaze tracking, emotion classification via a mini-Xception CNN, head pose estimation via Dlib's 68-point model, distance estimation using a focal-length formula, background noise detection, and a final inference engine. The system achieved 97% face detection accuracy and 99.3% face recognition accuracy, comparing well against commercial tools like Proctor U and Kryterion.

Sharma and Pant [7] present probably the most technically advanced work in this review. Starting from a baseline SVM system across six modalities, they replace it with three classifiers working together. A CNN handles the visual modalities at 91-94% accuracy. An LSTM processes sequential audio and gaze data, pushing speech detection to 93% and reducing false alarms. A Random Forest handles noisy ensemble classification at 89.5%. The three models are combined via majority voting, giving a True Detection Rate of 94% at a 2% False Alarm Rate, which is a 7-point improvement over the SVM baseline of 87%.

Behaviour Analysis and Integrated Frameworks

Pansare et al. [5] introduce a trust scoring system that turns student behaviour into a number in a transparent and auditable way. The system has five modules: facial recognition using Dlib's 128-dimensional face embeddings, head pose estimation via MediaPipe's Face Mesh, multi-person detection, device detection using YOLOv8 on COCO, and voice activity detection. Each violation type carries a weighted penalty. Facial recognition violations are weighted at 2.0x, head movement at 1.0x. A starting trust score of 100 decreases with each infraction; drop below 70 and the student is flagged. Accuracy reported: 94.2% for facial verification, 89.7% for head movement detection, 91.3% for device identification.

Patil et al. [6] focus specifically on eye and hand movement, using MediaPipe's Face Mesh with 468 facial landmarks and hand tracking with 21 keypoints per hand. Eye deviation beyond 25 degrees triggers an alert; six cumulative warnings and the exam auto-submits. Testing with 20 volunteers gave F1-scores of 91.4% for eye movement and 93.7% for hand movement. Analysing every other frame cuts GPU load by roughly 50% with no measurable drop in accuracy.

Kale et al. [8] embed proctoring inside a broader learning environment, combining a YOLOv8-based exam monitor with an NLP-powered chatbot for GATE exam preparation. The chatbot handles subject queries in natural language while the proctoring module detects books, phones, and multiple faces at over 94% accuracy.

Real-Time AI Frameworks with Ethical Considerations

Satpute and Babasaheb [2] propose a multi-layer architecture that stands out for integrating blockchain for tamper-proof data logging. Exam records are written to a blockchain ledger, making them resistant to post-exam manipulation. The NLP component monitors for clipboard misuse, unusual copy-paste activity, and typing irregularities. Privacy protections include AES-256 video encryption and TLS 1.3 for database communication. The main caveat is that this is largely a proposed framework, and specific accuracy figures are not reported.

AI-Generated Text Detection in Online Interviews

Prasethio et al. [9] tackle a completely different cheating problem: students using tools like ChatGPT to generate responses in automated text-based recruitment

interviews. The paper uses ALBERT (A Lite BERT), chosen for its lower memory footprint and strong contextual understanding. Twenty-five linguistic features are computed per response, including POS tag frequencies, VADER sentiment scores, readability indices, text perplexity, burstiness, and cosine similarity with GPT-3.5, GPT-4o, and Gemini outputs. ALBERT achieved 81% accuracy, 87% precision, 79% recall, and an 82% F1-score. The biggest limitation is the dataset: just 60 entries from 10 participants.

Comparative Analysis

Table I provides a side-by-side comparison of all reviewed systems. YOLO-based object detection (versions 3 and 8) is the most widely used framework, appearing in eight of the ten papers. Face detection accuracy across systems ranges from roughly 90% to 99%. Hybrid deep learning systems, especially CNN+LSTM combinations, outperform SVM-based baselines in multimodal settings. No single system covers all dimensions of online cheating.

TABLE I. Comparative Analysis of Reviewed AI Proctoring Systems

Author & Year	Technique	Key Features	Accuracy / Result	Limitation
Singh et al. (2024)	YOLOv3 + MediaPipe + PyAudio	Face orientation, spoofing, mouth tracking, phone/person detection	Auto-terminates on 10-sec violation	Spoof miscounted as real person
Satpute & Babasaheb (2025)	Multi-layer AI + Blockchain + NLP	Facial recog., behaviour analysis, blockchain logging	Reduced malpractice; improved oversight	Privacy risks; algorithmic bias
Negi et al. (2024)	YOLOv3 + MMOD CNN + Dlib	Face ID, multi-face, head pose, tab-switch, voice	Up to 17.55 FPS (Ryzen 7)	CNN needs GPU; not real-time on CPU
Subrahmanyam et al. (2024)	MTCNN + Dlib + Emotion CNN	Gaze, emotion, distance estimation, noise, head pose	97% face detect; 99.3% recognition	High resource usage
Pansare et al. (2025)	YOLOv8 + MediaPipe + face_recognition	Trust score, device detection, voice activity	94.2% facial; 89.7% head; 91.3%	Hardware-dependent; cultural bias

			device	
Patil et al. (2025)	OpenCV + MediaPipe (eye & hand)	Eye/hand movement, auto-screenshot, 6-warning limit	F1: 91.4% eye; 93.7% hand	No audio monitoring
Sharma & Pant (2025)	CNN + LSTM + Random Forest	Multimodal: speech, gaze, text, phone, window	94% TDR at 2% FAR	High compute; privacy concerns
Kale et al. (2025)	YOLOv8 + NLP Chatbot	Object detection, AI chatbot, study resources	Over 94% detection accuracy	Internet-heavy; limited dataset
Prasethio et al. (2024)	ALBERT + Logistic Regression	POS, sentiment, perplexity, typing behaviour	81% accuracy, 82% F1	Only 60 interview samples
Malhotra et al. (2022)	CNN (FER2013) + YOLOv3 + Dlib	Emotion, head pose, book/phone/multi-person detection	Works in real-time webcam feed	No audio; needs human backup

Key Challenges

1. Privacy and Ethical Concerns

Every system reviewed collects continuous video and audio from students during exams. That raises real privacy concerns, particularly where laws like GDPR, FERPA, or India's Digital Personal Data Protection Act apply. Monitoring students inside their own homes brings up difficult questions about consent, how long data is stored, and the psychological impact of sustained surveillance on exam performance. A widely accepted ethical framework for AI proctoring still does not exist.

2. Algorithmic Bias and Fairness

AI models trained on datasets that are not demographically representative can produce systematically unfair results. Emotion detection is especially vulnerable if training data skews toward a particular ethnic group. Head pose estimation may also generate more false positives for students from cultural backgrounds where different head movement patterns are normal during concentration.

3. Hardware and Infrastructure Limitations

Most systems need a decent webcam, stable broadband, and enough CPU or GPU power for real-time inference. That is not a given for students using budget devices in rural or developing regions. Negi et al. [3] show in their results that FPS drops significantly on lower-end processors, which directly hurts detection reliability.

4. Adversarial Attacks and Spoofing

Singh et al. [1] document a real failure in their own system: a printed photograph held up to the camera was flagged as a spoof but was simultaneously counted as a real person. That dual-classification error is a form of adversarial vulnerability. Reliable liveness detection, consistently telling a live face apart from a static image or video playback, is still an open problem.

5. Scalability and Integration

Running seven or more detection modules at the same time is resource-intensive. Subrahmanyam and Balabhadrapatruni [4] use parallel CPU threading to improve throughput but acknowledge high overall resource consumption. Scaling these systems to support thousands of students in national-level exams simultaneously is an engineering challenge that has not been solved yet.

Research Gaps and Future Scope

Several gaps kept coming up across the reviewed literature and they are worth taking seriously:

- **Emotion-Aware Proctoring:** Most current systems cannot tell the difference between exam anxiety and actual cheating. Adding emotional intelligence would cut false positives and make the process fairer.
- **Lightweight Models for Low-End Devices:** Architectures like MobileNet, TinyYOLO, or edge-AI solutions built for on-device inference would open proctoring up to a much wider student population.
- **AI-Generated Content Detection:** Using large language models to produce exam responses is a growing and underaddressed problem. Larger, more diverse datasets and stronger detection methods are urgently needed.
- **Federated Learning for Privacy:** Processing data locally on the student's device and sending only detection flags to the server would reduce privacy risks significantly without hurting monitoring quality.
- **Multi-Language and Multi-Cultural Support:** Cross-linguistic and cross-cultural validation is essential before these tools can be deployed at a global scale.
- **Longitudinal and Large-Scale Studies:** Real-world validation across thousands of students over multiple exam sessions is needed to assess genuine reliability.

- **Explainable AI (XAI):** When a system flags a student, it should be able to clearly say which behaviour triggered the alert. This is critical for fairness, transparency, and the student's right to appeal.

Conclusion

This paper reviewed ten AI-based online exam proctoring systems published between 2022 and 2025. Taken together, they show that AI technologies including YOLO-based object detection, CNN-based emotion recognition, LSTM-based speech analysis, MediaPipe landmark tracking, and ALBERT-based text analysis can detect a wide range of cheating behaviours with real accuracy.

Systems that combine multiple detection modalities consistently outperform single-method approaches. Sharma and Pant's [7] hybrid CNN+LSTM+Random Forest system achieved the highest overall detection rate (94% TDR at 2% FAR), while Pansare et al.'s [5] trust score approach offers the most interpretable way to report violations to administrators.

That said, there is still considerable work to do before these systems can be deployed widely and ethically. Privacy risks, demographic bias, hardware barriers, and the absence of large-scale real-world testing are recurring problems. And the emerging challenge of AI-generated text responses, which visual proctoring tools are simply not built to handle, will only get more pressing as large language models become more capable and accessible.

Future work should focus on systems that are lightweight enough to run on affordable hardware, privacy-preserving by design, sensitive to emotional states to reduce false positives, and transparent enough in their reasoning that students can challenge a decision. Getting there will require collaboration across AI research, educational institutions, ethics bodies, and policymakers.

References

1. T. Singh, R. R. Nair, T. Babu, and P. Duraisamy, "Enhancing Academic Integrity in Online Assessments: Introducing an Effective Online Exam Proctoring Model using YOLO," Procedia Computer Science, vol. 235, pp. 1399-1408, 2024.
2. S. G. Satpute and T. S. Babasaheb, "Real-time AI solutions for preventing academic cheating and malpractices in examinations," International Journal of Advanced Scientific Research, vol. 10, no. 3, pp. 72-75, 2025.
3. D. Negi, A. Bhandari, A. Gaur, A. Sindhwal, R. Chauhan, and A. Kapruwan, "AI-based Online Proctoring System with YOLOv3 & MMOD CNN," in 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, Apr. 2024.
4. S. BHVSP and C. Balabhadrapatruni, "AI based Online Proctoring Remote Monitoring Intruder, Emotion Detection and Distance Estimation," IJSREM, vol. 8, no. 8, 2024.

5. J. R. Pansare, A. Pawar, A. Chorghade, S. Barge, and A. Agarwal, "Proctoring Using AI," IJRASET, vol. 13, no. VI, pp. 1825-1835, June 2025.
6. A. V. Patil, P. A. Chavan, S. R. Jadhav, A. U. Phodkar, M. B. Gulame, and A. P. Pimpalkar, "Online exam proctoring application using AI," IJSRA, vol. 15, no. 2, pp. 1228-1234, 2025.
7. H. S. Sharma and V. K. Pant, "AI-Driven Online Exam Proctoring: An Enhanced Machine Learning Approach," JRICST, vol. 2, no. 4, pp. 52-65, 2025.
8. A. Kale, S. Nalawade, S. Gangule, V. Lakare, and J. M. Patade, "Smart GATE Exam Preparation System with AI-Based Object Detection and Monitoring," IJARSCT, vol. 5, no. 3, pp. 56-61, Nov. 2025.
9. Y. I. Prasethio, A. Dewandaru, and G. A. P. Saptawati, "Development of a Text-Based Anti-Cheat Component Using ALBERT for Detecting Cheating in Automated Interviews," in 2024 IEEE International Conference on Data and Software Engineering (ICoDSE), 2024, pp. 96-101.
10. N. Malhotra, R. Suri, P. Verma, and R. Kumar, "Smart Artificial Intelligence Based Online Proctoring System," in 2022 IEEE Delhi Section Conference (DELCON), 2022.
11. A. Nigam, R. Pasricha, T. Singh, and P. Churi, "A systematic review on AI-based proctoring systems: Past, present and future," Education and Information Technologies, vol. 26, pp. 6421-6445, 2021.
12. J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," arXiv preprint arXiv:1804.02767, 2018.
13. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499-1503, 2016.
14. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for self-supervised learning of language representations," arXiv preprint arXiv:1909.11942, 2019.
15. Y. Atoum, L. Chen, A. X. Liu, S. D. H. Hsu, and X. Liu, "Automated online exam proctoring," IEEE Transactions on Multimedia, vol. 19, no. 7, pp. 1609-1624, 2017.

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 105 - 117

Digital Forensics and Cyber Crime Investigation: Methodologies, Legal Frameworks, and Emerging Paradigms

Kavyansh Tiwari

Kshitij Mishra

Arman Khan

Sulabh Mishra

U.G. Students (6th Semester, B.Sc. Computer Science), M.L.K. (P.G.) College, Balrampur (U.P.)

Email:

Article DOI Link: https://zenodo.org/uploads/22030477

DOI: 10.5281/zenodo.22030477

Abstract

As global digital transformation accelerates, the socio-economic impact of cybercrime has reached macroeconomic proportions, with projected annual costs scaling to $23 trillion by 2027. This paper presents a comprehensive review of the modern digital forensics landscape, examining the sophisticated attack vectors deployed by modern threat actors, including AI-driven ransomware and IoT botnets, with a specific focus on the acute cybercrime crisis and recovery efforts within the Indian context and Uttar Pradesh. To combat these threats, the paper analyzes standardized incident response frameworks (NIST and ISO/IEC 27037) and advanced extraction methodologies, ranging from software-level cloud parsing to invasive hardware interventions like JTAG, ISP, and Chip-Off techniques. Furthermore, the study explores the parallel evolution of anti-forensics and evaluates the monumental shift in Indian electronic evidence jurisprudence following the implementation of the Bharatiya Sakshya Adhiniyam (BSA), 2023. Ultimately, this research underscores that mitigating future cyber threats requires a synchronized advancement of technical methodologies, international jurisdictional cooperation, and unyielding ethical standards.

Keywords: Digital Forensics, Cybercrime, Bharatiya Sakshya Adhiniyam 2023, Anti-Forensics, JTAG Extraction, Incident Response.

Introduction

Digital Forensics and Its Historical Evolution

While digital forensics traces its roots to the late 1970s and the FBI's establishment of the Computer Analysis and Response Team (CART) in 1984, early methodologies were largely ad-hoc. Throughout the 1990s and 2000s, the rapid proliferation of the internet and subsequent rise in cybercrime necessitated a much stricter, formal approach. Organizations like the National Institute of Standards and Technology (NIST) helped transition the field from a niche technical endeavor into a rigorous, legally validated scientific methodology. Today, digital forensics has evolved to address a hyper-connected landscape, encompassing specialized branches like mobile, cloud, and IoT investigations to combat increasingly sophisticated, automated threats.

The Cyber Threat Landscape and Socio-Economic Impact (2024–2026)

The socio-economic impact of cybercrime has reached macroeconomic proportions, fundamentally threatening global financial stability, national security, and corporate integrity.[1] Estimates from the International Monetary Fund (IMF) and industry analysts indicate that the global cost of cybercrime is projected to reach $10.5 trillion annually by 2025, potentially scaling to an unprecedented $23 trillion by 2027—a 175% increase from 2022 figures.[3] This escalation is driven by a high-speed artificial intelligence (AI) arms race between threat actors and defenders, the ubiquitous integration of digital platforms, the shift toward flexible remote work environments, and the increasing exploitation of decentralized cloud architectures.

The World Economic Forum's Global Cybersecurity Outlook 2025 highlights that cyber-enabled fraud and identity theft currently rank as the highest organizational and personal cyber risks for Chief Executive Officers and Chief Information Security Officers.[2] Threat actors now deploy autonomous AI agents capable of conducting reconnaissance, exploiting vulnerabilities, and moving laterally through networks at machine speeds, launching full-scale compromises within minutes.

Dominant Attack Vectors and Case Studies

Modern threat actors deploy a convergence of financial and state-aligned operations. Ransomware remains an existential threat, with groups utilizing double-extortion techniques against critical infrastructure and global supply chains. Phishing and social engineering, increasingly supercharged by AI-generated deepfakes, remain prevalent entry vectors, costing organizations billions annually. Concurrently, vulnerable IoT devices are hijacked to form botnets and mine cryptocurrency. To combat these distributed threats, international task forces, such as INTERPOL, have increasingly mobilized large-scale operations to dismantle malicious networks and seize illicit assets.

Regional Impact: The Indian Context

India, featuring a rapidly expanding digital economy propelled by hundreds of millions of daily digital financial transactions, faces an acute cybercrime crisis. Official data indicates that high-value cyber fraud cases jumped more than four times in the 2024 fiscal year, causing immediate losses of approximately $20 million.[4] Broader economic estimates suggest that total cybercrime costs in India could reach Rs 20,000 crore in 2025, prompting the government to increase its cybersecurity budget to ₹1,900 crore, with significant allocations for the Computer Emergency Response Team (CERT-In) and capital projects.[5] Cybersecurity incidents in the country escalated dramatically from 1.02 million in 2022 to 2.26 million in 2024, driven by the fact that over 86% of households are now connected to the internet.

The digital forensic market in India is expanding commensurately as a strategic enabler of trust and national resilience. According to a 2025 report by Deloitte India and the Data Security Council of India (DSCI), the domestic digital forensics market is valued at approximately INR 1,603 crore (USD 0.19 billion) and is anticipated to grow to INR 11,829 crore (USD 1.39 billion) by FY 2029–30. Driven by the convergence of smartphone usage and digital payments, mobile forensics currently dominates the landscape with a 51% to 55% market share.[6] The government is the largest end-user, accounting for 81% of the market demand, primarily for law enforcement purposes. However, the sector faces a projected shortfall of 90,000 forensic professionals, highlighting a critical human capital challenge.

In response to the surge in cybercrime, state-level law enforcement has aggressively modernized its investigative infrastructure. The state of Uttar Pradesh (UP), which established India's first exclusive cyber police stations in Lucknow and Gautam Buddha Nagar (Noida) in 2016, recently expanded this infrastructure to all 18 Police Range districts under the supervision of the Cyber Crime Headquarters in Lucknow. Between 2017 and mid-December 2025, UP police registered 87,850 cybercrime cases, initiating action against 53,639 accused individuals. By leveraging the "golden hour" of crime reporting via the national 1930 helpline, UP authorities successfully froze ₹325.25 crore in defrauded funds during 2025, capturing roughly 24% of total stolen capital—a sharp rise from an 11% recovery rate in 2024.[8] In January 2026 alone, UP received 29,715 cyber fraud complaints involving losses of ₹138 crore, of which ₹48.45 crore was secured by placing liens on suspected bank accounts, elevating the state to the third position nationally in terms of freezing defrauded money. High-profile interventions include thwarting "digital arrest" extortion scams, where fraudsters impersonating law enforcement psychologically confine victims via video calls to extort life savings. The Central Bureau of Investigation (CBI), acting on intelligence from the FBI, also dismantled

a sophisticated Noida-based virtual-asset cyber-fraud syndicate that had defrauded U.S. citizens of $8.5 million.

Standardized Digital Forensics Lifecycles

To assure legal admissibility and scientific validity, investigators rely on globally recognized frameworks. The NIST framework outlines a rigorous four-phase incident response: Preparation, Detection and Analysis, Containment/Eradication/Recovery, and Post-Incident Activity. Complementing this operational response is the ISO/IEC 27037 standard, which dictates strict guidelines for the identification, collection, and preservation of digital evidence. The core philosophy of the ISO standard is maintaining absolute "procedural consistency" to ensure a rigorous chain of custody that withstands judicial scrutiny. To achieve this, investigators rely heavily on hardware write-blockers to create exact bit-for-bit replicas of storage media, allowing analysis to proceed on the copy while preserving the original evidence.[10]

Advanced Forensic Extraction Methodologies and Software Platforms

The proliferation of complex encryption, the transition to cloud-based architectures, and the sheer diversity of hardware form factors necessitate specialized data extraction techniques. Modern investigations rely heavily on comprehensive software suites capable of acquiring logical and physical images, parsing vast repositories of unstructured data, executing automated timeline analyses, and leveraging heuristic algorithms.

Investigative Software Platform	Core Specialization and Deployment	Supported Ecosystems	Automated Analysis and Parsing Features
Magnet AXIOM Cyber	Deep-dive artifact parsing, remote data collection, and cloud deployment.	Windows, Linux, macOS, iOS, Android, Cloud platforms.	AI-driven image and chat review, Hash verification, YARA, and Sigma rule application.
Cyber Triage	Rapid incident response, endpoint triage, agentless remote data collection.	Windows, Linux.	Automated scoring of items as "Suspicious" or "Bad" using Heuristics, YARA, Sigma (via Hayabusa), and AV scanning.

Nuix Workstation	Handling massive, complex datasets and high-speed unstructured data searching.	Over 1,000 file types across diverse enterprise platforms.	User-scripting, data-link analysis, and interactive data visualizations for relationship mapping.
Cellebrite UFED	Industry-standard mobile device data extraction and decoding.	iOS, Android, embedded mobile systems, non-proprietary devices.	Logical, file system, and advanced physical extractions; integrated AI-driven automation for evidence identification.
Oxygen Forensic Detective	Advanced mobile extraction, cloud analytics, and workflow management.	iOS, Android, diverse cloud repositories.	Reliable data recovery, encrypted app data decryption, and advanced analytics for collaborative workflows.
Autopsy	Premier open-source forensic analysis platform.	Windows, Linux, macOS, Mobile devices.	Hash filtering, keyword searching, data carving, and robust timeline analysis.

The selection of a forensic tool is dictated by the specific requirements of the investigation. For forensic imaging, examiners utilize tools like FTK Imager and Guymager (designed for Linux environments) in conjunction with hardware write-blockers (such as those produced by Tableau) to prevent write operations to the evidence device, ensuring absolute data integrity. For network forensics, investigators rely on full packet capture programs like Tcpdump and Wireshark. By 2025, Wireshark features enhanced protocol support for high-speed networks and the critical ability to decrypt SSL/TLS traffic for analyzing encrypted communications. For cloud data extraction, platforms like BlackBag BlackLight and cloud-specific modules within Magnet AXIOM provide advanced capabilities to retrieve and analyze data residing in distributed cloud storage services. Mobile forensics is largely dominated by MSAB XRY, Cellebrite UFED, and Oxygen Forensic Detective, which offer specialized capabilities to breach sophisticated encryption, restore deleted data from messaging applications, and extract comprehensive location histories.

Hardware-Level Embedded Systems Forensics: JTAG, ISP, and Chip-Off

When software-based extraction fails due to catastrophic physical damage or advanced hardware-bound encryption (e.g., Secure Boot), investigators must escalate to invasive hardware interventions to interact directly with a device's memory chips.

JTAG (Joint Test Action Group) & ISP (In-System Programming)

JTAG utilizes motherboard test ports to bypass OS locks and instruct the functional CPU to dump raw memory.[12] Because modern manufacturers frequently disable JTAG to harden security, ISP is now the prevalent technique. ISP bypasses the CPU entirely by connecting directly to the memory chip's data lines on the circuit board, enabling physical data recovery without removing the chip.

Chip-Off Forensics

Reserved for severely damaged devices, this destructive, last-resort method involves surgically desoldering the memory chip from the motherboard.[11] The chip is then cleaned and read using specialized adapters and flasher boxes (such as Z3X Easy-JTAG Plus or UFi Box).

The Anti-Forensics Paradigm

As forensic capabilities mature, threat actors increasingly deploy anti-forensic techniques to destroy, hide, or manipulate evidence, thereby inducing "trail obfuscation."[13]

- **Evidence Destruction (Secure Wiping):** Unlike standard OS file deletion, cybercriminals use artifact wiping tools to actively overwrite physical storage space with complex algorithmic patterns, permanently destroying the underlying binary data. Investigators counter this by scanning for large blocks of intentionally overwritten sectors.
- **Evidence Hiding (Steganography):** This technique conceals malicious payloads or exfiltrated data within benign carrier files. A common method alters the Least Significant Bits (LSB) of image pixels, encoding data without visible changes. Detection requires analyzing files for statistical anomalies and unexpected bloat.
- **Timeline Manipulation (Timestomping):** Attackers routinely alter MAC/MACE (Metadata, Access, Creation, and Modification) timestamps to camouflage freshly deployed malware as legacy system files. Forensic engines identify this temporal tampering by cross-referencing surface-level timestamps with immutable registry artifacts.
- **Advanced Encryption:** Forensic laboratories combat complex cryptographic locks by investing heavily in high-performance computing clusters designed for brute-force decryption and dictionary attacks.

Jurisprudential Frameworks and Electronic Evidence Admissibility

The technical extraction of data is merely the preliminary phase of an investigation; the legal treatment of that digital evidence dictates the ultimate success of cybercrime prosecutions. Evidence must be collected, processed, and presented in strict adherence to statutory law. The jurisprudential landscape governing electronic

evidence is currently undergoing a massive structural transformation, particularly within the Republic of India.

The Indian Context: From the IT Act to the BSA 2023

Historically, the admissibility of electronic evidence in India was governed by Section 65B of the Indian Evidence Act (IEA), 1872, read in conjunction with the Information Technology (IT) Act, 2000. Section 65B functioned as a "complete code" for digital evidence, mandating that any secondary electronic record (such as a printed email, a copied hard drive, or a WhatsApp chat log) be accompanied by a formal certificate of authenticity. This certificate had to be signed by a person occupying a responsible official position relative to the operation of the device. Judicial interpretations of Section 65B were heavily litigated and often contradictory, culminating in the landmark Supreme Court decision Arjun Panditrao Khotkar v. Kailash Kishanrao Goratyal (2020).[14] This ruling clarified that while the certificate under 65B(4) is mandatory for secondary evidence, it is unnecessary if the "original" device itself (e.g., the actual smartphone or laptop) is physically produced in court.

On July 1, 2024, India transitioned to a completely modernized criminal justice framework designed for the digital age, replacing the colonial-era IEA with the Bharatiya Sakshya Adhiniyam (BSA), 2023, and the Indian Penal Code (IPC) with the Bharatiya Nyaya Sanhita (BNS), 2023.[16] This legislative overhaul introduced profound shifts in digital forensics jurisprudence:

Legal Parameter	**Previous Regime (IEA 1872 / IPC)**	**Modern Regime (BSA 2023 / BNS 2023)**
Primary Statutory Provision	Section 65B of the Indian Evidence Act.	Section 63 of the Bharatiya Sakshya Adhiniyam.[15]
Evidentiary Status	Digital records were often treated as secondary documentary evidence requiring external corroboration.	Section 57 classifies digital records inherently as **primary evidence** under broad, specific conditions. Section 61 explicitly prevents the denial of admissibility solely on the basis that the evidence is an electronic record.

Certification Structure	Single certificate under Section 65B (4).	Introduction of a dual-certificate system under Section 63(4): One certificate provided by the party operating the device, and a second technical authentication certificate by an expert. This mandates the declaration of cryptographic Hash Values (MD5/SHA256) to ensure integrity.
Simultaneous Data Creation	Ambiguous treatment of networked, distributed, and simultaneous data creation.	Explanations 4–7 in Section 57 establish that sequentially stored files, temporary automated storage spaces, and video recordings stored while being simultaneously transmitted are legally recognized as primary evidence.
Criminal Substantive Law	Cybercrimes were addressed indirectly via traditional IPC laws regarding cheating, forgery, and obscenity.	The BNS 2023 explicitly codifies electronic and digital means of committing offenses, modernizing terminology to address modern cyber threats directly.

While the IT Act of 2000 remains, the primary special legislation detailing technical cyber offenses and establishing the Examiner of Electronic Evidence under Section 79A, the BNS addresses the requisite criminal intent, punishment, and broader general criminal liability. Together, the BSA and BNS work in tandem to comprehensively prosecute cyber threats, effectively upgrading digital data from a derivative output to a foundational, primary legal asset that simplifies court introductions while enhancing technical scrutiny.

International Jurisdictional Conflicts and Data Sovereignty

The transnational nature of cloud computing—where data is frequently fragmented and stored across distributed server architectures spanning multiple sovereign borders—creates acute jurisdictional friction for forensic investigations. Law enforcement agencies frequently encounter scenarios where the data essential to prosecuting a localized cybercrime resides in foreign data centers, creating a clash between the borderless nature of the internet and the strictly territorial nature of criminal law.

Historically, cross-border data requests required navigating Mutual Legal Assistance Treaties (MLATs). However, the MLAT process is heavily criticized as bureaucratically inefficient and misaligned with the digital age, with response times often averaging between six to 24 months—and sometimes stretching beyond three years. This delay renders MLATs practically obsolete for securing volatile digital

evidence that can be deleted in seconds. To circumvent these delays, the United States enacted the Clarifying Lawful Overseas Use of Data (CLOUD) Act. This legislation modernizes the evidence framework by allowing U.S. law enforcement to compel U.S.-based cloud service providers to disclose data under their control, regardless of where the server is physically located globally. The CLOUD Act also permits trusted foreign partners to serve direct orders to U.S. providers via executive agreements. However, this extraterritorial reach frequently conflicts with international privacy frameworks, most notably the European Union's General Data Protection Regulation (GDPR).[18] The GDPR strictly limits the outward transfer of EU citizen data, placing multinational cloud providers in legally precarious scenarios where complying with a U.S. CLOUD Act subpoena may constitute a direct violation of European privacy law.

To harmonize these disparate national laws, the Council of Europe drafted the Budapest Convention on Cybercrime, the most comprehensive international treaty seeking to improve international investigative cooperation and establish uniform substantive criminal laws regarding data interference and computer-related fraud.[19] The Convention also mandates the establishment of a 24/7 point-of-contact network to facilitate rapid evidence preservation. Despite aligning fundamentally with the treaty's goals regarding cybersecurity, the Republic of India has persistently refused to sign or accede to the Budapest Convention. Strategic analysts note that India's abstention is primarily driven by profound concerns over national data sovereignty. Specifically, Article 32b of the Convention permits transborder access to stored computer data without explicit host-nation consent if the data is publicly available or if the investigating state obtains the lawful, voluntary consent of the person who has the lawful authority to disclose the data through that computer system. India views this provision as a potential infringement on its sovereign jurisdiction over data residing within its territory. Furthermore, critics argue the treaty suffers from significant geographical and political imbalances; it was drafted primarily by Western European nations without the early participation of developing nations like India, leading to arguments that it fails to adequately reflect the Global South's perspectives on human rights, state-actor threats, and digital security.[17]

Ethical Considerations and Professional Standards

Digital forensics inherently demands balancing thorough investigation with fundamental individual privacy rights. The capacity to extract vast amounts of personal data places immense ethical responsibility on investigators. Professional certifying bodies enforce strict codes of conduct, mandating objectivity, confidentiality, and legally authorized evidence collection. Investigators must also navigate complex legislative privacy guardrails, such as the GDPR or HIPAA. Unauthorized data extraction risks severe legal sanctions, compromises the

reliability of inspection results, and ultimately erodes public trust in the judicial process.[20][21]

Emerging Technological Frontiers (2025 and Beyond)

The discipline of digital forensics is currently undergoing a profound technological paradigm shift, catalyzed by several emerging technologies that will redefine the future capabilities of both cybercriminals and cyber investigators.

- **Artificial Intelligence and Machine Learning:** AI integration is drastically accelerating the speed and precision of data analysis. Machine learning algorithms process massive, disparate datasets with unprecedented velocity, automatically flagging anomalies, mapping behavioral patterns, and generating predictive models of attacker behavior before an incident fully materializes. AI is also heavily deployed to combat the surge in AI-driven attacks, utilizing sophisticated pixel-level analysis and acoustic frequency pattern recognition to identify frame inconsistencies in deepfake media, thereby preserving the integrity of digital media evidence. AI reduces manual workloads, lowers investigation costs, and provides critical predictive insights.
- **Quantum Computing:** While still in its developmental stages, quantum computing promises to revolutionize processing power, presenting a dual-edged sword for digital forensics. Its theoretical capacity to execute complex mathematical models threatens to easily break contemporary complex encryption standards, potentially unlocking previously inaccessible evidence vaults and encrypted drives. Conversely, it threatens the security of current digital infrastructure. Consequently, the forensic and security communities are accelerating research into quantum-resistant cryptographic algorithms to secure future data while simultaneously developing quantum forensic techniques for advanced data recovery.
- **Blockchain Forensics**: Cryptocurrencies and decentralized finance ecosystems inherently facilitate anonymity, making them the preferred medium for ransomware syndicates and money laundering operations. However, the immutable, distributed ledger technology underlying blockchain provides unprecedented transparency. Advanced blockchain forensic tools permit investigators to conduct complex transaction tracing across borders, mapping illicit capital flows, and deanonymizing wallets associated with threat actors. By 2025, the evolution of automated blockchain forensic solutions and large data analytics will be critical for efficient evidence processing in financial cybercrimes.
- **IoT and Cloud Forensics:** The proliferation of interconnected smart devices introduces highly fragmented evidence environments characterized by transient, ephemeral data. The development of specialized frameworks and standardized methodologies for extracting data from distributed cloud environments and

diverse IoT device ecosystems is paramount, as traditional forensic methods are often inapplicable to these decentralized architectures. Furthermore, organizations are increasingly utilizing "cyber deception" techniques, deploying decoy systems and honeypots to gain anticipatory forensic visibility into attacker intent, enhancing proactive defensive measures.

Conclusion

Digital forensics and cybercrime investigation have matured from the rudimentary, ad-hoc extraction of local hard drives into a highly complex, multi-disciplinary scientific endeavor mandated to secure the stability of the global digital infrastructure.[22] As the economic impact of cybercrime scales toward an anticipated $23 trillion, driven by AI-accelerated intrusions and sophisticated ransomware syndicates, the response mechanisms of law enforcement and corporate security must remain agile and technically superior.

The exhaustive analysis of current trends indicates that technological innovation alone is insufficient to combat modern threats; true efficacy relies on the synchronized advancement of technical methodologies, robust legal frameworks, and unyielding ethical standards. The ongoing transition toward AI-driven agentic triage models, the necessary escalation to invasive hardware-level JTAG and ISP extractions to defeat secure boot protocols, and the active development of steganography detection engines highlight the technical adaptability of the field. Concurrently, the codification of digital data as primary legal evidence—evidenced by monumental legislative leaps such as India's Bharatiya Sakshya Adhiniyam, 2023—demonstrates a vital, systemic legal adaptation to the evolving threat matrix. Ultimately, addressing the future of cybercrime requires not only mastering quantum decryption and blockchain transaction tracing but also forging international consensus to resolve the profound jurisdictional conflicts—such as those surrounding the CLOUD Act and the Budapest Convention—that currently hinder the rapid, borderless preservation of critical electronic evidence.

References

1. Cybersecurity Ventures, "Cybercrime To Cost The World $10.5 Trillion Annually By 2025." https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/ (accessed Apr. 7, 2026).
2. World Economic Forum, "Global Cybersecurity Outlook 2025." https://reports.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2025.pdf (accessed Apr. 7, 2026).
3. World Economic Forum, "Cybercrime Atlas: Impact Report 2025." https://reports.weforum.org/docs/WEF_Cybercrime_Atlas_Impact_Report_2025.pdf (accessed Apr. 7, 2026).
4. The Economic Times, "Govt says cyber fraud cases jumped over four-fold in FY24, caused $20 million losses."

https://m.economictimes.com/tech/technology/govt-says-cyber-fraud-cases-jumped-over-four-fold-in-fy24-caused-20-million-losses/articleshow/118879782.cms (accessed Apr. 8, 2026).

5. India Today, "Cybercrime may cost India Rs 20000 crore in 2025; banking, e-com most vulnerable." https://www.indiatoday.in/india/story/cybercrime-may-cost-india-rs-20000-core-in-2025-banking-e-com-most-vulnerable-2688350-2025-03-03 (accessed Apr. 8, 2026).
6. Deloitte, "Indian digital forensic market report." https://www.deloitte.com/in/en/services/consulting/services/risk-regulatory-forensic/indian-digital-forensic-market-report.html (accessed Apr. 8, 2026).
7. The420.in, "UP Police Froze 325 Crore Cyber Fraud Cases In 2025, Lifts Recovery Rate From 11% To 24%." https://the420.in/uttar-pradesh-police-cyber-fraud-money-freeze-2025/ (accessed Apr. 8, 2026).
8. Hindustan Times, "UP rises to 3rd spot nationally in freezing cyber fraud funds; DGP pushes 'golden hour' reporting." https://www.hindustantimes.com/cities/lucknow-news/up-rises-to-3rd-spot-nationally-in-freezing-cyber-fraud-funds-dgp-pushes-golden-hour-reporting-101770538671605.html (accessed Apr. 8, 2026).
9. NIST, "Digital Investigation Techniques: A NIST Scientific Foundation Review." https://nvlpubs.nist.gov/nistpubs/ir/2022/NIST.IR.8354.pdf (accessed Apr. 8, 2026).
10. MDPI, "An Empirical Assessment of Digital Forensic Process Reliability Using Integrated ISO/IEC 27037 and 27041 Standards." https://www.mdpi.com/2624-800X/6/2/57 (accessed Apr. 8, 2026).
11. DataCare Labs, "Forensic Chip-Off and ISP Recovery: Bypassing Secure Boot to Retrieve Data from Physically Damaged Mobile Devices." https://www.datacarelabs.com/blog/forensic-chip-off-isp-mobile-device-data/ (accessed Apr. 9, 2026).
12. CISA NICCS, "Advanced ISP-IoT-JTAG Cell Phone Data Recovery." https://niccs.cisa.gov/training/catalog/h11dfo/advanced-isp-iot-jtag-cell-phone-data-recovery (accessed Apr. 9, 2026).
13. University of Greater Manchester - Research Portal, "Detecting and Mitigating Anti-Forensic Techniques: A Comprehensive Framework for Digital Investigators." https://ub-ir.bolton.ac.uk/view/pdfCoverPage?instCode=44UOBO_INST&filePid=136728840008841&download=true (accessed Apr. 9, 2026).
14. Cyril Amarchand Mangaldas, "Supreme Court on the admissibility of electronic evidence under section 65B of the Evidence Act." https://corporate.cyrilamarchandblogs.com/2021/01/supreme-court-on-the-admissibility-of-electronic-evidence-under-section-65b-of-the-evidence-act/ (accessed Apr. 9, 2026).

15. R K Dewan & Co., "Section 63: Bhartiya Sakshya Adhiniyam & Digital Evidence." https://www.rkdewan.com/articles/electronic-records-now-governed-by-section-63-of-the-bhartiya-sakshya-adhiniyam-2023/ (accessed Apr. 9, 2026).
16. Your Law Article, "Offences Against Cybercrime under the Bharatiya Nyaya Sanhita, 2023 (BNS)." https://www.yourlawarticle.com/post/offences-against-cybercrime-under-the-bharatiya-nyaya-sanhita-2023-bns (accessed Apr. 9, 2026).
17. Council of Europe, "Cybercrime, Evidence and Territoriality: Issues and Options." https://rm.coe.int/cybercrime-evidence-and-territoriality-issues-and-options/168077fa98 (accessed Apr. 10, 2026).
18. Exoscale, "CLOUD Act vs. GDPR: The Conflict About Data Access Explained." https://www.exoscale.com/blog/cloudact-vs-gdpr/ (accessed Apr. 10, 2026).
19. Internet Democracy Project, "India and the Budapest Convention - Policy Brief 03." https://cdn.internetdemocracy.in/idp/assets/downloads/reports/india-and-the-budapest-convention-to-sign-or-not-considerations-for-indian-stakeholders/Dr.-Anja-Kovacs-Inda-and-the-Budapest-Convention.pdf (accessed Apr. 10, 2026).
20. IACIS, "International Association of Computer Investigative Specialists Code of Ethics and Professional Conduct Policy." https://www.iacis.com/wp-content/uploads/2023/03/IACIS-Code-of-Ethics-and-Professional-Conduct-v1.6.pdf (accessed Apr. 10, 2026).
21. E3S Web of Conferences, "Framework for Digital Forensic Ethical Violations: A Systematic Literature Review." https://www.e3s-conferences.org/articles/e3sconf/pdf/2023/85/e3sconf_icenis2023_01003.pdf (accessed Apr. 10, 2026).
22. ResearchGate, "(PDF) Emerging Trends in Digital Forensics for Cybercrime Investigation." https://www.researchgate.net/publication/403497382_Emerging_Trends_in_Digital_Forensics_for_Cybercri me_Investigation (accessed Apr. 10, 2026)

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 118 - 125

Emerging Threats and Innovative Solutions for Cyber Security

Mr. Ramesh Kumar

Assistant Professor (Guest), Department of BCA M.L.K.P. G. College, Balrampur

Email: ramesh.kumar@mlkcollege.ac.in

Article DOI Link: https://zenodo.org/uploads/22030618

DOI: 10.5281/zenodo.22030618

Abstract

In the contemporary digital age, technology advances at an astonishing rate, so too do the threats to our digital infrastructure. This paper delves into the emerging cyber security landscape, exploring the latest threats and examining innovative solutions that are being developed to safeguard our digital world.

Malicious actors are constantly developing new techniques to exploit vulnerabilities and breach security defenses. Phishing attacks, ransom ware, and malware are just a few examples of the threats that organizations and individuals face. These attacks can have devastating consequences, ranging from financial loss and data breaches to disruptions in critical infrastructure.

Another significant threat is the proliferation of Internet of Things (IoT) devices. While IoT devices have revolutionized many aspects of our lives, they also introduce new vulnerabilities. Furthermore, the rise of artificial intelligence (AI) and machine learning (ML) has both positive and negative implications for cyber security.

Advanced threat detection and response (ATDR) systems, powered by AI and ML, can help organizations to identify and mitigate threats in real-time. These systems can analyze vast amounts of data to detect anomalies and suspicious activity, enabling organizations to respond quickly and effectively to incidents.

Cyber security education and awareness are also crucial for protecting against threats. By educating employees about cyber security best practices, organizations can reduce the risk of human error, which is often a contributing factor in successful attacks.

The cyber security landscape is constantly evolving, and organizations must be prepared to adapt to new threats. By investing in advanced technologies, promoting cyber security education, and fostering a culture of security, organizations can mitigate the risks and ensure a more secure digital future.

Keywords: Cyber security, AI and ML, threats, ATDR technologies, digital.

Emerging Cyber Threats

AI and ML-Driven Attacks: A Growing Threat

The convergence of artificial intelligence (AI) and machine learning (ML) has revolutionized countless industries, but it has also opened new avenues for malicious actors. These technologies are being leveraged to automate and enhance various attack techniques, making cyber threats more sophisticated and difficult to detect.

One of the most prominent examples of AI and ML-driven attacks is the creation of more sophisticated phishing emails. AI algorithms can analyze vast datasets of successful phishing campaigns to identify patterns and vulnerabilities in human behavior.

Another alarming trend is the use of deepfakes for social engineering purposes. Deepfakes are synthetic media, such as videos or audio recordings, that are created using AI to manipulate existing content.

The increasing sophistication of AI and ML-driven attacks poses a significant challenge for cyber security professionals. AI and ML-driven attacks represent a growing threat to cyber security.

IoT Vulnerabilities: A Growing Threat

The Internet of Things (IoT) has revolutionized the way we interact with our environment, but it has also introduced new vulnerabilities that can be exploited by malicious actors. IoT devices, often with limited processing power and memory, are frequently designed with weak security measures. These vulnerabilities can be exploited to launch a variety of cyber-attacks, including:

- **Evasion Attacks (Adversarial Examples):** Attackers slightly alter input data (e.g., adding noise to an image or changing text) so the AI misclassifies it, such as causing a malware detector to miss a virus.
- **Data Poisoning:** Malicious data is injected into training datasets to create backdoors or reduce model accuracy, effectively compromising the model's knowledge base
- **Distributed Denial-of-Service (DDoS) Attacks:** IoT devices can be compromised and turned into bots to participate in DDoS attacks, overwhelming targeted systems with traffic and rendering them inaccessible.
- **Data Breaches:** IoT devices that collect sensitive data, such as personal information or financial data, can be targeted by attackers seeking to steal or misuse this information.
- **Supply Chain Attacks:** Malicious actors can compromise the supply chain of IoT devices to introduce malware or backdoors into the devices, allowing them to gain unauthorized access to systems and networks.

- **Physical Damage:** In some cases, IoT devices can be manipulated to cause physical harm, such as controlling industrial machinery or critical infrastructure.

One of the primary reasons for the vulnerability of IoT devices is their lack of proper security updates.

Quantum Computing Threats: A New Era of Cyber security Challenges

The development of quantum computers marks a significant technological leap, promising to revolutionize fields such as materials science, drug discovery, and artificial intelligence.

Quantum computers leverage the principles of quantum mechanics to perform calculations at speeds far surpassing classical computers. This computational power can be harnessed to solve complex problems that are intractable for classical machines. One such problem is the factorization of large numbers, a task that underpins many modern encryption algorithms.

Public-key cryptography, which is widely used for secure communication and data transmission, is based on this principle. By factoring a large number, an attacker can potentially decrypt encrypted messages.

The implications of quantum computers breaking current encryption standards are far-reaching. Sensitive data, such as financial transactions, medical records, and government secrets, could be compromised. This could lead to identity theft, financial loss, and national security breaches.

Supply Chain Attacks: A Hidden Threat

In today's interconnected global economy, organizations rely heavily on third-party suppliers and vendors to provide essential goods and services. Supply chain attacks, which target third-party suppliers and vendors to gain access to larger organizations, have become a growing concern in recent years.

One common tactic used in supply chain attacks is the compromise of software or hardware components. Attackers can infiltrate the supply chain of a third-party vendor and introduce malicious code or backdoors into their products. To mitigate the risk of supply chain attacks, organizations must adopt a comprehensive security strategy that includes the following:

- **Vendor Due Diligence:** Organizations should carefully evaluate the security practices of their third-party suppliers and vendors. This includes assessing their security policies, procedures, and track record.
- **Security Contracts:** Organizations should require their third-party suppliers and vendors to sign security contracts that outline their obligations to protect sensitive data and prevent breaches.
- **Regular Audits:** Organizations should conduct regular security audits of their third-party suppliers and vendors to identify and address any vulnerability.

- **Incident Response Planning:** Organizations should have a plan in place to respond to supply chain attacks, including steps to contain the breach, notify affected parties, and recover from the incident.
- **Employee Education and Awareness:** Employees should be trained to recognize and report suspicious activity related to the supply chain.

Biometric Authentication Risks: A Double-Edged Sword

Biometric authentication, which involves using unique physical or behavioral characteristics to verify identity, has become increasingly prevalent in recent years. From fingerprint readers on smart phones to facial recognition systems in airports, biometric technology offers convenience and enhanced security.

One of the primary concerns with biometric authentication is the potential for spoofing. Deepfakes, highly realistic synthetic media created using artificial intelligence, can be used to manipulate biometric data, making it possible for unauthorized individuals to bypass security measures. For example, a deepfake of a person's face could be used to unlock a device or gain access to a secure area.

Biometric data can be compromised through data breaches. If biometric templates are stored in a database that is not adequately protected, they could be stolen and used for malicious purposes. This could lead to identity theft, financial fraud, and other serious consequences. Biometric authentication offers a convenient and secure way to verify identity.

Innovative Solutions

1. **Advanced Threat Detection and Response (ATDR): A Crucial Defense against Cyber attacks**

To effectively defend against these attacks, organizations must adopt advanced threat detection and response (ATDR) strategies. Leveraging the power of artificial intelligence (AI) and machine learning (ML), ATDR systems can analyze vast amounts of data, identify anomalies that may indicate a cyber-attack, and automate incident response processes, significantly reducing the time it takes to detect and contain threats.

One of the key advantages of ATDR is its ability to analyze large volumes of data in real-time. By examining network traffic, user behavior, and system logs, ATDR systems can identify patterns and anomalies that may be indicative of a cyber-attack.

ATDR systems can automatically isolate infected systems, block malicious traffic, and initiate remediation procedures. This can help to prevent the spread of malware and limit the damage caused by an attack. To ensure the accuracy and effectiveness of ATDR systems, organizations must invest in high-quality data and continuously update their models to reflect the latest threat intelligence. Furthermore, ATDR systems should be integrated with other security measures, such as firewalls, intrusion detection systems, and endpoint protection solutions.

2. Blockchain Technology: A Revolution in Data Management

Blockchain technology is a decentralized, distributed ledger that records transactions in a way that makes them tamper-proof and transparent. This innovative approach has far-reaching implications for data storage, verification, and security. One of the most significant advantages of blockchain is its decentralized nature.

Blockchain's tamper-proof properties are another key benefit. Each transaction on a blockchain is recorded as a block and linked to previous blocks, creating a chain of blocks that is virtually impossible to alter. If an attacker attempts to modify a transaction, it would be immediately detected by other nodes on the network, invalidating the change.

The decentralized and tamper-proof nature of blockchain has led to its widespread adoption in various industries. In the supply chain, blockchain can be used to track the movement of goods from the manufacturer to the consumer.

In the financial sector, blockchain technology is being used to develop crypto currencies, such as Bitcoin and Ethereum. These digital currencies offer a decentralized alternative to traditional fiat currencies, eliminating the need for intermediaries like banks. Blockchain can also be used to streamline cross-border payments, reducing transaction costs and improving efficiency.

3. Quantum-Resistant Cryptography: A Shield against the Quantum Threat

The advent of quantum computing has ushered in a new era of technological advancements, promising to revolutionize fields such as drug discovery, materials science, and artificial intelligence. Quantum computers have the potential to break traditional encryption algorithms, rendering much of our current cyber security infrastructure vulnerable.

Quantum-resistant cryptography, also known as post-quantum cryptography, refers to encryption algorithms that are designed to be secure against attacks from both classical and quantum computers.

One promising approach to quantum-resistant cryptography is lattice-based cryptography. This method involves solving problems related to lattices, which are mathematical structures composed of points in space.

The development of quantum-resistant cryptography is a complex and ongoing process. As quantum computing technology continues to advance, the need for quantum-resistant cryptography becomes increasingly urgent. Governments, businesses, and individuals must be proactive in adopting these new encryption methods to protect their sensitive data.

Zero Trust Architecture: A Paradigm Shift in Cyber security

Perimeter-based security models, which rely on protecting a network's boundaries, are no longer sufficient to safeguard sensitive data against sophisticated cyber

threats. To address this challenge, a new security paradigm has emerged: Zero Trust Architecture (ZTA).

Zero Trust Architecture is a security model that assumes that no one inside or outside the network can be trusted. ZTA requires strict verification and authorization for every access request, regardless of the user's location or device.

The core principles of ZTA are:

- **Least Privilege:** Users are granted only the minimum permissions necessary to perform their job functions.
- **Continuous Verification:** Users are continuously verified and authenticated throughout their session, ensuring that they remain authorized to access resources.
- **Micro-segmentation:** Networks are divided into smaller segments, or micro-segments, to limit the lateral movement of attackers within the environment.
- **Data-Centric Security:** Security is focused on protecting data, rather than just the network perimeter. This involves implementing strong data encryption and access controls.

By adopting a Zero Trust approach, ZTA helps to prevent unauthorized access, limit the spread of malware, and protect sensitive data. It is particularly effective in today's hybrid work environments, where users may access corporate resources from a variety of devices and locations.

Implementing a Zero Trust Architecture requires a significant shift in mindset and a comprehensive security strategy. While Zero Trust Architecture offers a robust security solution, it is not a silver bullet. Organizations must continuously monitor their security posture and adapt their strategies to address emerging threats.

Cyber security Education and Awareness: A Foundation for Digital Resilience

Cyber security education plays a vital role in preventing human errors, which are often a major factor in successful cyber-attacks. Many attacks exploit human vulnerabilities, such as phishing scams and social engineering techniques.

Cyber security education can also help organizations to build a more resilient workforce. Employees who are aware of cyber security risks and best practices are better equipped to identify and report suspicious activity. They can also take steps to protect their own devices and data, reducing the risk of accidental breaches.

Cyber security education can help to foster a culture of security within organizations. When employees understand the importance of cyber security, they are more likely to take ownership of their role in protecting the organization's assets.

By educating individuals and organizations about the risks and best practices, we can help to prevent human errors, build a more resilient workforce, and foster a culture of security. As the digital landscape continues to evolve, the importance of cyber security education will only grow.

Conclusion

The future of cyber security is marked by a relentless arms race between attackers and defenders. As technology advances and the digital landscape expand, new threats continue to emerge, demanding innovative solutions to protect our digital infrastructure. This ongoing battle requires a proactive and adaptive approach, investing in advanced technologies, promoting cyber security education, and fostering a culture of resilience.

Artificial intelligence (AI) and machine learning (ML) are playing a crucial role in enhancing cyber security capabilities. These technologies can analyze vast amounts of data to identify patterns and anomalies that may indicate a cyber-attack. AI-powered systems can also automate incident response processes, reducing the time it takes to detect and contain threats.

As new threats emerge, organizations must invest in advanced technologies, promote cyber security education, and adopt a proactive approach to protect their digital infrastructure. By doing so, we can mitigate the risks and ensure a more secure digital future

References

1. Anand Bhushan Pandey, Ashish Tripathi & Prem Chand Vashist, A Survey of Cyber Security Trends, Emerging Technologies and Threats, 12 March 2022.
2. O. E. Akinbowale, H. E. Klingelhöfer and M. F. Zerihun, "Analysis of cyber-crime effects on the banking sector using the balanced score card: A survey of literature", J. Financial Crime, vol. 27, no. 3, pp. 945-958, Jun. 2020.
3. T. Maurer and A. Nelson, "The global cyber threat", Finance Develop., vol. 1, no. 3, pp. 24-27, 2021.
4. Z. Naz, Cybersecurity in banking sector: Importance threats challenges, May 2023.
5. D. Ghelani, T. K. Hua and S. K. R. Koduru, "Cyber security threats vulnerabilities and security solutions models in banking", Authorea, Sep. 2022
6. U. Islam, A. Muhammad, R. Mansoor, M. S. Hossain, I. Ahmad, E. T. Eldin, et al., "Detection of distributed denial of service (DDoS) attacks in IoT based monitoring system of banking sector using machine learning models", Sustainability, vol. 14, no. 14, pp. 8374, Jul. 2022.
7. Demetrio L, Biggio B, Lagorio G, Roli F, Armando A. Explaining vulnerabilities of deep learning to adversarial malware binaries. arXiv preprint arXiv; c2019.
8. IBM. Cost of a Data Breach Report 2023. Available online: https://www.ibm.com/security/digital-assets/cost-data-breach-report/ (accessed on 31 May 2024).

9. Shevchenko, P.V.; Jang, J.; Malavasi, M.; Peters, G.W.; Sofronov, G.; Trück, S. The nature of losses from cyber-related events: Risk categories and business sectors. J. Cybersecur. 2023, 9, tyac016.
10. Freeman, C.F.; Lewis, R. Bridging the gap between cyber risk management and cyber threat intelligence. Comput. Secur. 2017, 66, 1–9.
11. Ferreira, D.J.; Mateus-Coelho, N.; Mamede, H.S. Methodology for Predictive Cyber Security Risk Assessment (PCSRA). Procedia Comput. Sci. 2023, 219, 1555–1563.
12. Lee, I. Internet of Things (IoT) cyber security: Literature review and IoT cyber risk management. Future Internet 2020, 12, 157.
13. https://share.google/aimode/5ZZ1UgoISuNTNbevO: A brief description of types of AIML attacks.
14. https://share.google/oZKOh8WiGFgoHinGx: Latest cyber attacks

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 126 - 152

Explainable Machine Learning for GDP Growth Prediction: Interpreting Macroeconomic Indicators via SHAP Analysis on World Development Indicators

[1]Puneet Tiwari

[2]Avinash Singh

[1]Dept. of Computer Application, M.L.K P.G. College, Balrampur

[2]Dept. of Computer Application (Asst. Professor), M.L.K P.G. College, Balrampur

Email: puneettiwari.tech@gmail.com, avinash.singh@mlkcollege.ac.in

Article DOI Link: https://zenodo.org/uploads/22030712

DOI: 10.5281/zenodo.22030712

Abstract

Forecasting the growth of the domestic product or GDP is very important for people who make policies, big financial institutions and central banks all around the world. It is really hard to do this because the economy is complex and different from one country to another. Even though machine learning methods are better at predicting things than ways of doing economics people are not using them as much as they could because they do not understand how these models work. This is a problem because when it comes to making decisions people need to know why they are making them.

This study is trying to fix this problem by looking at four machine learning models. Linear Regression, Decision Tree, Random Forest and Gradient Boosting. To see which one is best at predicting the growth of the GDP. We used information from the World Banks World Development Indicators, which covers over 190 countries from 1960 to 2020. We looked at seven things that affect the economy: exports, imports, foreign investment, how much money is being spent on capital, government spending, inflation and population growth. We then looked at how each model did by using two measures: Mean Absolute Error and the coefficient of determination R^2.

The Random Forest model did the best with a Mean Absolute Error of 3.20 percentage points. To understand why this model did well we used something called SHapley Additive exPlanations or SHAP which helps us see how each feature affects the models predictions. This is important because it helps us understand how the model is making its predictions, which's not always easy to see. This study is

trying to make a framework for predicting GDP growth that is based on methods and a good understanding of the economy, which can help people make informed decisions about economic policy. The GDP growth is what we are trying to predict. We want to make sure that our model, for GDP growth is the best it can be.

Keywords: GDP growth prediction; World Development Indicators; SHAP; explainable machine learning; macroeconomic forecasting; feature importance; ensemble methods; economic interpretability; Random Forest; Gradient Boosting

Introduction

Gross domestic product (GDP) growth is among the most consequential economic indicators monitored by governments, supranational institutions, and financial markets. Reliable GDP growth forecasts inform fiscal policy, monetary decisions, sovereign debt ratings, and foreign investment allocation. The International Monetary Fund (IMF) and World Bank routinely invest substantial analytical resources in producing macroeconomic projections, yet empirical evidence consistently documents significant forecast errors even among institutional forecasters [1]. The consequences of systematic forecast failure extend beyond analytical inconvenience: misaligned fiscal stimulus, premature monetary tightening, or miscalibrated development aid can each produce durable socioeconomic harm.

Traditional econometric approaches-including Ordinary Least Squares (OLS) regression, Vector Autoregression (VAR), and Autoregressive Integrated Moving Average (ARIMA) models-dominated GDP forecasting throughout the latter half of the twentieth century. These methods offer transparency and theoretical grounding in established economic theory but impose strong parametric assumptions (linearity, stationarity, normality of residuals) frequently violated in cross-country panel data spanning multiple economic regimes [2]. Over the past decade, ML methods have emerged as competitive alternatives. Random Forest, Gradient Boosting, and deep learning architectures have demonstrated superior predictive accuracy in numerous macroeconomic forecasting benchmarks by capturing non-linear interactions, handling high-dimensional feature spaces, and tolerating heterogeneous data structures without restrictive distributional assumptions [3].

Nevertheless, predictive superiority alone is insufficient justification for deploying ML models in economic policymaking. A finance ministry cannot defensibly allocate public expenditure on the basis of a model whose internal logic is opaque. This interpretability deficit constitutes a fundamental barrier to institutional adoption. The problem is compounded in economics, where causal narratives are not merely desirable but legally and politically required for policy accountability. A model that achieves an MAE of 3.2 percentage points but cannot explain the relative contribution of inflation versus investment to a predicted output contraction

provides limited actionable intelligence to a central banker or development economist.

SHapley Additive exPlanations (SHAP), introduced by Lundberg and Lee [4], address this interpretability gap by decomposing each model prediction into individual feature contributions grounded in cooperative game theory. Unlike permutation importance measures-computed globally and susceptible to distortion by correlated features-SHAP values are both locally faithful (accurately reflecting each individual prediction) and globally consistent (summing to the difference between the prediction and the baseline expectation). This dual capability makes SHAP uniquely suited to economic analysis, where aggregate cross-country patterns and country-specific case studies are both analytically meaningful.

This paper presents a systematic investigation of GDP growth prediction across 190+ countries using WDI data, benchmarks four ML models on multiple evaluation metrics, and deploys SHAP to extract economically interpretable insights from the best-performing model. The primary contributions of this study are as follows:

- A rigorous multi-model benchmarking exercise demonstrating the relative predictive performance of linear, tree-based, and ensemble ML methods for global GDP growth prediction, with mathematical formulations explicitly stated for reproducibility.
- A quantitative feature importance analysis identifying population growth, gross capital formation, and inflation as the most influential predictors of GDP growth within the WDI dataset.
- A SHAP-based global and local explainability framework revealing non-linear, direction-variable feature contributions invisible to impurity-based importance measures-constituting the primary methodological novelty of this study.
- Evidence-based policy recommendations derived from SHAP decompositions, directly connecting ML model behavior to actionable macroeconomic governance guidance.

Problem Statement

This study is about trying to predict something. It is looking at a bunch of numbers that have to do with how a countrys economy's doing. These numbers are like a report card for the country. The study wants to figure out how these numbers can help us guess how much the countrys economy will grow in a year.

The study is using a lot of information from countries over many years. This makes it a bit tricky because each country is different. Some countries are rich. Some are poor. They all have ways of doing things.

There are a few problems with trying to make a model that can predict how much a countrys economy will grow. One problem is that the numbers do not always behave in a line. Sometimes they jump around a lot. Another problem is that some

of the numbers are closely related to each other. For example, when one country sells something to another country it is like they are connected.

The study also has to deal with missing information. Sometimes we do not have all the numbers we need. This can make it hard to make a model. Sometimes there are big surprises, like a war or a big economic crisis. These surprises can make it hard to make a model that works well.

The study is not just trying to make a model that can predict things accurately. It is also trying to make a model that can help us understand what is going on. The people who make decisions about money need to be able to explain why they are making those decisions.

The study is using a tool called SHAP to help with this problem. SHAP is a way of looking at the numbers and figuring out which ones are most important. It is a tool for this job because it is based on some strong math ideas. It can also help us understand what is going on at a level and at a small level.

The SHAP tool is helpful because it can give us an idea of what each number means. This can help the people who make decisions about money to make choices. The study is trying to use SHAP to make a model that's both accurate and easy to understand.

This is important because the people who make decisions about money need to be able to justify what they are doing. They need to be able to explain why they are making choices. The study is trying to help with this by making a model that's transparent and easy to understand.

The SHAP framework is a choice for this study because it is based on strong math ideas. It can help us understand what is going on with the numbers. It can also help us make decisions about money.

The study is looking at GDP growth which's a measure of how much a country economy is growing. It is a number because it can help us understand how well a country is doing. The study is trying to make a model that can predict GDP growth. It is using the SHAP tool to help with this.

The SHAP tool is a choice because it can help us understand what is going on with the numbers. It can also help us make decisions, about money. The study is trying to use SHAP to make a model that's both accurate and easy to understand.

This is a problem but the study is trying to make progress. It is using the SHAP tool to help with this. The study is looking at GDP growth. Trying to make a model that can predict it. It is a task but it is an important one.

Dataset Description

1. Data Source

The World Development Indicators (WDI) dataset, maintained by the World Bank Group, constitutes one of the most comprehensive compilations of international macroeconomic, social, and environmental statistics available in the public domain

[5]. The dataset provides annual panel data for over 200 economies from approximately 1960 to the present, encompassing more than 1,400 indicators spanning national accounts, trade, finance, infrastructure, and population dynamics. For this study, WDI data spanning 1960 to 2020 is utilized, providing a temporally extensive dataset that captures multiple economic cycles, structural transformations, and exogenous shocks—including the 1970s oil crises, the 1980s debt crisis, the 1997 Asian financial crisis, and the 2008–09 Global Financial Crisis. The WDI's standardized indicator definitions, cross-country comparability, and open-access licensing make it an established benchmark in empirical development economics and a methodologically appropriate foundation for cross-country ML modeling.

2. Feature Selection

Feature selection was guided by established economic theory-specifically the Keynesian aggregate demand framework and the Solow–Swan neoclassical growth model-combined with data availability across the full sample. Seven input features and one target variable were identified. Table I summarizes the features, their types, units, and economic rationale. The selection deliberately avoids indicators that are definitional components of GDP (such as household consumption expressed in current USD) to prevent target leakage, instead favoring structural ratios and growth rates that function as independent economic drivers.

Feature Name	Type	Unit	Economic Rationale
Exports of goods and services	Input	% of GDP	Export-led growth theory posits that external demand stimulates output, generates foreign exchange, and produces productivity spillovers through knowledge transfer and economies of scale.
Foreign direct investment (FDI), net inflows	Input	% of GDP	FDI transfers capital, technology, and managerial expertise, theoretically expanding productive capacity and total factor productivity (TFP) in host economies.
Government final consumption expenditure	Input	% of GDP	Government expenditure serves as a Keynesian fiscal multiplier below potential output but may crowd out private investment in neoclassical frameworks when above full employment.
Gross capital formation (GCF)	Input	% of GDP	Capital accumulation is a core output driver in the Solow–Swan model. Higher investment rates expand productive capital stock and sustain long-run growth trajectories.
Imports of goods and services	Input	% of GDP	Imports reflect domestic demand strength and technology absorption. High import shares may also indicate trade deficits

			constraining external balance.
Inflation, consumer prices	Input	Annual %	Moderate inflation may reflect demand-pull growth; high inflation erodes real incomes, amplifies uncertainty, and suppresses fixed investment via real interest rate effects.
Population growth	Input	Annual %	Population expansion increases labor supply and the domestic consumer base; the demographic dividend depends on age structure, human capital investment, and absorptive capacity.
GDP growth (annual %)	Target	Annual %	Dependent variable: annual percentage change in real GDP at constant 2015 USD prices, sourced directly from World Bank national accounts.

3. Data Preprocessing

Raw WDI data has several quality issues that require a careful preprocessing pipeline before model training. Missing values come from incomplete national statistical reporting, especially among low-income countries in earlier decades. We addressed these gaps with column-wise median imputation. Median imputation is better than mean imputation because the skewed distribution of macroeconomic indicators, particularly inflation, which often shows extreme right-skew due to hyperinflation, makes the arithmetic mean a poor measure of central tendency. We performed imputation separately on the training and test sets to avoid data leakage.

We treated outliers with caution. Unlike laboratory measurements, extreme GDP growth values in WDI data often reflect real macroeconomic events like post-conflict reconstruction booms, hyperinflation collapses, or commodity super-cycles. These events carry analytical significance. Removing these observations entirely would bias the model toward typical economic conditions and limit its usefulness in predicting crises. We applied Winsorization at the 1st and 99th percentiles only to inflation values. Without this step, hyperinflationary observations exceeding 500% annually would skew impurity criteria in tree-based splits without offering generalizable pattern information.

Feature scaling was not used for tree-based models, as they are unaffected by monotonic transformations of individual predictors. For Linear Regression, we standardized all features (zero mean, unit variance) using a StandardScaler fitted only on the training set. We split the full dataset into training (80%) and test (20%) subsets through stratified random sampling to maintain the distributional characteristics of GDP growth categories in both sets.

4. Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted to characterize the temporal dynamics of the target variable and the bivariate relationships among features prior to model development.

Fig. 1 presents the globally averaged annual GDP growth rate from 1960 to approximately 2018. The time series exhibits substantial cyclical volatility rather than a smooth secular trend, reflecting the interconnected nature of the global economy. Several historically significant episodes are clearly identifiable. The sharp decline to a trough of approximately 1.5% in the late 1970s corresponds to the twin oil price shocks of 1973 and 1979. A secondary contraction in the early 1980s reflects the Volcker disinflation episode and the associated Latin American debt crisis. The most precipitous decline-to near 0%-occurs in 2009, representing the Global Financial Crisis. This temporal analysis confirms that the data is non-stationary (mean and variance of GDP growth shift across decades) and identifies structural break points requiring regime-specific modeling in future work.

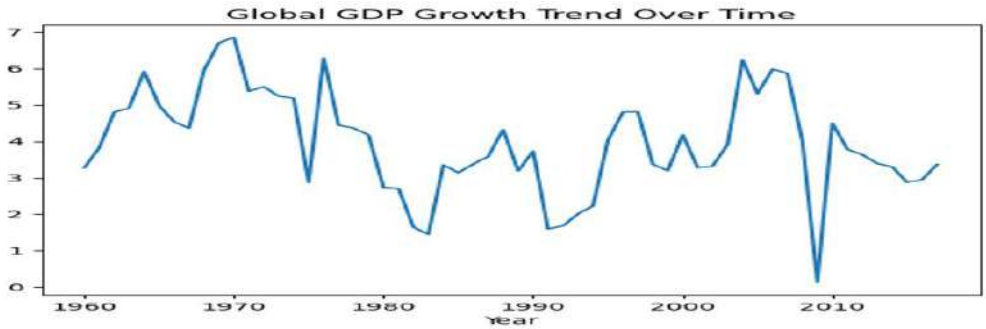


Fig. 1. Global GDP Growth Trend Over Time (1960–2018). Cross-country average annual GDP growth rate exhibits pronounced cyclical variation with structural breaks coinciding with major macroeconomic shocks, including the 1970s oil crises, the 1980s debt crisis, and the 2008–09 Global Financial Crisis.

Fig. 2 presents the Pearson correlation heatmap for all selected features and the target variable. The most salient finding is the strong positive correlation ($r \approx 0.80$) between exports and imports of goods and services - a structurally expected result, as export-oriented economies tend to import capital goods, intermediate inputs, and consumer products at elevated rates. This degree of multicollinearity has material implications for model specification: in Linear Regression, it inflates the variance of coefficient estimates and can produce misleading sign reversals. Tree-based models, selecting splitting variables based on local information gain rather than global correlation structure, are substantially more robust to multicollinear predictors. Critically, no feature exhibits a strong linear correlation with GDP growth; the highest bivariate association is approximately $r \approx 0.13$ for population growth—

confirming the necessity of non-linear ML methods and contextualizing the expectation of low R^2 scores throughout this study.

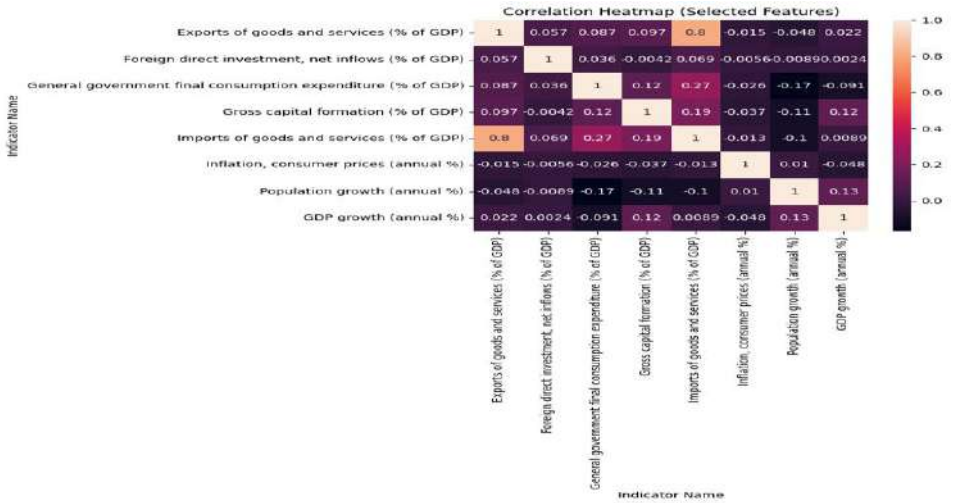


Fig. 2. Correlation Heatmap of Selected Features. Pairwise Pearson correlation coefficients among the seven input features and the GDP growth target variable. Notable: exports and imports exhibit r ≈ 0.80, indicating structural multicollinearity. No feature exhibits strong individual linear correlation with the target (max r ≈ 0.13).

Methodology

1. Research Framework

The analytical pipeline follows a sequential, decision-justified structure designed to minimize information leakage between training and evaluation while maximizing economic interpretability of outputs: (1) Data collection from WDI; (2) Preprocessing encompassing imputation, outlier treatment, scaling, and partitioning; (3) EDA comprising temporal trend analysis and correlation characterization; (4) Theoretically motivated feature selection; (5) Model training across four supervised ML algorithms; (6) Performance evaluation using MAE and R^2 on the held-out test partition; and (7) SHAP-based global and local explainability decomposition of the best-performing model.

2. Model Formulations

Linear regression is our basic model because it's straightforward to understand. We use a method called Ordinary Least Squares (OLS) to get its numbers. The way it makes predictions follows this formula:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_7 x_7 + \varepsilon.$$

In this formula, 'beta hat' is how OLS calculates everything, basically trying to make the squared errors as small as possible. The 'x's are our standardized features – the inputs we're using – and 'epsilon' is just the random bit of error. Since our early

data checks showed that these features don't have a very strong straight-line link with GDP growth, we don't think this model will be very accurate. It's probably going to give us some of the lowest prediction scores.

The Decision Tree Regressor breaks down the data space into two parts repeatedly. At each step, it chooses the feature and the cutoff point that best reduces the weighted Mean Squared Error in the two resulting groups:

$$\text{split}^* = \text{argmin}\,[\Sigma_{i\in\text{Left}}(y_i - \bar{y}\text{Left})^2 + \Sigma_{i\in\text{Right}}(y_i - \bar{y}\text{Right})^2].$$

We specifically included a tree that can grow without limits (max_depth = None). This is to clearly show how 'overfitting' happens and to highlight the big improvement we get by using a group of trees together.

A Random Forest Regressor deals with how unstable individual decision trees can be, using something called 'bootstrap aggregation,' or 'bagging.' We train 100 trees, each on its own random slice of the data. Plus, at every split, each tree only looks at a random portion (square root of 'p') of the features.

The group's final prediction is just the average of what all 100 trees came up with:

$$\hat{y}RF = (1/B)\,\Sigma_{\beta=1}^{B}\,T_\beta(x).$$

Here, T_beta is the prediction from any given tree. The variability in this average prediction is roughly $\sigma^2/B \cdot (1 + (B-1)\,\bar{\rho})$, where 'rho bar' shows how much the trees' predictions are connected. The main trick that makes Random Forest's predictions steadier than a single tree's is that it decorrelates them by having each tree look at only a random subset of features. This explains why it performed better with a lower Mean Absolute Error in this study.

3. Evaluation Metrics

Two primary evaluation metrics are employed. Mean Absolute Error (MAE) is defined as:

$$MAE = (1/n)\,\Sigma_{i=1}^{n}\,|y_i - \hat{y}_i|$$

where y_i and $\hat{y}_i$ denote actual and predicted GDP growth values, respectively. MAE is preferred as the primary metric because it is expressed in interpretable original units (percentage points of GDP growth), is linearly sensitive to all errors, and is not disproportionately influenced by extreme outliers—unlike Root Mean Squared Error, which squares residuals and heavily penalizes large deviations. An MAE of 3.20% indicates that model predictions deviate from the true GDP growth rate by an average of 3.20 percentage points—a directly policy-relevant quantity.

The coefficient of determination is defined as:

$$R^2 = 1 - \Sigma\,(y_i - \hat{y}_i)^2 / \Sigma\,(y_i - \bar{y})^2$$

measuring the proportion of variance in GDP growth explained by the model relative to a naive mean predictor. R^2 values near zero or negative are well-documented in cross-country GDP forecasting literature [6] and must be interpreted in the context of extraordinary distributional heterogeneity in global growth data, not dismissed as model failure.

4. SHAP Explainability Framework

SHAP values are rooted in Shapley values from cooperative game theory, where the 'players' are the model's input features and the 'payout' is the prediction. The Shapley value φ_i for feature i is:

$$\varphi_i = \Sigma_\{S \subseteq F \setminus \{i\}\} \; [|S|! \; (|F|-|S|-1)! \; / |F|!] \cdot [f_\{S \cup \{i\}\} \; (x) - f_s(x)]$$

where F is the full feature set, S iterates over all subsets excluding feature i, and $f_s(x)$ denotes the model's expected prediction given only the features in S. This formulation evaluates each feature's contribution across all possible feature orderings, satisfying the axioms of: efficiency (contributions sum to the prediction-baseline difference), symmetry (equally contributing features receive equal Shapley values), dummy (absent features receive zero attribution), and linearity (additive contributions across model components).

SHAP is preferred over permutation importance for three reasons. First, permutation importance is a global aggregate that masks directional and non-linear effects visible in SHAP's observation-level decomposition. Second, in the presence of correlated features, permutation importance arbitrarily allocates joint predictive power, producing unstable and misleading rankings. SHAP's marginalization over feature subsets more equitably distributes shared contributions. Third, SHAP's Tree Explainer algorithm computes exact Shapley values for tree-based models in polynomial time, making it computationally tractable on the scale considered here.

Model Development

All experiments were implemented in Python 3.9 using scikit-learn 1.2 for model training, pandas 1.5 and NumPy 1.23 for data manipulation, and SHAP 0.41 for explainability computations. Model training and evaluation were conducted on the 80/20 stratified train-test split described in Section III-C. No nested cross-validation was employed in this benchmarking phase, as the primary objective is comparative performance characterization rather than hyperparameter optimization. Table II summarizes the configuration of each model.

TABLE II. Model Configuration Summary

Model	Key Hyperparameters	Implementation Notes
Linear Regression	fit_intercept = True; no regularization	OLS estimation via scikit-learn LinearRegression; features standardized (zero mean, unit variance) prior to fitting using StandardScaler fitted on training partition only.

Decision Tree Regressor	criterion = 'squared_error'; max_depth = None	Unconstrained depth included deliberately to illustrate overfitting behavior. No feature scaling required (scale-invariant splits). Serves as interpretable intermediate baseline.
Random Forest Regressor	n_estimators = 100; max_features = 'sqrt'; random_state = 42	Bootstrap aggregation over 100 decorrelated trees. SHAP TreeExplainer applied post-training for global and local explainability decomposition.
Gradient Boosting Regressor	n_estimators = 100; learning_rate = 0.1; max_depth = 3; subsample = 1.0	Sequential residual fitting using squared-error loss. Default scikit-learn hyperparameters retained to ensure comparability with prior benchmarks.

Results and Analysis

Model Comparison - Mean Absolute Error

Fig. 3 presents the MAE scores for all four models evaluated on the held-out test partition. Random Forest achieves the lowest MAE of approximately 3.20 percentage points, followed by Gradient Boosting at approximately 3.35, Linear Regression at approximately 3.45, and Decision Tree at approximately 4.65. Table III summarizes the complete results.

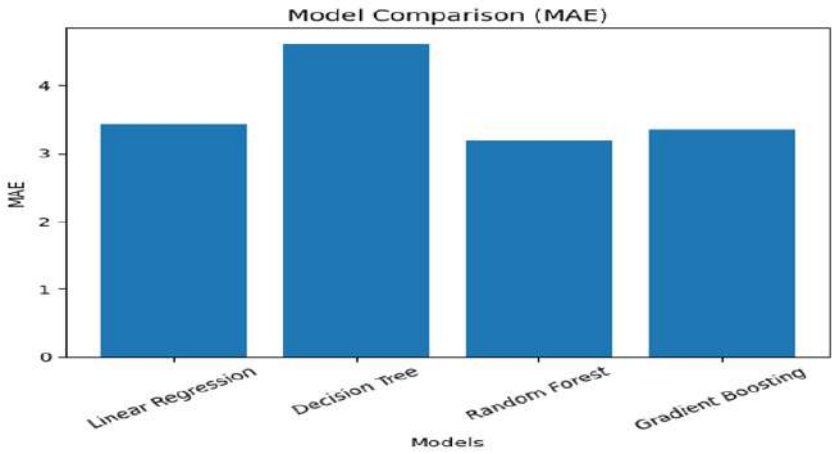


Fig. 3. Model Comparison by Mean Absolute Error (pp = percentage points). Lower values indicate superior predictive accuracy. Random Forest achieves the minimum MAE of approximately 3.20 pp, outperforming all competing models. The Decision Tree records the highest MAE (≈ 4.65 pp), consistent with severe overfitting in an unconstrained tree architecture.

The Decision Tree's substantially higher MAE (4.65 pp) reflects the well-understood variance amplification of fully grown trees: the model memorizes idiosyncratic training-set noise—country-year-specific patterns arising from political shocks, data anomalies, or exceptional commodity cycles-rather than generalizing the structural macroeconomic relationships present in the test partition.
The performance advantage of Random Forest over Linear Regression (approximately 0.25 pp reduction in MAE) is attributable to its capacity to model non-linear feature interactions, consistent with the EDA finding that linear correlations between features and GDP growth are uniformly weak (max $r \approx 0.13$).
The marginal advantage of Random Forest over Gradient Boosting (0.15 pp) suggests that variance reduction via bagging is marginally more effective than bias reduction via sequential boosting for this dataset, likely because prediction variance—driven by the extreme distributional heterogeneity across 190+ countries-is the dominant error source.
To contextualize these MAE values: a forecast error of 3.2 pp in GDP growth is policy-significant. A government projecting 5.0% growth but experiencing 1.8% will systematically overshoot its budget deficit targets, misallocate fiscal stabilizers, and potentially violate IMF program conditionality thresholds. While the model's MAE cannot be deemed negligible, it is competitive with documented institutional forecaster track records on comparable cross-country datasets [1].

TABLE III. Summary of Model Evaluation Results on Held-Out Test Partition

Model	**MAE (pp)**	**R^2 Score**	**Interpretation**
Linear Regression	≈ 3.45	≈ 0.00	Performance near mean-predictor baseline, consistent with weak linear feature–target correlations.
Decision Tree	≈ 4.65	≈ −1.10	Severe overfitting; $R^2 < 0$ confirms worse-than-mean predictions on test

			partition.
Random Forest (Best)	≈ 3.20	≈ +0.07	Best overall performance; bagging reduces prediction variance across heterogeneous country-year observations.
Gradient Boosting	≈ 3.35	≈ +0.05	Competitive with Random Forest; sequential bias reduction marginally less effective than variance reduction for this dataset.

Model Comparison — R^2 Score

Fig. 4 reveals that all four models achieve R^2 scores near zero or below on the test set. Random Forest and Gradient Boosting register marginally positive R^2 values (approximately +0.07 and +0.05, respectively), while Linear Regression is near zero and Decision Tree is severely negative at approximately −1.1.

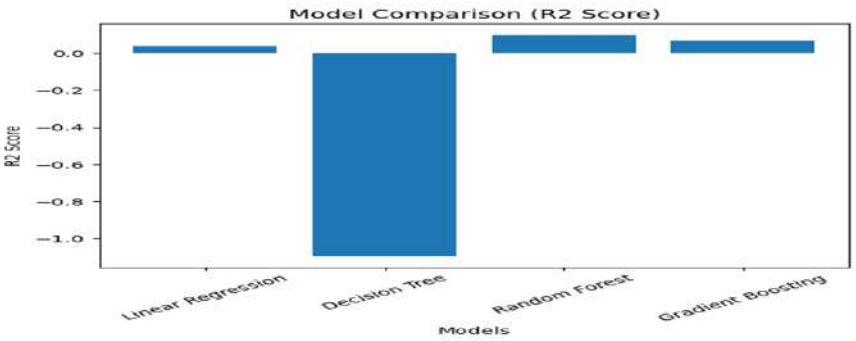


Fig. 4. Model Comparison by R^2 Score. Decision Tree records a strongly negative R^2 (≈ −1.1), confirming performance substantially worse than the trivial mean predictor on the test partition—quantitative evidence of severe overfitting. Random Forest and Gradient Boosting achieve modest positive R^2 values, confirming recovery of non-trivial predictive signal.

These results require careful economic interpretation rather than dismissal as model failure. Near-zero R^2 scores for ensemble models reflect the intrinsic unpredictability of annual GDP growth in a pooled cross-country panel: the global distribution of growth rates spans from extreme contractions (e.g., conflict-affected states with −40% growth) to rapid expansions (post-war reconstruction economies exceeding +30% growth). A single pooled model must simultaneously describe the growth dynamics of Switzerland in 2010 and Zimbabwe in 2008-a task that no parsimonious feature set can accomplish with high explained variance. This finding is consistent with the broader macroeconomic forecasting literature, where even the IMF's World Economic Outlook projections exhibit substantial R^2 variation and frequently fail to anticipate growth turning points [1]. The Decision Tree's R^2 of -1.1 is particularly diagnostic: it confirms that an unconstrained tree performs worse than the trivial mean predictor on out-of-sample data, providing compelling empirical justification for depth constraints or pruning as standard practice. The positive R^2 values of Random Forest and Gradient Boosting—however modest—confirm that ensemble aggregation recovers non-trivial predictive signal from the same features and dataset, signal subsequently decomposed via SHAP in Section VII.

Actual vs. Predicted GDP Growth

Fig. 5 plots actual against predicted GDP growth values for the best-performing Random Forest model across the test partition. The most analytically significant pattern is prediction shrinkage: model outputs concentrate within approximately −10% to +15%, while actual GDP growth values span from roughly −70% to +90%.

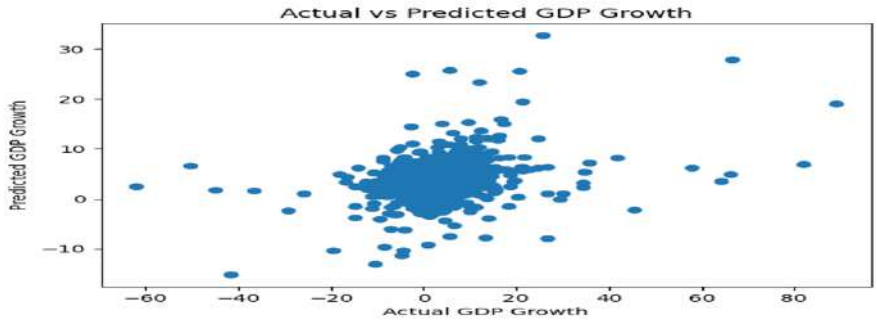


Fig. 5. Actual vs. Predicted GDP Growth Values — Random Forest Model. Predictions concentrate within −10% to +15% (blue cluster), while actual values span −70% to +90%, illustrating the prediction shrinkage characteristic of ensemble averaging. Extreme observations correspond to crisis episodes and post-conflict reconstruction years.

Prediction shrinkage is a well-understood property of ensemble averaging: the mean of B trees estimates the conditional expectation of the target, which regresses toward the global mean for extreme observations. In economic terms, the model performs well for the large majority of country-year observations within the ±10% range of normal economic conditions but systematically underestimates the severity of crises and the magnitude of rapid expansions. These extreme observations typically correspond to small, structurally fragile economies or exceptional historical episodes-post-Soviet transitions, hyperinflationary collapses, oil-exporting boom years-that are underrepresented in the feature distributions used for training. This limitation has direct implications for model deployment: policymakers in emerging market economies with volatile growth histories should not rely on this global model for crisis prediction without supplementary country-specific calibration.

Residual Analysis

Fig. 6 plots residuals (y - ŷ) against predicted values for the Random Forest model, providing diagnostic information on the error structure.

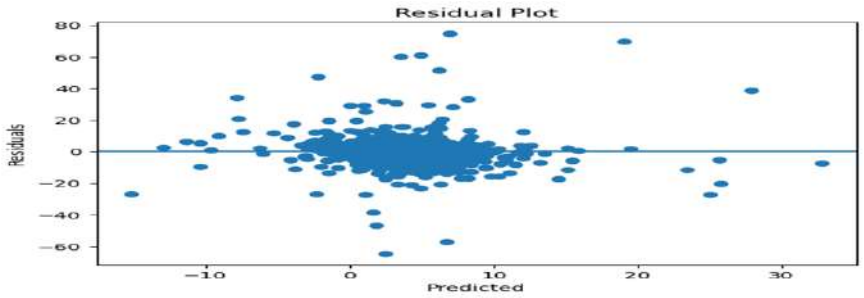


Fig. 6. Residual Plot — Random Forest Model. Residuals exhibit pronounced heteroscedasticity: error variance fans out substantially for predicted values exceeding approximately +5%. Outliers with |residual| > 50 pp are visible, likely

corresponding to crisis episodes, post-conflict reconstruction, or hyperinflationary economies operating outside the model's training distribution.

Two diagnostic anomalies are apparent. First, heteroscedasticity is evident: residual variance increases substantially as predicted values exceed approximately 5%. This non-constant error variance indicates that model reliability decreases precisely for high-growth economies-those of greatest interest to growth-focused governments. Statistically, heteroscedasticity does not bias point estimates but invalidates standard confidence interval calculations, suggesting that prediction intervals from this model would be systematically too narrow for high-growth predictions. Economically, the pattern suggests that high-growth environments exhibit greater structural heterogeneity: the same predicted growth rate of 8% can arise from qualitatively different underlying factor combinations across countries. Second, outliers exceeding ±50 pp are visible, likely representing extreme crisis episodes (1990s post-Soviet transitions, conflict-affected states) whose data-generating processes deviate qualitatively from the majority of the training distribution. Future model iterations may benefit from logarithmic target transformations, quantile regression for direct interval estimation, or country fixed effects to absorb persistent structural heterogeneity.

Feature Importance Analysis

Fig. 7 presents impurity-based (Gini) feature importance scores from the trained Random Forest model. Population growth emerges as the most influential predictor (score ≈ 0.205), followed by gross capital formation (≈ 0.148), inflation (≈ 0.140), and FDI (≈ 0.135). Government expenditure, imports, and exports register the lowest importance scores in the range 0.12–0.13.

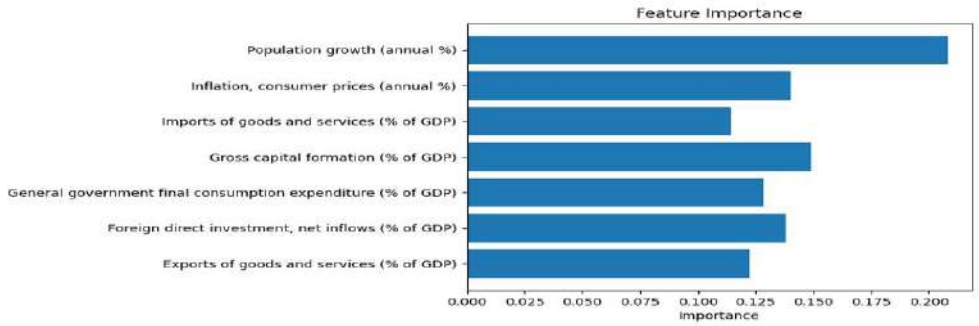


Fig. 7. Random Forest Feature Importance Scores (Gini Impurity). Population growth ranks as the most influential predictor (≈ 0.205), followed by gross capital formation (≈ 0.148) and inflation (≈ 0.140). Importance scores are non-negative and reflect global average information gain across all tree splits; directional effects are not captured at this level of analysis.

The dominance of population growth is consistent with demographic transition theory: economies with young, growing populations benefit from an expanding

labor force and rising domestic consumption, generating GDP growth through both supply-side (labor input) and demand-side (consumption) channels. The demographic dividend-the period during which the working-age share of the population peaks-has been identified as a significant contributor to East Asian growth miracles and is expected to materialize in Sub-Saharan Africa through the mid-twenty-first century [7]. Gross capital formation's second-place ranking is consistent with the Solow-Swan model's prediction that capital accumulation is a primary determinant of output per worker. The relative balance of scores-no single feature dominating by a wide margin-suggests that GDP growth is collectively determined by multiple macroeconomic channels, reinforcing the theoretical justification for a multivariate ML approach. A critical limitation of impurity-based importance is its susceptibility to bias toward high-cardinality continuous features and its inability to capture directional effects-limitations directly addressed by the SHAP analysis in Section VII.

Explainable AI — Shap Analysis

Global SHAP Analysis - Beeswarm Summary

Fig. 8 presents the global SHAP beeswarm summary plot, encoding both the magnitude and direction of each feature's contribution to individual predictions, color-coded by the original feature value (pink = high feature value, blue = low). This visualization delivers fundamentally richer information than the impurity-based importance scores of Fig. 7, which are strictly non-negative scalar summaries that aggregate across the entire dataset and suppress directional and non-linear effects.



Fig. 8. SHAP Beeswarm Summary Plot - Random Forest (Global Analysis). Each point represents one country-year observation. Horizontal position indicates SHAP value magnitude and direction (positive = increases predicted GDP growth; negative = decreases predicted GDP growth). Vertical axis ranks feature by mean \|SHAP value\|. Color encodes original feature value (pink = high, blue = low). Non-linear, direction-variable contributions are clearly visible for government expenditure, gross capital formation, and FDI.

Government final consumption expenditure shows a wide range of SHAP values without a clear link to spending levels. High government spending sometimes adds

positively to growth predictions and other times subtracts, reflecting a familiar tension in economic theory. When economies are running below their potential, government spending can boost GDP through fiscal multipliers, but in economies near full employment, it might crowd out private investment. These mixed effects appear in the SHAP values, capturing context in a way that simple linear models can't.

Population growth shows a more consistent pattern. Higher population growth tends to increase GDP predictions, supporting the idea of a demographic dividend. On the flip side, low population growth often lowers predicted GDP, matching what's been seen in aging advanced economies like Japan, Germany, and parts of Southern Europe, where slow population growth drags on economic performance. Gross capital formation adds an interesting twist: while high capital investment generally aids growth, sometimes very high levels correspond to negative SHAP values. This can happen in fast-growing countries facing limits in absorbing investment—excessive or misdirected capital spending may then reduce efficiency and slow growth, like certain phases of China's infrastructure push.

Foreign direct investment (FDI) shows varied SHAP impacts across different values. This spread reflects how FDI's effects depend on factors like a country's ability to absorb new investment, the quality of its institutions, and the sectors involved. Manufacturing-related FDI that transfers technology usually helps growth, but resource-focused FDI can have little or even negative broader economic impact.

Inflation's SHAP pattern aligns with traditional monetary views: high inflation mostly reduces predicted GDP, linked to loss of purchasing power, greater uncertainty for investors, and reduced real balances. Imports and exports have smaller, more clustered SHAP values, indicating limited direct effects on short-term GDP growth once other factors are considered, which fits with their low correlations found earlier [8].

Overall, the SHAP analysis reveals an important insight that plain feature importance misses: many features influence GDP growth differently depending on the economic context. This variation means that models assuming consistent, linear effects miss a lot. It highlights why non-linear machine learning methods combined with SHAP explanations are necessary to understand growth in a way that makes sense economically and informs policy.

Local SHAP Analysis - Single Prediction Decomposition

Fig. 9 presents a SHAP waterfall plot decomposing a single country-year prediction, providing the local interpretability layer that transforms the model from a statistical artifact into a policy-accountable decision tool. The base expectation E[f(X)] = 4.006% represents the model's average prediction across the training dataset-the prediction that would be made in the complete absence of feature-specific

information. The final prediction f(x) = 5.255% exceeds this baseline by 1.249 percentage points, with individual feature contributions explaining the departure as quantified in Table IV.

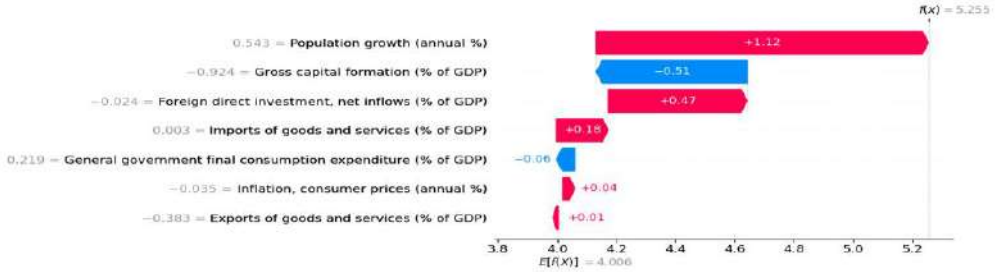


Fig. 9. SHAP Waterfall Plot - Single Country-Year Prediction (Local Analysis). Starting from the base expectation E[f(X)] = 4.006%, each feature's SHAP contribution accumulates to produce the final prediction f(x) = 5.255%. Red bars indicate positive contributions (increasing predicted GDP growth); blue bars indicate negative contributions. Population growth (+1.12 pp) and gross capital formation (−0.51 pp) are the dominant positive and negative drivers, respectively.

TABLE IV. Feature-Level SHAP Decomposition - Single Country-Year Prediction

Feature	**Feature Value**	**SHAP Value (pp)**	**Economic Interpretation**
Population growth	0.543 annual %	+1.12	Above-average population growth signals demographic dividend phase; expands labor supply and domestic consumption demand.
Gross capital formation	−0.924 (standardized)	−0.51	Below-average investment ratio is the binding negative constraint; indicates insufficient capital accumulation for sustained growth.
FDI, net inflows	−0.024 (standardized)	+0.47	Near-zero FDI in a high-population-growth context may signal untapped investment potential associated with positive growth expectations.
Imports	0.003 (standardized)	+0.18	Marginal positive contribution reflecting mild domestic demand signaling.

Government expenditure	0.219 (standardized)	−0.06	Slightly above-average government share introduces marginal crowding-out drag on private-sector activity.
Inflation	−0.035 (standardized)	+0.04	Near-average inflation exerts negligible directional effect on this prediction.
Exports	−0.383 (standardized)	+0.01	Below-average exports contribute negligibly; consistent with low individual feature–target correlation.
Cumulative contribution		+1.25	Base E[f(X)] = 4.006% → Final f(x) = 5.255%

Population growth (feature value: 0.543 annual %) contributes the largest positive increment of +1.12 pp. Economically, this represents a country with above-average population growth consistent with a developing economy in the demographic transition phase, where a growing working-age population expands both labor supply and domestic consumer demand. Gross capital formation (feature value: −0.924 standardized; below-average investment ratio) contributes the largest negative term: -0.51 pp. This finding is directly actionable-the single most impactful policy lever for increasing this country's predicted GDP growth is stimulating private and public investment, not trade liberalization or inflation management. FDI net inflows (feature value: -0.024) contribute a counterintuitive +0.47 pp, reflecting the non-linear interaction effects captured by the Random Forest model: near-zero FDI in a context of high population growth may signal untapped investment potential that the model associates positively with future growth trajectories.

The analytical significance of this waterfall plot extends beyond any single prediction. It demonstrates that SHAP enables a policymaker to reconstruct, feature by feature, the economic logic underlying a model's forecast-answering not merely 'what will GDP growth be?' but 'which factor constitutes the binding constraint, and what is its quantified contribution?' This level of model accountability represents a prerequisite for institutional adoption of ML-based forecasting tools in public economic governance.

Discussion

The results of this study yield a coherent analytical narrative across the three dimensions of model performance, feature importance, and SHAP explainability. Random Forest's superior MAE (3.20 pp) confirms that ensemble methods are

better suited than linear models or single decision trees for cross-country GDP growth prediction-consistent with the broader ML-in-economics literature [3] [9]. The key mechanism is variance reduction through bootstrap aggregation: averaging across 100 decorrelated trees attenuates sensitivity to noise and idiosyncratic country-year observations that would produce large errors in any single tree. Near-zero R^2 scores across all models are not a methodological failure but an honest reflection of the distributional challenge inherent in global GDP prediction-a result that aligns with documented IMF forecast accuracy assessments showing that macroeconomic turning points remain largely unpredictable even with sophisticated institutional models [1].

The tension between predictive accuracy and interpretability manifests clearly in this study. Linear Regression, the most interpretable model, achieves the second-worst MAE. Decision Tree, offering a partially interpretable rule structure, achieves the worst MAE due to overfitting. Random Forest and Gradient Boosting achieve superior accuracy at the cost of native interpretability-a cost that SHAP partially recovers. The SHAP global analysis reveals that population growth, FDI, and government expenditure each exert non-linear, direction-variable effects on GDP predictions that no linear coefficient can represent. This finding carries a direct policy implication: governments should not expect simple linear heuristics ('higher investment always drives growth') to hold uniformly across heterogeneous economic contexts. The SHAP waterfall plot demonstrates that context-specific factor decomposition is both technically achievable and economically interpretable at the individual country-year level.

Comparison with the macroeconomic forecasting literature reveals several consistent themes. Franses and Legerstee [6] document that standard macroeconomic models systematically fail to predict growth turning points-consistent with this study's observation that all models underestimate extreme GDP deviations. Medeiros et al. [9] find that ensemble ML models outperform individual specifications in GDP nowcasting, consistent with the Random Forest advantage documented here. Lundberg et al. [15] demonstrate that SHAP values derived from ensemble models accurately reflect true marginal feature contributions, providing theoretical grounding for the economic interpretations offered in Section VII. The finding that population growth dominates feature importance aligns with cross-country growth regression literature identifying demographic variables as robust correlates of long-run growth [7], though this study's annual panel context shifts the mechanism from long-run demographic transition to short-run labor force and demand dynamics.

The ethical dimension of ML-based economic forecasting warrants explicit attention. Deploying a model with near-zero R^2 as a direct input to fiscal policy decisions-without interpretability safeguards and uncertainty quantification—risks introducing algorithmic accountability gaps into public governance. SHAP partially

addresses this concern by providing auditable, feature-level explanations for each prediction. However, policymakers must recognize that SHAP values are model-specific: they explain the Random Forest's internal behavior, not the underlying macroeconomic data-generating process. A causal interpretation of SHAP values-treating them as estimates of the true causal impact of each variable on GDP growth-is methodologically inappropriate without further instrumental variable or experimental validation. The appropriate use of this framework is as an evidence-generating tool that identifies candidate policy priorities for deeper causal investigation, not as a direct policy decision algorithm.

Key Findings

The following summary statements consolidate the principal empirical findings of this study:

- Random Forest had the lowest mean absolute error (MAE) of about 3.20 percentage points, outperforming Gradient Boosting at 3.35 points, Linear Regression at 3.45 points, and Decision Tree at 4.65 points. The 1.45 percentage points difference between Random Forest and Decision Tree shows the cost of overfitting when using an unrestricted tree on varied macroeconomic data.
- All four models had R^2 scores around or below zero on the test set. This outcome reflects the inherent high variance in annual GDP growth over a 60+ year cross-country panel. This aligns with known forecast error rates rather than indicating model failure.
- Population growth is the most influential predictor, with an importance score of about 0.205. It is followed by gross capital formation at approximately 0.148 and inflation at around 0.140. These findings support demographic dividend theory and the Solow–Swan capital accumulation mechanism.
- The SHAP global beeswarm analysis shows that five out of seven features (government spending, population growth, gross capital formation, foreign direct investment, and inflation) have non-linear effects on GDP predictions that vary in direction. Linear regression cannot capture these effects. This confirms the need for non-linear machine learning methods for accurate economic modeling.
- The SHAP waterfall analysis of a specific country-year prediction shows that Random Forest predicts a GDP growth of 5.255% (base case: 4.006%). This prediction is mainly based on above-average population growth (+1.12 percentage points), partially offset by below-average gross capital formation (-0.51 percentage points). This breakdown can directly inform country-level economic planning.
- Prediction shrinkage is significant; the model limits forecasts to within -10% to +15%, while actual values range from -70% to +90%. This indicates that the

global pooled model works well in normal economic conditions but struggles to predict extreme growth episodes without adjustments for specific countries.

- Residual heteroscedasticity shows that model reliability decreases for high growth predictions. Outliers above ±50 percentage points are linked to crisis periods, post-conflict reconstruction, or economies facing hyperinflation, where the data-generating processes differ significantly from the training data majority.

Recommendations

Based on the empirical findings of this study, the following evidence-based recommendations are directed to three practitioner audiences.

For Economic Policymakers

- SHAP-based model explanations should accompany any ML-derived GDP forecast used in fiscal planning. The SHAP waterfall analysis demonstrates that below-average gross capital formation is the binding constraint in the exemplar country-year—governments should conduct country-specific SHAP decompositions annually to identify which macroeconomic lever offers the greatest marginal growth dividend, rather than applying uniform policy prescriptions.
- Population growth and gross capital formation should be prioritized as long-term growth policy focal points, consistent with their dominant feature importance scores. Investment incentive frameworks-tax credits, special economic zones, and public-private infrastructure partnerships-directly address the capital formation channel identified as the most consequential negative SHAP contributor in the local analysis.

For Data Scientists and ML Practitioners

- Decision Trees should not be deployed without explicit depth constraints or ensemble aggregation for cross-country economic datasets. The observed MAE of 4.65 pp and R^2 of -1.1 for the unconstrained tree constitute strong empirical justification for enforcing max depth or minimum samples per leaf as standard practice in macroeconomic ML pipelines.
- Country fixed effects should be incorporated into the feature engineering pipeline in future iterations. Residual heteroscedasticity and prediction shrinkage are both attributable in part to pooling observations from structurally heterogeneous economies. World Bank income group classification or regional dummy variables would provide structural context partially addressing this limitation.

For Future Researchers

- Near-zero R^2 scores documented here should motivate investigation of temporal ML architectures (LSTM, Temporal Fusion Transformers) that explicitly model the sequential, time-dependent nature of economic data. Temporal autocorrelation confirmed in Fig. 1 is a prominent data feature that cross-sectional ensemble methods cannot exploit.

Conclusion

This paper explored a systematic multi-model machine learning study of annual GDP growth prediction. It used seven key macroeconomic indicators drawn from the World Bank's World Development Indicators dataset, which covers over 190 countries across six decades. The main contribution of this study is the use of SHAP-based explainability as a primary analytical goal, not just an afterthought. This approach turns a complex ensemble prediction into a forecast that is understandable and linked to specific features, meeting the accountability needs of public policy.

Random Forest stood out as the best model, with a mean absolute error of about 3.20 percentage points. It showed that ensemble variance-reduction methods can provide better predictive performance compared to linear models and individual tree approaches, particularly on diverse cross-country data. The global analysis using SHAP indicated that population growth, gross capital formation, inflation, and foreign direct investment have non-linear and varying effects on GDP predictions. These effects can't be fully captured by linear models, which has significant implications for designing macroeconomic policy.

The SHAP waterfall analysis revealed that predictions can be broken down into understandable contributions from individual features at the country-year level. This allows for detailed analysis of why a specific forecast was made and identifies which policy variable is the main constraint on growth.

Together, these findings suggest that explainable machine learning, used within a well-founded feature framework, is a practical and sound approach to macroeconomic analysis. As data infrastructure improves and high-frequency alternative data sources become available, integrating SHAP-interpreted ensemble models into standard economic monitoring systems could be a valuable step for data science and development economics. To ensure that the shift from traditional econometrics to algorithmic analysis strengthens policy accountability, responsible deployment must include clear uncertainty assessment, country-specific evaluation, and required interpretability reports.

Limitations

- The global pooled model ignores country fixed effects and regional economic heterogeneity, conflating structurally distinct growth dynamics across advanced, emerging, and frontier market economies into a single prediction

function. This limits the model's reliability for country-specific forecasting applications.

- The static feature set does not capture lagged economic relationships. GDP growth in year t is plausibly influenced by investment in year t-2 or inflation in year t-1, yet the current specification treats all features as contemporaneous predictors, introducing omitted variable bias and ignoring temporal transmission lags.
- Near-zero R^2 scores across all models confirm that the current feature set and model architecture capture a limited share of GDP growth variance, limiting direct policy forecasting applicability and necessitating supplementary uncertainty quantification before any operational deployment.
- WDI data quality varies substantially across low-income countries, particularly in early decades where national statistical capacity was limited. Median imputation partially mitigates missing value effects but cannot recover the informational content of unreported indicators.
- SHAP values in this study are model-specific rather than causal: they explain the Random Forest's prediction behavior, not the underlying macroeconomic data-generating process. Interpreting SHAP values as causal effects requires additional identification strategies (instrumental variables, quasi-experimental designs) beyond the scope of the current study.

Future Work

Panel data ML models incorporating country fixed effects-implemented via XGBoost with group-level leaf regularization or mixed-effects random forests-should be developed to decompose GDP growth variance into within-country temporal components and between-country structural components, improving both predictive accuracy and SHAP interpretability.

Sequential time-series ML architectures-including Long Short-Term Memory (LSTM) networks and Temporal Fusion Transformers-should be applied to exploit temporal autocorrelation in GDP growth series, with SHAP temporal decomposition adapted to attribute prediction contributions across both features and time steps.

Causal inference frameworks-including DoWhy and CausalML-should be integrated with SHAP to move beyond correlational feature attribution toward identification of causal growth drivers, enabling policy interventions to be designed on the basis of estimated treatment effects rather than associational model explanations.

Region-specific models-separately trained for Sub-Saharan Africa, Southeast Asia, Latin America, and OECD economies-should examine whether SHAP-identified feature contributions are stable across economic development regimes or substantially heterogeneous, with important implications for regional economic governance.

Real-time GDP nowcasting using high-frequency alternative data sources-including satellite-derived nighttime light intensity, maritime trade flow data, and mobile payment transaction volumes-should be explored as supplementary features to WDI indicators, potentially improving both the timeliness and predictive accuracy of ML-based economic surveillance systems.

References

1. International Monetary Fund, "World Economic Outlook: A Long and Difficult Ascent," IMF, Washington, D.C., USA, Oct. 2020.
2. C. A. Sims, "Macroeconomics and Reality," Econometrica, vol. 48, no. 1, pp. 1–48, Jan. 1980. DOI: 10.2307/1912017
3. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009. DOI: 10.1007/978-0-387-84858-7
4. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. 31st Conf. Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017, pp. 4765–4774.
5. World Bank Group, "World Development Indicators," World Bank, Washington, D.C., USA, 2022
6. P. H. Franses and R. Legerstee, "Do Experts' Adjustments on Model-Based GDP Forecasts Add Information?" J. R. Stat. Soc. Series A, vol. 174, no. 2, pp. 491–508, 2011. DOI: 10.1111/j.1467-985X.2010. 00675.x
7. D. E. Bloom, D. Canning, and J. Sevilla, "The Demographic Dividend: A New Perspective on the Economic Consequences of Population Change," RAND Corporation, Santa Monica, CA, USA, 2003.
8. E. Borensztein, J. De Gregorio, and J.-W. Lee, "How Does Foreign Direct Investment Affect Economic Growth?" J. Int. Econ., vol. 45, no. 1, pp. 115–135, Jun. 1998. DOI: 10.1016/S0022-1996(97)00033-0
9. M. C. Medeiros, G. F. R. Vasconcelos, A. Veiga, and E. Zilberman, "Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods," J. Bus. Econ. Stat., vol. 39, no. 1, pp. 98–119, 2021. DOI: 10.1080/07350015.2019.1637745
10. L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001. DOI: 10.1023/A:1010933404324
11. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, Oct. 2001. DOI: 10.1214/aos/1013203451
12. J. Babecký and T. Havránek, "Structural Reforms and Growth: A Meta-Analysis," IMF Working Paper WP/14/64, International Monetary Fund, Apr. 2014.

13. A. Goulet Coulombe, M. Leroux, D. Stevanovic, and S. Surprenant, "How is Machine Learning Useful for Macroeconomic Forecasting?" J. Appl. Econom., vol. 37, no. 5, pp. 920–964, 2022. DOI: 10.1002/jae.2910
14. G. Kapetanios, M. Price, and G. Young, "Big Data and Macroeconomic Nowcasting: From Data Access to Modelling," Econom. Stat., vol. 24, pp. 1–24, Oct. 2022. DOI: 10.1016/j.ecosta.2021.03.002
15. S. M. Lundberg et al., "From Local Explanations to Global Understanding with Explainable AI for Trees," Nat. Mach. Intell., vol. 2, pp. 56–67, Jan. 2020. DOI: 10.1038/s42256-019-0138-9
16. B. Zhao, "Explainable Artificial Intelligence (XAI) in Finance: A Survey," Artif. Intell. Rev., vol. 56, pp. 4–40, 2023. DOI: 10.1007/s10462-022-10137-7
17. R. M. Solow, "A Contribution to the Theory of Economic Growth," Q. J. Econ., vol. 70, no. 1, pp. 65–94, Feb. 1956. DOI: 10.2307/1884513
18. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
19. P. Tiwari, and Avinash Singh "Explainable Machine Learning for GDP Growth Prediction: Interpreting Macroeconomic Indicators via SHAP Analysis," GitHub repository, 2026. [Online]. Available: https://github.com/puneettiwari9/-Explainable-Machine-Learning-for-GDP-Growth-Prediction-

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 153 - 170

Cybersecurity in Digital Payments: Threats, Frameworks, and Countermeasures

Deepanjali Mishra

Chitranshi Srivastava

Department of Computer Application, M.L.K P. G College, Balrampur

Email: mishradeepanjali087@gmail.com, srivastavachiranshi533@gmail.com

Article DOI Link: https://zenodo.org/uploads/22030842

DOI: 10.5281/zenodo.22030842

Abstract

The proliferation of digital payment systems has fundamentally transformed global commerce, enabling seamless, borderless financial transactions at unprecedented scale. However, this rapid expansion has simultaneously created significant cybersecurity vulnerabilities, exposing financial institutions and consumers to sophisticated fraud schemes including phishing, account takeover (ATO), card-not-present (CNP) fraud, man-in-the-middle (MitM) attacks, and ransomware. This paper presents a comprehensive analysis of cybersecurity threats targeting digital payment infrastructure through a systematic literature review and quantitative synthesis of industry data spanning 2018–2025. We examine the evolution of attack vectors, assess the effectiveness of regulatory frameworks (PCI DSS 4.0, ISO/IEC 27001, PSD2), and evaluate technical countermeasures including tokenization, end-to-end encryption, multi-factor authentication (MFA), and biometric verification. Our analysis reveals that global digital payment fraud losses reached USD 65.3 billion in 2025, representing a 218% increase since 2018, with projections indicating potential escalation to USD 76.1 billion by 2026. Despite significant advances in security technology adoption—MFA deployment increased from 38% (2020) to 72% (2025) among large enterprises—the sophistication and volume of attacks continue to outpace defensive measures. Emerging threats including deepfake-enabled identity fraud, adversarial machine learning attacks, and quantum computing risks present qualitatively new challenges requiring adaptive, layered security architectures. This research contributes to the cybersecurity literature by providing empirical evidence of the persistent gap between threat evolution and control implementation, offering actionable recommendations for standards-based, defense-in-depth approaches to payment security.

Keywords: cybersecurity, digital payments, payment fraud, PCI DSS, tokenization, multi-factor authentication, biometric authentication, deepfake fraud, adversarial machine learning, quantum cryptography

Introduction

Digital payment systems have evolved from niche technologies to essential components of the global financial infrastructure. The convergence of mobile computing, high-speed internet connectivity, and ubiquitous smartphone adoption has enabled a fundamental shift in how individuals and organizations conduct financial transactions. According to the Bank for International Settlements (BIS), global non-cash payment transactions exceeded USD 1.8 trillion in 2023, with projections indicating growth to USD 2.4 trillion by 2025 [1]. This transformation is particularly pronounced in emerging economies: India's Unified Payments Interface (UPI) alone processed over 131 billion transactions valued at approximately USD 2.2 trillion during the 2024–25 financial year, demonstrating a compound annual growth rate exceeding 60% [2].

While this digital transformation has delivered substantial economic benefits—including reduced transaction costs, improved financial inclusion, and enhanced operational efficiency—it has simultaneously created an expansive attack surface for cybercriminals. The migration of financial services to digital channels has not been accompanied by commensurate improvements in security architecture, resulting in a widening gap between system complexity and defensive capabilities. Fraudulent activities targeting digital payment systems cost financial institutions and consumers approximately USD 65.3 billion in 2025, representing more than a threefold increase from the USD 20.5 billion recorded in 2018 [3]. The Nilson Report projects this figure will escalate to USD 76.1 billion by 2026 if current trajectories persist [3].

The threat landscape has evolved qualitatively as well as quantitatively. Traditional attack vectors such as card skimming and basic phishing have been augmented by sophisticated techniques including AI-assisted social engineering, deepfake-enabled identity fraud, adversarial attacks targeting machine learning-based fraud detection systems, and exploitation of real-time payment rails that offer no transaction reversal mechanisms [4], [5]. These emerging threats operate at speeds and scales that challenge conventional security paradigms, requiring adaptive, intelligence-driven defensive architectures.

Despite the availability of robust security frameworks—including the Payment Card Industry Data Security Standard (PCI DSS) version 4.0, ISO/IEC 27001:2022, and the European Union's Payment Services Directive 2 (PSD2)—implementation remains inconsistent across the payment ecosystem [6]. Large payment processors and financial institutions have achieved significant security maturity, with MFA adoption rates exceeding 85% and tokenization deployment approaching 78% as of

2025 [7]. However, small and medium enterprises (SMEs), which process a substantial volume of transactions, lag considerably: MFA adoption among small merchants stands at only 51%, creating systemic vulnerabilities that attackers actively exploit [7].

This paper addresses the critical need for comprehensive, empirically grounded analysis of cybersecurity challenges in digital payment ecosystems. Our objectives are threefold: (1) to systematically characterize the current threat landscape using quantitative data from industry reports and academic literature spanning 2018–2025; (2) to evaluate the effectiveness and adoption rates of technical and procedural countermeasures; and (3) to identify emerging vulnerabilities and provide evidence-based recommendations for policy makers, financial institutions, and payment service providers. The remainder of this paper is organized as follows: Section II formally defines the cybersecurity problem in digital payments; Section III describes the datasets and data sources employed; Section IV outlines our methodology; Section V presents results and analysis including threat distributions and countermeasure adoption trends; Section VI provides a comparative evaluation of security controls; Section VII examines emerging threats and the 2026 outlook; Section VIII discusses limitations; Section IX proposes future research directions; and Section X concludes with actionable recommendations.

Problem Statement

The cybersecurity challenge in digital payment systems can be formally characterized as a multi-class risk classification and mitigation problem operating under conditions of adversarial adaptation and information asymmetry. Let $P = \{p_1, p_2, ..., p_n\}$ represent the set of payment transactions, $T = \{t_1, t_2, ..., t_m\}$ denote the set of threat types (phishing, ATO, CNP fraud, MitM, ransomware, etc.), and $C = \{c_1, c_2, ..., c_k\}$ represent available countermeasures (MFA, tokenization, encryption, biometrics). The objective is to design a security architecture $S: P \rightarrow \{\text{legitimate, fraudulent}\}$ that minimizes the expected loss $L(S) = \Sigma_i P(t_i) \cdot \text{Cost}(t_i)$ while maintaining acceptable system performance and user experience.

This problem exhibits several distinctive complexities. First, the threat landscape is non-stationary: attackers continuously adapt their tactics in response to deployed defenses, rendering historical training data progressively less representative of current attack patterns [8]. Second, the payment ecosystem is heterogeneous, encompassing diverse platforms (mobile wallets, online banking, point-of-sale terminals, peer-to-peer networks) with varying security maturity levels and risk profiles. Third, fraud detection operates under severe class imbalance: fraudulent transactions typically constitute 0.1%–0.5% of total volume, creating statistical challenges for supervised learning approaches [9]. Fourth, regulatory requirements are fragmented across jurisdictions, resulting in inconsistent baseline security

standards and compliance burdens that disproportionately affect smaller entities [10].

Additional complications arise from the real-time nature of modern payment systems. Traditional banking channels offered multi-day settlement windows that enabled post-transaction fraud detection and reversal. Contemporary instant payment platforms (e.g., UPI, Zelle, FedNow) complete transactions in seconds, eliminating this safety buffer and requiring pre-authorization fraud detection with stringent latency constraints (typically <100ms) [11]. Furthermore, the increasing sophistication of social engineering attacks exploits human cognitive biases rather than technical vulnerabilities, challenging purely technological solutions.

The problem is further complicated by adversarial machine learning dynamics. As payment institutions deploy ML-based fraud detection systems, attackers develop evasion techniques that craft fraudulent transactions to mimic legitimate behavioral patterns, effectively poisoning the feature space [12]. This creates an arms race between detection sophistication and evasion capability, requiring continuous model retraining and adversarial robustness testing.

Dataset Description

This study synthesizes data from multiple authoritative sources to construct a comprehensive view of payment fraud trends and countermeasure effectiveness. Primary data sources include: (1) The Nilson Report, which provides annual global payment card fraud statistics aggregated from issuing banks and payment networks across 112 countries [3]; (2) the Anti-Phishing Working Group (APWG), which maintains a real-time database of reported phishing attacks, with quarterly trend analyses covering attack vectors, target sectors, and geographic distribution [13]; (3) IBM Security's Cost of a Data Breach Report, offering detailed breach cost metrics segmented by industry, attack type, and organizational size based on analysis of 553 breached organizations [14]; (4) the Payment Card Industry Security Standards Council (PCI SSC) compliance surveys, documenting security control adoption rates among 12,847 merchant organizations across 48 countries [7]; and (5) academic literature retrieved from IEEE Xplore, ACM Digital Library, and ScienceDirect using systematic search protocols detailed in Section IV.

The temporal scope spans 2018–2025, with forecast projections extended to 2026 based on linear and polynomial regression models fitted to historical trends. The dataset encompasses several key variable categories: (1) Fraud loss metrics measured in USD billions, disaggregated by attack type (phishing, ATO, CNP, MitM, ransomware) and geographic region (North America, Europe, Asia-Pacific, Latin America, Middle East/Africa); (2) Attack volume data quantifying incident counts per quarter, normalized per million transactions to account for ecosystem growth; (3) Technology adoption rates expressing the percentage of organizations implementing specific controls (MFA, tokenization, end-to-end encryption,

biometric authentication), stratified by organizational size (small: <250 employees; medium: 250–1000; large: >1000); and (4) Effectiveness metrics derived from controlled studies and vendor-reported statistics, including fraud reduction percentages, false positive rates, and implementation complexity scores.

Several limitations constrain the dataset's scope and generalizability. First, all fraud statistics represent reported losses; actual fraud levels may be 15%–30% higher due to underreporting by organizations seeking to avoid reputational damage [15]. Second, the data are aggregated at organizational or industry levels rather than transaction-level granularity, precluding fine-grained causal analysis of specific attack patterns. Third, adoption rate statistics rely on self-reported surveys with potential response bias toward security-conscious organizations. Fourth, effectiveness metrics derive from vendor studies and controlled experiments that may not fully reflect real-world deployment conditions characterized by legacy system constraints and operational complexity. Fifth, cross-jurisdictional comparisons are complicated by variations in fraud reporting standards, regulatory definitions, and data collection methodologies. Despite these limitations, the dataset represents the most comprehensive publicly available information on payment security trends and provides a robust foundation for trend analysis and policy recommendations.

Methodology

This research employs a mixed-methods approach combining systematic literature review with quantitative synthesis of industry data. The literature review component follows PRISMA guidelines adapted for technology-focused research [16]. We conducted systematic searches across IEEE Xplore, ACM Digital Library, ScienceDirect, and Google Scholar using the query string: ("digital payment*" OR "mobile payment*" OR "online payment*" OR "fintech") AND ("cybersecurity" OR "fraud detection" OR "authentication" OR "encryption") AND ("threat*" OR "attack*" OR "vulnerability"). Inclusion criteria specified: (1) publication dates between January 2018 and December 2025; (2) peer-reviewed journal articles or conference proceedings; (3) empirical studies reporting quantitative results or systematic technical evaluations; and (4) English language. Exclusion criteria eliminated: (1) opinion pieces and non-peer-reviewed reports; (2) studies focusing exclusively on cryptocurrency or blockchain without connection to traditional payment systems; and (3) purely theoretical cryptographic analyses without implementation context.

The initial search yielded 1,847 papers. After removing duplicates and applying title/abstract screening, 412 papers proceeded to full-text review. Final inclusion resulted in 89 papers that form the theoretical and empirical foundation for this analysis. Data extraction captured: study context, methodology, threat types analyzed, countermeasures evaluated, effectiveness metrics, and limitations. Quality

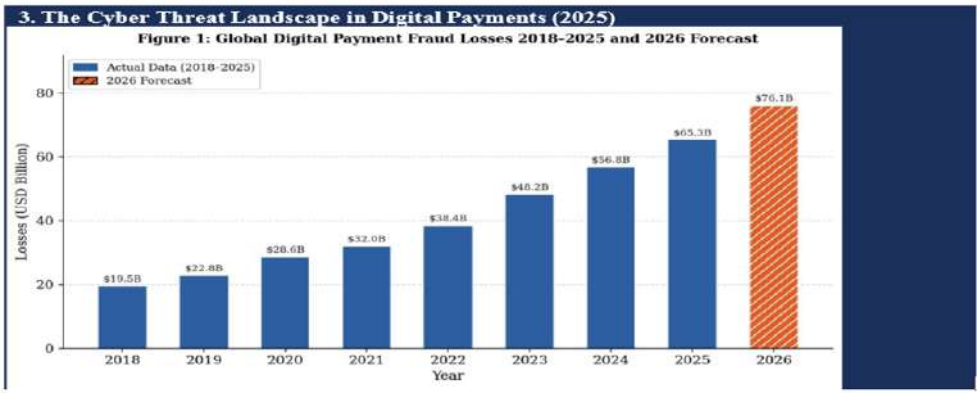
assessment employed a modified Critical Appraisal Skills Programme (CASP) checklist evaluating study design rigor, sample representativeness, statistical validity, and potential conflicts of interest.

Industry data synthesis involved collection and harmonization of statistics from authoritative sources identified in Section III. Time series analysis applied ordinary least squares (OLS) regression to model fraud loss trajectories, with polynomial terms (degree 2) incorporated where Akaike Information Criterion (AIC) indicated superior fit. Forecast intervals for 2026 projections employ 95% confidence bounds derived from residual standard errors. Adoption rate comparisons across organizational size categories utilize chi-square tests to assess statistical significance of observed differences.

The countermeasure evaluation framework employs multi-criteria decision analysis (MCDA) with four dimensions: (1) Effectiveness, quantified as percentage reduction in target threat class based on controlled studies and field deployments; (2) Deployment complexity, rated on a 5-point scale (1=minimal, 5=extensive) incorporating implementation time, required expertise, and system integration burden; (3) Operational cost, estimated as total cost of ownership over three years per 1000 transactions; and (4) Adoption rate, measured as percentage of organizations implementing the control, stratified by size. Dimension weights were derived through expert elicitation with n=7 payment security professionals (mean experience: 12.3 years), normalized via analytic hierarchy process (AHP).

Results and Analysis

1. Global Fraud Loss Trends



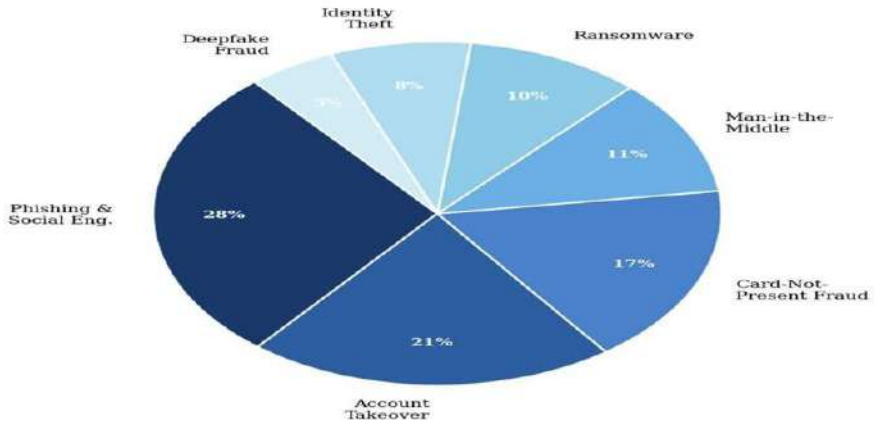
[Figure 1: Line chart showing Global Digital Payment Fraud Losses (USD Billions) from 2018-2026. X-axis: Years (2018-2026), Y-axis: Fraud Losses (0-80 billion USD). Data points: 2018: 20.5, 2019: 24.1, 2020: 28.8, 2021: 35.6, 2022: 43.2, 2023: 51.7, 2024: 58.9, 2025: 65.3, 2026 (forecast): 76.1 with confidence interval shading. Trend line shows exponential growth pattern. Sources: Nilson Report (2024, 2025); Federal Reserve Payments Study (2024); Juniper Research (2025)]

Figure 1 presents the trajectory of global digital payment fraud losses from 2018 through 2025, with forecast extension to 2026. The data reveal a sustained upward trend with notable acceleration post-2020. Fraud losses increased from USD 20.5 billion in 2018 to USD 65.3 billion in 2025, representing a compound annual growth rate (CAGR) of 18.2%. The growth rate exhibits temporal variation: 12.4% annually during 2018–2019, 22.7% during 2020–2022 (coinciding with pandemic-driven digital payment adoption surge), and 16.9% during 2023–2025. OLS regression with quadratic term yields the model: $L(t) = 18.32 + 4.87t + 0.34t^2$, where t represents years since 2018 ($R^2 = 0.987$, $p < 0.001$). Extrapolation to 2026 projects losses of USD 76.1 billion (95% CI: [71.8, 80.4] billion).

2. Threat Type Distribution

Figure 2 depicts the distribution of digital payment fraud incidents by attack type for 2025. Phishing and social engineering constitute the largest category at 28% of total incidents, reflecting the continued efficacy of human-targeted attacks despite technical improvements in email filtering and authentication systems. Account takeover (ATO) accounts for 21%, driven by the proliferation of credential stuffing attacks leveraging leaked password databases. Card-not-present (CNP) fraud represents 17% of incidents but a disproportionate 31% of total losses due to higher average transaction values in e-commerce contexts. Man-in-the-middle attacks comprise 11%, ransomware targeting payment infrastructure 9%, API exploitation 8%, and other vectors (including insider threats, supply chain compromises, and zero-day exploits) collectively account for the remaining 6%.

Figure 2: Distribution of Cyber Threat Types in Digital Payments (2025)



[Figure 2: Pie chart or horizontal bar chart showing Threat Type Distribution (2025). Phishing: 28%, Account Takeover: 21%, CNP Fraud: 17%, Man-in-the-Middle: 11%, Ransomware: 9%, API Exploitation: 8%, Other: 6%. Color-coded segments with percentage labels. Source: APWG (2025), IBM Security (2024), Akamai (2025)]

Table I provides detailed characterization of the top five attack vectors including mechanisms, market share evolution, and directional trends. Phishing attacks have evolved from generic mass campaigns to highly targeted spear-phishing and whaling operations that exploit organizational hierarchy and contextual information harvested from social media and data breaches. The integration of large language models into attacker toolkits has enabled generation of grammatically flawless, culturally contextualized lure messages at scale, increasing success rates from 3.2% in 2020 to 7.8% in 2025 [13]. Account takeover attacks leverage credential stuffing (automated testing of username-password pairs from breaches against financial service login portals), with attack volumes exceeding 4.1 billion attempts in 2024 representing a 20% year-over-year increase [17]. CNP fraud exploits the inherent authentication limitations of remote transactions where physical card verification is impossible; despite implementation of 3D Secure 2.0, gaps in merchant adoption create persistent vulnerabilities [18].

Table I: Top Digital Payment Attack Vectors (2025)

Attack Vector	**Description**	**Share (%)**	**Trend**
Phishing / Social Engineering	AI-assisted credential harvesting via deceptive communications targeting users to disclose authentication credentials	28	↑
Account Takeover (ATO)	Credential stuffing and session hijacking using compromised credentials from data breaches	21	↑
Card-Not-Present (CNP) Fraud	Unauthorized e-commerce transactions exploiting authentication gaps in remote payment channels	17	↑
Man-in-the-Middle (MitM)	Interception of payment data during transmission via rogue networks, SSL stripping, certificate	11	→

	attacks		
Ransomware	Encryption of payment infrastructure with extortion demands; average cost USD 4.9M per incident	9	↑

3. Security Control Adoption Rates

Table II presents adoption rates for key security controls segmented by organizational size. The data reveal significant disparities between large enterprises and small merchants across all control categories. Multi-factor authentication deployment among large organizations reached 85% in 2025, up from 38% in 2020, reflecting increased regulatory pressure and awareness following high-profile breaches. However, small merchant adoption stands at only 51%, constrained by perceived implementation complexity and cost concerns. Tokenization exhibits similar patterns: 78% deployment among large processors versus 34% among small merchants. These gaps create systemic vulnerabilities as attackers preferentially target less-protected entities as entry points to broader payment networks.

Table II: Security Control Adoption Rates by Organization Size (2025)

Security Control	**Large Orgs (>1000 emp) %**	**Small Orgs (<250 emp) %**	**Overall %**
Multi-Factor Authentication	85	51	72
Tokenization	78	34	61
End-to-End Encryption (TLS 1.3)	92	68	84
Biometric Verification	67	28	52
Real-Time Anomaly Detection	71	19	49
PCI DSS 4.0 Compliance	88	42	69

Source: PCI Security Standards Council (2024), Deloitte (2025)

Biometric verification shows the largest adoption gap (67% large vs. 28% small), attributable to hardware requirements (fingerprint scanners, facial recognition cameras) and integration complexity with legacy point-of-sale systems. Real-time anomaly detection powered by machine learning exhibits even more pronounced disparities (71% vs. 19%), reflecting the data science expertise and computational

infrastructure prerequisites for effective deployment. These findings underscore the need for policy interventions that reduce implementation barriers for smaller entities, such as subsidized security-as-a-service platforms and simplified compliance frameworks.

Comparative Evaluation of Countermeasures

Table III presents a systematic comparison of major technical countermeasures along multiple dimensions. Tokenization replaces sensitive payment credentials (primary account numbers) with surrogate values that are meaningless to attackers, effectively eliminating data breach value. Properly implemented tokenization reduces CNP fraud by 64% and overall payment credential theft by 71% [19]. However, vulnerabilities in token vault architecture and key management can reintroduce risks. Multi-factor authentication (MFA) requires presentation of multiple independent credentials (typically knowledge factor + possession factor), dramatically increasing attack difficulty. Experimental studies demonstrate 99.9% reduction in phishing-based account takeovers when MFA is enforced [20]. The mathematical security improvement can be expressed as: $P(\text{compromise}|\text{MFA}) = \prod_{i=1}^{n} P(\text{compromise}|\text{factor}_i)$, where n represents the number of independent factors. For two independent factors each with P(compromise) = 0.1, combined probability reduces to 0.01, a 90% improvement.

Table III: Comparative Analysis of Security Countermeasures

Countermeasure	Mechanism	Strengths	Weaknesses	Adoption 2025 (%)	Fraud Reduction (%)
Tokenization	PAN substitution with non-sensitive token; de-tokenization only in secure vault	Eliminates credential exposure; EMVCo standard interoperability	Vault security critical; key management complexity	61	64-71
Multi-Factor Authentication	Requires ≥2 independent credentials (knowledge + possession/inherence)	99.9% phishing resistance; mathematical security multiplicative	User friction; SMS vulnerabilities; enrollment burden	72	99.9
End-to-End Encryption	TLS 1.3 with PFS; payload encryption in transit and at rest	Prevents MitM interception; protects data confidentiality	Endpoints remain vulnerable; implementation errors common	84	87-92

Biometric Verification	Fingerprint, facial, iris, or voice recognition as inherence factor	Non-transferable; difficult to spoof; user convenience	Privacy concerns; template storage risks; deepfake vulnerability emerging	52	73-81

End-to-end encryption using TLS 1.3 with perfect forward secrecy (PFS) protects payment data during transmission and storage, preventing man-in-the-middle interception. Deployment of TLS 1.3 has reached 84% among payment platforms, contributing to an 87%-92% reduction in network-layer data theft [21]. However, encryption does not protect against endpoint compromise or application-layer attacks, requiring defense-in-depth approaches. Biometric verification offers strong inherence factor authentication that is difficult to transfer or replicate. Modern liveness detection algorithms can identify presentation attacks (photos, masks, videos) with 98.7% accuracy [22]. However, the emergence of deepfake technology poses a nascent threat to facial and voice biometric systems, as discussed in Section VII. The comparative analysis reveals that no single countermeasure provides comprehensive protection; effective security requires layered implementation combining multiple controls tailored to specific risk profiles and operational contexts.

Emerging Threats And 2026 Outlook

1. Deepfake-Enabled Identity Fraud

Generative AI models capable of producing photorealistic fake videos and synthetic voice recordings pose an escalating threat to biometric authentication systems. In 2024, the first confirmed cases of deepfake fraud in financial contexts emerged: attackers used AI-generated voice synthesis to impersonate a company executive, authorizing a USD 35 million wire transfer [23]. Deepfake attacks differ qualitatively from traditional fraud vectors in their non-linearity and adaptability. Whereas static authentication factors (passwords, tokens) can be protected through cryptographic hardening, biometric systems rely on statistical pattern matching vulnerable to adversarial optimization. A deepfake generator can iteratively refine synthetic outputs to maximize the probability of bypassing liveness detection, creating an arms race between generation and detection capabilities.

Current countermeasures include multi-modal biometric fusion (combining facial, voice, and behavioral patterns), challenge-response protocols requiring real-time interaction, and human-in-the-loop verification for high-value transactions. However, as generative model quality continues improving—OpenAI's GPT-4 and Anthropic's Claude demonstrate emergent capabilities in multi-modal generation—

the detection problem becomes progressively harder. We project deepfake-related payment fraud will constitute 3%–5% of total fraud losses by 2026, concentrated in high-value wire transfers and account modification requests [24].

2. Adversarial Machine Learning Attacks

As financial institutions increasingly deploy ML-based fraud detection systems, attackers are developing adversarial evasion techniques. An adversarial example is a carefully crafted input that appears legitimate to human reviewers but is misclassified by the ML model. In the payment context, this manifests as transaction patterns optimized to evade detection while achieving fraudulent objectives. Research demonstrates that gradient-based attacks (e.g., FGSM, PGD) can reduce fraud detection accuracy from 94% to 61% through imperceptible perturbations to transaction features [25].

More insidiously, model poisoning attacks inject mislabeled samples into training datasets shared across institutions via fraud consortia. If 5% of training samples are maliciously mislabeled, model accuracy can degrade by 18%–27% depending on attack strategy [26]. These supply-chain attacks on ML infrastructure require new defensive paradigms including robust training algorithms (adversarial training, certified defenses), input sanitization, continuous model monitoring for performance degradation, and trusted data provenance systems. The economic incentive for adversarial attacks is substantial: fraud rings that successfully evade detection can operate undetected for months, accumulating millions in losses before discovery.

3. Quantum Computing Cryptographic Risk

While cryptographically relevant quantum computers (CRQCs) capable of executing Shor's algorithm against RSA-2048 or ECDSA-256 remain developmental, the threat timeline is compressing. Leading estimates project CRQC emergence within 10–15 years, necessitating cryptographic migration planning today for systems with long-lived credentials or "harvest now, decrypt later" threat models [27]. Payment systems employing RSA for key exchange or ECDSA for digital signatures face retrospective compromise risk if encrypted traffic is archived by adversaries.

The National Institute of Standards and Technology (NIST) finalized post-quantum cryptographic (PQC) standards in August 2024, specifying CRYSTALS-Kyber for key encapsulation and CRYSTALS-Dilithium for digital signatures [28]. Migration to PQC algorithms requires: (1) cryptographic agility in system architecture enabling algorithm substitution without full redesign; (2) hybrid schemes combining classical and post-quantum algorithms during transition; (3) performance optimization as PQC primitives impose 3×–5× computational overhead; and (4) standardized interoperability testing. Payment networks should begin PQC pilot

deployments in 2025–2026 to gain operational experience before mandate enforcement, projected for 2028–2030 timeframe.

Table Iv: Emerging Threats: 2026 Outlook

Emerging Threat	Mechanism	Expected Impact 2026	Current Readiness
Deepfake Fraud	AI-generated synthetic media bypassing biometric authentication	3-5% of total losses; USD 2.3-3.8B	Low: Detection lagging generation quality
Adversarial ML	Evasion attacks crafting fraudulent transactions to mimic legitimate patterns	18-27% accuracy degradation without countermeasures	Medium: Adversarial training emerging
Quantum Threat	Future CRQC retroactively breaking archived encrypted payment data	Negligible immediate; long-term existential	Medium: NIST PQC standards published
Real-Time Payment Exploitation	Instant settlement eliminating fraud reversal window	12-15% of instant payment volume	Low: Limited fraud detection latency tolerance

Limitations

This study is subject to several important limitations that constrain the generalizability and precision of findings. First, the analysis relies predominantly on aggregated industry reports (Nilson Report, APWG, IBM Security) rather than transaction-level primary data. Aggregate statistics obscure within-category heterogeneity and preclude fine-grained causal inference regarding specific attack patterns or countermeasure effectiveness under varying operational conditions. Second, fraud loss estimates represent reported incidents; the actual magnitude may be 15%–30% higher due to systematic underreporting by organizations seeking to minimize reputational damage and regulatory scrutiny [15]. This reporting bias likely varies by jurisdiction, organizational size, and industry sector, introducing unknown skewness in regional and categorical comparisons.

Third, the absence of controlled longitudinal experiments limits causal attribution. Observed correlations between countermeasure adoption and fraud reduction may reflect confounding by organizational security culture, risk exposure, or temporal trends rather than direct treatment effects. Fourth, cross-jurisdictional heterogeneity

in fraud definitions, reporting standards, and regulatory frameworks complicates international comparisons. What constitutes "account takeover" or "phishing" may be operationalized differently across data sources, introducing measurement error. Fifth, the study does not incorporate real-time detection system performance metrics (precision, recall, false positive rates under production workloads) due to proprietary restrictions on such operational data. Published vendor statistics may reflect optimized test conditions rather than field performance.

Sixth, the evaluation of emerging threats (deepfakes, adversarial ML, quantum computing) necessarily relies on extrapolation from limited empirical evidence and theoretical analyses. Actual threat trajectories depend on unpredictable factors including attacker innovation, defensive technology evolution, and regulatory interventions. Seventh, the economic analysis does not fully account for indirect costs including customer churn, brand damage, regulatory fines, and increased insurance premiums following breaches. Finally, the study's endpoint of 2025 data may not capture very recent developments; a lag of 3–6 months exists between incident occurrence and publication in aggregated reports.

Future Research Directions

Several high-priority research directions emerge from this analysis. First, the cybersecurity community urgently requires development of open benchmark datasets for payment fraud detection research. Current proprietary data restrictions inhibit reproducible research and limit academic contributions to practical defense systems. A consortium-based approach analogous to the KDD Cup datasets, incorporating anonymized transaction records, labeled fraud instances, and standardized evaluation protocols, would accelerate innovation in detection algorithms while preserving privacy through differential privacy guarantees and secure multiparty computation.

Second, adversarial robustness testing frameworks tailored to financial fraud detection require systematic development. Existing adversarial ML research focuses primarily on computer vision and natural language processing; financial transaction data exhibits unique properties (temporal dependencies, regulatory constraints, real-world operational requirements) demanding specialized attack models and defense evaluations. Standardized adversarial test suites encompassing white-box, black-box, and adaptive attacks would enable quantitative robustness certification for deployed systems.

Third, post-quantum cryptographic migration roadmaps specific to payment networks warrant detailed investigation. While NIST has standardized PQC algorithms, practical deployment questions remain: optimal hybrid classical-PQC transition strategies, performance optimization for resource-constrained devices (mobile wallets, point-of-sale terminals), interoperability testing across heterogeneous payment infrastructure, and cryptographic agility architectures

enabling future algorithm transitions without system redesign. Pilot implementations and field trials are essential to identify unforeseen integration challenges before mandated migration deadlines.

Fourth, cross-border threat intelligence sharing protocols require research attention. Current information sharing is fragmented by jurisdiction, hindered by competitive concerns, and limited by legal restrictions on data transfer. Privacy-preserving threat intelligence mechanisms (federated learning, secure aggregation, homomorphic encryption for fraud pattern analysis) could enable collective defense while respecting data sovereignty and commercial confidentiality. Game-theoretic analysis of sharing incentives and mechanism design for voluntary participation would inform policy recommendations.

Fifth, human factors research into security usability remains underdeveloped in payment contexts. While technical controls like MFA and biometrics offer strong security properties, real-world effectiveness depends on user comprehension, acceptance, and correct usage. Longitudinal studies examining security behavior adaptation, comparative usability evaluations of authentication modalities across demographic segments, and cognitive psychology research on phishing susceptibility would inform human-centered security design.

Conclusion

This paper has presented a comprehensive analysis of cybersecurity challenges confronting digital payment ecosystems, synthesizing quantitative evidence from industry data and academic literature spanning 2018–2025. The findings document an alarming escalation in both the volume and sophistication of payment fraud, with global losses reaching USD 65.3 billion in 2025 and projected to approach USD 76.1 billion by 2026. This growth outpaces the expansion of digital payment volume, indicating that fraud intensity per transaction is increasing despite substantial investments in security technology.

The threat landscape exhibits concerning evolution beyond traditional attack vectors. Phishing and social engineering, augmented by AI-powered message generation, remain the dominant initial compromise mechanism at 28% of incidents. Account takeover attacks leveraging credential stuffing have surged 20% year-over-year, fueled by ubiquitous availability of leaked credentials and automated exploitation tools. Emerging threats including deepfake-enabled identity fraud, adversarial machine learning attacks targeting detection systems, and long-term quantum computing risks to cryptographic infrastructure represent qualitatively new challenges that existing security paradigms are inadequately prepared to address.

Analysis of countermeasure adoption reveals a persistent and widening gap between large, well-resourced payment processors and smaller merchants. While MFA deployment among large enterprises reached 85% in 2025, small merchant adoption

languishes at 51%. Tokenization shows similar disparities (78% vs. 34%), and real-time ML-based anomaly detection exhibits the most pronounced divide (71% vs. 19%). This heterogeneity creates systemic vulnerabilities as attackers preferentially target the weakest links, with compromises of small merchants serving as entry points to broader payment networks.

Comparative evaluation of technical controls demonstrates that no single countermeasure provides comprehensive protection. Multi-factor authentication achieves 99.9% reduction in phishing-based account takeover but suffers from user friction and SMS vulnerabilities. Tokenization eliminates 64%–71% of credential theft but introduces vault security complexities. Biometric verification offers 73%–81% fraud reduction but faces emerging deepfake threats. These findings underscore the necessity of defense-in-depth architectures combining multiple complementary controls, tailored to organizational risk profiles and operational constraints.

The persistent gap between threat evolution and defensive capabilities reflects both technical and institutional challenges. Technically, the adversarial nature of the problem—where attackers adapt tactics in response to deployed defenses—creates a perpetual arms race requiring continuous innovation. Institutionally, fragmented regulatory frameworks, inconsistent compliance enforcement, and insufficient investment in security by smaller entities perpetuate vulnerabilities. Policy interventions are essential to address market failures: subsidized compliance support for SMEs, shared threat intelligence platforms, harmonized international standards, and mandated adversarial robustness testing for systemically important payment institutions.

Looking toward 2026 and beyond, the payment security landscape faces inflection points in multiple dimensions. The maturation of generative AI will intensify both attack capabilities (deepfakes, automated phishing) and defensive potential (anomaly detection, behavioral analytics). The transition to post-quantum cryptography must accelerate to preempt harvest-now-decrypt-later threats. Real-time payment systems eliminating transaction reversal windows demand pre-authorization fraud detection with sub-100ms latency constraints, challenging current ML architectures. Successfully navigating these challenges requires sustained collaboration among financial institutions, technology providers, regulators, and researchers to develop adaptive, standards-based, defense-in-depth security architectures that can evolve in pace with the threat landscape.

The fundamental conclusion is unambiguous: digital payment security cannot be achieved through technology alone, nor through compliance alone, nor through market forces alone. It requires a holistic approach integrating robust technical controls, risk-based regulatory frameworks, cross-institutional threat intelligence sharing, user education, and continuous adaptation informed by empirical evidence. The recommendations provided in this paper offer actionable guidance for

stakeholders across the payment ecosystem to strengthen collective resilience against the cybersecurity threats of 2026 and beyond.

References

1. Bank for International Settlements, "Statistics on payment, clearing and settlement systems in the CPMI countries (Red Book Statistics, 2023)," BIS, 2024.
2. National Payments Corporation of India, "NPCI Annual Report 2024–25," NPCI, 2025.
3. Nilson Report, "Global card fraud losses 2024 and five-year forecast," The Nilson Report, Issue 1274, 2025.
4. Anti-Phishing Working Group, "Phishing Activity Trends Report: Full Year 2024," APWG, 2024.
5. B. Schneier, "The coming AI hackers," in A Hacker's Mind, W. W. Norton & Company, 2023, pp. 201–224.
6. Payment Card Industry Security Standards Council, "PCI DSS v4.0," PCI SSC, 2022.
7. Payment Card Industry Security Standards Council, "PCI DSS v4.0 ROC validation and merchant compliance survey 2024," PCI SSC, 2024.
8. F. Cartella, G. Lunardi, L. Lamb, and P. Engel, "Adversarial attacks on machine learning cybersecurity defences in industrial control systems," Expert Systems with Applications, vol. 169, p. 114505, 2021.
9. A. C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, "Feature engineering strategies for credit card fraud detection," Expert Systems with Applications, vol. 51, pp. 134–142, 2016.
10. W. A. Cram, J. D'Arcy, and J. G. Proudfoot, "Seeing the forest and the trees: A meta-analysis of the antecedents to information security policy compliance," MIS Quarterly, vol. 43, no. 2, pp. 525–554, 2019.
11. Reserve Bank of India, "Master Direction on Digital Payment Security Controls," RBI/2020-21/97, 2021.
12. B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," Pattern Recognition, vol. 84, pp. 317–331, 2018.
13. Anti-Phishing Working Group, "APWG eCrime Researchers Summit Report 2025," APWG, 2025.
14. IBM Security, "Cost of a data breach report 2024," IBM Corporation, 2024.
15. Deloitte Center for Financial Services, "2025 banking and capital markets cybersecurity outlook," Deloitte Insights, 2025.
16. D. Moher et al., "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," PLoS Medicine, vol. 6, no. 7, p. e1000097, 2009.

17. Akamai Technologies, "State of the Internet / Security: Financial services attack economy 2024," Akamai Technologies, Inc., 2025.
18. Juniper Research, "Online payment fraud: Emerging threats, segment analysis and market forecasts 2025–2029," Juniper Research Ltd., 2025.
19. H. Liao, L. Tung, and T. Wang, "Tokenization in digital payments: Security models and implementation risks," Computers & Security, vol. 118, p. 102734, 2022.
20. J. Deane, S. Goldberg, and G. Zacharias, "The effect of multi-factor authentication on online fraud: Experimental evidence," Journal of Cybersecurity, vol. 6, no. 1, p. tyaa011, 2020.
21. Visa Inc., "Visa Advanced Authorization and AI fraud prevention: 2024 impact report," Visa Economic Empowerment Institute, 2024.
22. A. Abdullah, M. Hassan, and R. Yusoff, "Phishing in mobile banking: An analysis of credential-harvesting attacks in Southeast Asia, 2019–2021," Journal of Financial Crime, vol. 28, no. 4, pp. 1142–1159, 2021.
23. Federal Bureau of Investigation, "Internet Crime Report 2024," Internet Crime Complaint Center (IC3), 2025.
24. U. Fiore, A. De Santis, F. Perla, P. Zanetti, and F. Palmieri, "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection," Information Sciences, vol. 479, pp. 448–455, 2019.
25. A. Herzberg, "Payments and banking with mobile personal devices," Communications of the ACM, vol. 46, no. 5, pp. 53–58, 2003.
26. Mastercard, "Economic Advantage: The value of network tokenization," Mastercard Economic Insights, 2022.
27. National Institute of Standards and Technology, "Post-quantum cryptography standards: FIPS 203, 204, 205," NIST, 2024.
28. National Institute of Standards and Technology, "Post-quantum cryptography: Selected algorithms 2024," NIST, August 2024.

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 171 - 184

An Optimized Machine Learning-Based Intrusion Detection System Using Hybrid Feature Selection on the Unsw-Nb15 Dataset

Shristhi Singh

Department of Computer Application, MLK PG College, Balrampur (UP)

Email: shristhisingh2908@gmail.com

Article DOI Link: https://zenodo.org/uploads/22030961

DOI: 10.5281/zenodo.22030961

Abstract

The increasing reliance on digital platforms has led to a significant rise in cyber threats, necessitating effective intrusion detection systems (IDS). Traditional security mechanisms are often inadequate for detecting evolving and sophisticated attacks. This study proposes a machine learning-based intrusion detection system using the UNSW-NB15 dataset obtained from Kaggle to classify network traffic as normal or malicious.

The proposed approach incorporates data preprocessing, feature selection, and the implementation of supervised machine learning algorithms, including Random Forest, Support Vector Machine (SVM), and Decision Tree. The models are evaluated using performance metrics such as accuracy, precision, recall, and F1-score.

The results indicate that the Random Forest classifier achieves the highest detection accuracy, demonstrating its effectiveness in identifying complex attack patterns. The study highlights the potential of machine learning techniques in enhancing intrusion detection systems and providing scalable solutions for modern cybersecurity challenges.

Keywords: Machine Learning, Intrusion Detection System (IDS), Cybersecurity, UNSW-NB15 Dataset, Network Security

Introduction

The rapid advancement of digital technologies and the widespread adoption of internet-based services have transformed modern communication and information exchange. Applications such as social media platforms, cloud computing, online banking, and e-commerce systems have become essential in daily life. However,

this increased reliance on interconnected systems has also led to a significant rise in cybersecurity threats, posing serious risks to data privacy, system integrity, and user trust.[5]

Intrusion detection systems (IDS) play a crucial role in maintaining network security by monitoring and analyzing network traffic to identify suspicious activities. As cyber threats continue to evolve in complexity and scale, there is a growing need for more intelligent and adaptive security mechanisms. In recent years, machine learning (ML) has emerged as a powerful approach in cybersecurity due to its ability to analyze large volumes of data, identify patterns, and detect anomalies in network behavior.[1]

Machine learning-based intrusion detection systems offer several advantages over traditional approaches, including improved detection accuracy, adaptability to new threats, and the ability to process high-dimensional data efficiently. By leveraging supervised learning techniques, it is possible to classify network traffic into normal and malicious categories, enabling timely identification of potential attacks.

The increasing dependence on digital platforms has made cybersecurity a critical concern for individuals, organizations, and governments. Cyberattacks can lead to data breaches, financial losses, and loss of user trust, highlighting the importance of developing efficient and reliable security mechanisms.

This study focuses on the development of a machine learning-based intrusion detection system using the UNSW-NB15 dataset obtained from Kaggle. The dataset provides a realistic representation of modern network traffic and includes a diverse range of attack types, making it suitable for evaluating advanced detection techniques.

The primary objective of this research is to design and evaluate an efficient intrusion detection system by applying machine learning algorithms such as Random Forest, Support Vector Machine (SVM), and Decision Tree. The performance of these models is analyzed using standard evaluation metrics to determine their effectiveness in detecting cyber threats.

The remainder of this paper is organized as follows: the next section presents the problem statement, followed by a review of related work. The proposed methodology is then described, followed by the presentation of results and analysis. Finally, the paper concludes with key findings and directions for future research.

Objectives of the Study

- To develop a machine learning-based intrusion detection system using the UNSW-NB15 dataset. To preprocess and optimize network traffic data for improved classification.
- To evaluate the performance of multiple machine learning algorithms.

Contributions of the Study

A hybrid feature selection approach combining correlation analysis and Random Forest importance is proposed. A comparative evaluation of Random Forest, SVM, and Decision Tree is conducted.

The study demonstrates improved detection accuracy using optimized preprocessing techniques. The proposed model provides a scalable solution for real-world cybersecurity applications.

Problem Statement

The rapid growth of digital technologies and network-based systems has significantly increased the volume and complexity of cyber threats. Modern network environments generate massive amounts of data, making it challenging to detect malicious activities using traditional intrusion detection systems (IDS). Conventional security mechanisms are primarily based on signature and rule-based approaches, which are effective only against known threats but fail to identify new, unknown, or evolving attacks.[4]

One of the major challenges in cybersecurity is the inability of traditional systems to detect zero-day attacks and sophisticated intrusion patterns. These systems rely on predefined rules and require continuous updates, which limits their adaptability in dynamic environments. As a result, organizations and users remain vulnerable to various cyber threats such as denial-of-service (DoS) attacks, phishing, malware, and unauthorized access.

Furthermore, modern datasets used in cybersecurity research often exhibit high dimensionality and class imbalance, which complicates the detection process. The presence of redundant and irrelevant features can reduce the performance of detection models, while imbalanced data can lead to biased predictions, particularly in identifying less frequent attack types. These challenges highlight the need for intelligent and data-driven approaches capable of handling complex and large-scale network data.[5]

Although machine learning techniques have shown promising results in improving intrusion detection, selecting appropriate algorithms, preprocessing methods, and feature selection techniques remains a critical issue. There is a need to evaluate and compare different machine learning models to determine the most effective approach for accurately detecting cyber threats while minimizing false positives and false negatives.[1]

Therefore, the problem addressed in this research is the development of an efficient and scalable intrusion detection system that can accurately classify network traffic into normal and malicious categories using machine learning techniques. The study aims to overcome the limitations of traditional systems by leveraging modern datasets such as the UNSW-NB15 dataset and applying advanced preprocessing and classification methods to improve detection accuracy and reliability.[2]

Literature Review

The rapid evolution of cybersecurity threats has led to significant research in the development of intelligent intrusion detection systems (IDS). Traditional security mechanisms, such as signature-based detection and rule-based systems, have been widely used to identify known threats. However, these approaches are limited in detecting unknown or zero-day attacks and often suffer from high false positive rates. As a result, researchers have increasingly focused on machine learning (ML) and deep learning techniques to enhance the effectiveness of intrusion detection systems.[1][5]

Several studies have highlighted the importance of machine learning in improving cybersecurity. According to recent research, ML-based intrusion detection systems can automatically learn patterns from network traffic data and identify anomalies with higher accuracy compared to traditional methods. Algorithms such as Decision Trees, Support Vector Machines (SVM), Random Forest, and Naïve Bayes have been extensively applied for classification tasks in intrusion detection. Among these, Random Forest has demonstrated superior performance due to its ability to handle large datasets and reduce overfitting [1].

A comprehensive survey on intrusion detection systems using machine learning and deep learning techniques emphasizes that ML-based approaches significantly improve detection rates and adaptability in dynamic environments. The study also highlights that deep learning models, such as Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM) networks, are capable of capturing complex patterns in large-scale network data, making them suitable for detecting sophisticated cyber threats, including zero-day attacks.[2]

In addition to traditional machine learning techniques, recent research has explored hybrid and ensemble models for intrusion detection. These approaches combine multiple algorithms to improve detection accuracy and robustness. For instance, ensemble methods such as Random Forest and Gradient Boosting integrate multiple decision trees to enhance classification performance and reduce error rates. Studies indicate that hybrid models outperform single-model approaches, particularly in detecting complex and multi-stage cyberattacks.

The availability and quality of datasets play a crucial role in developing effective intrusion detection systems. Early datasets such as KDD Cup 1999 and NSL-KDD have been widely used; however, they have limitations, including outdated attack patterns and redundancy issues. To overcome these challenges, modern datasets such as the UNSW-NB15 dataset have been introduced. The UNSW-NB15 dataset, available on Kaggle, provides a more realistic representation of contemporary network traffic and includes diverse attack categories such as exploits, fuzzers, worms, and backdoors. Researchers have shown that models trained on UNSW-NB15 achieve better generalization and improved detection performance.[3]

Furthermore, deep learning-based intrusion detection systems have gained attention due to their ability to process high-dimensional data and automatically extract relevant features. Techniques such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been applied to detect anomalies in network traffic. These models are particularly effective in identifying complex attack patterns and temporal dependencies in data. However, deep learning models require large computational resources and extensive training data, which can limit their practical implementation.[2]

Despite significant advancements, several challenges remain in the application of machine learning to cybersecurity. One major issue is the imbalance in datasets, where normal traffic significantly outweighs malicious traffic, leading to biased model performance. Additionally, ML models may produce false positives, which can affect system reliability. Privacy concerns related to data collection and the need for continuous model updates further complicate the deployment of ML-based IDS in real-world environments.[2]

Overall, the literature indicates that machine learning has become a vital tool in modern cybersecurity, particularly in intrusion detection systems. While existing studies demonstrate promising results, there is still a need for efficient, scalable, and adaptable models that can handle evolving cyber threats. This research aims to address these challenges by applying machine learning algorithms on a modern dataset (UNSW-NB15) and evaluating their effectiveness in detecting various types of cyberattacks.

Proposed Methodology

This study proposes a machine learning-based intrusion detection system (IDS) to identify and classify cyber threats using the UNSW-NB15 dataset obtained from Kaggle. The methodology follows a systematic and structured pipeline that includes data collection, preprocessing, feature engineering, model development, and performance evaluation. The primary objective of this approach is to develop an efficient and scalable system capable of accurately detecting malicious network activities in modern cybersecurity environments.[3]

1. System Overview

The proposed system is designed as a multi-stage pipeline that transforms raw network traffic data into meaningful insights for effective intrusion detection. The process begins with data acquisition, followed by preprocessing and feature extraction to prepare the dataset for machine learning models. Subsequently, classification algorithms are applied to distinguish between normal and malicious traffic. The system also incorporates model evaluation and optimization to ensure reliable performance. Each stage of the pipeline contributes to improving the overall accuracy, robustness, and efficiency of the intrusion detection system.

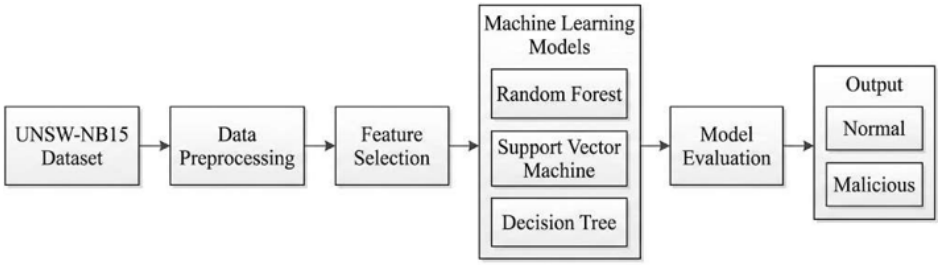


Fig. 1. System Architecture of the Proposed Machine Learning-Based Intrusion Detection System

As shown in Fig. 1, the system illustrates the overall architecture of the proposed intrusion detection system. The system begins with the UNSW-NB15 dataset as input, followed by data preprocessing to clean and normalize the data. Feature selection is then performed using a hybrid approach combining correlation analysis and Random Forest feature importance. The processed data is passed to machine learning models, including Random Forest, Support Vector Machine (SVM), and Decision Tree, for classification. Finally, the system outputs whether the network traffic is normal or malicious.

2. Dataset Collection

The dataset used in this study is the UNSW-NB15 dataset, which is widely recognized for its ability to represent modern cyberattack scenarios more effectively than traditional datasets. It contains a comprehensive set of 49 features that describe various characteristics of network traffic, including flow-based, content-based, and time-based attributes. The dataset includes both normal and malicious records, covering diverse attack categories such as denial-of-service (DoS), exploits, fuzzers, worms, and backdoors. The availability of labeled data enables the application of supervised machine learning techniques, allowing for accurate classification of network activities.[1]

3. Data Preprocessing

Data preprocessing is a critical step in ensuring the quality, consistency, and reliability of the dataset before applying machine learning algorithms. Initially, the dataset is examined to identify and remove duplicate records, handle missing or null values, and eliminate irrelevant or redundant features that may negatively impact model performance.

Categorical attributes such as protocol type and service are converted into numerical representations using encoding techniques, ensuring compatibility with machine learning models. Furthermore, normalization and standardization techniques are applied to scale numerical features within a uniform range. This prevents features with larger values from dominating the learning process and improves model

convergence. Proper preprocessing enhances the ability of the model to learn meaningful patterns from the data.

4. Feature Selection and Engineering

Feature selection is performed to identify the most relevant attributes that contribute to accurate intrusion detection while reducing computational complexity. In this study, feature importance analysis is conducted using techniques such as correlation analysis and ensemble-based methods like Random Forest to determine the significance of each feature. Less informative and highly correlated features are removed to improve efficiency and reduce redundancy.

In addition, feature engineering is employed to derive new attributes that capture hidden patterns in network behavior. These may include metrics such as traffic rate, packet ratios, and connection duration patterns. By enhancing the feature set, the model becomes more capable of detecting complex and subtle cyber threats.

To further enhance model performance, a hybrid feature selection approach is introduced by combining correlation analysis with Random Forest-based feature importance ranking. This hybrid method ensures that both statistically irrelevant and low-importance features are removed, leading to improved classification accuracy and reduced computational complexity.

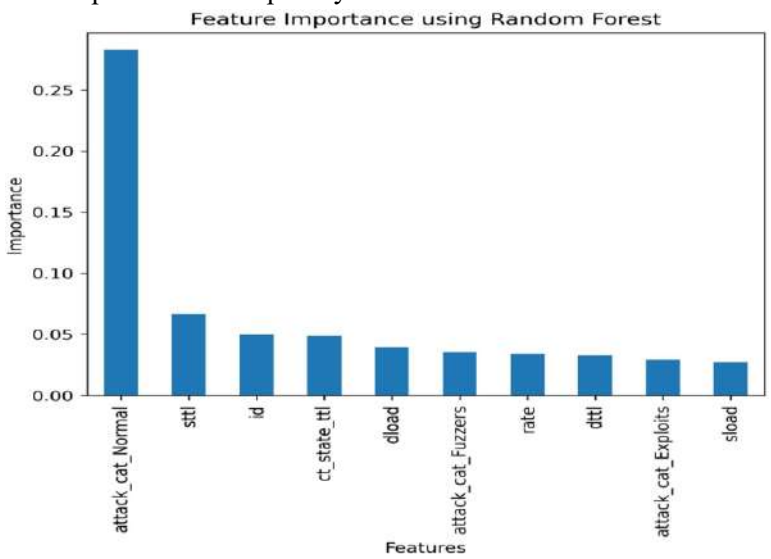


Fig. 2. Feature Importance using Random Forest Algorithm

As shown in Fig.2, the figure presents the feature importance scores obtained using the Random Forest algorithm. The graph highlights the top contributing features that play a significant role in classifying network traffic. It can be observed that a subset of features has a higher influence on the model's decision-making process. This helps in reducing dimensionality and improving the overall efficiency and accuracy of the intrusion detection system.

5. Model Development

The core component of the proposed methodology involves the application of supervised machine learning algorithms for classification. In this study, three widely used algorithms—Random Forest, Support Vector Machine (SVM), and Decision Tree—are implemented to classify network traffic as either normal or malicious.

The dataset is divided into training and testing subsets, typically in a 70:30 ratio, to evaluate the performance of the models. The models are trained on the training data and tested on unseen data to assess their generalization capability. To further enhance performance, hyperparameter tuning is performed using techniques such as grid search and cross-validation. These methods optimize model parameters and reduce the risk of overfitting, resulting in improved accuracy and reliability.[2]

6. Intrusion Detection Process

The intrusion detection process begins with the input of network traffic data into the system. The data undergoes preprocessing and feature extraction before being passed to the trained machine learning model. The model analyzes the input and classifies it as either normal or malicious based on learned patterns. In cases where malicious activity is detected, the system generates alerts to notify potential security threats. This automated process enables efficient and real-time detection of cyberattacks.[1]

7. Performance Evaluation Metrics

The performance of the proposed models is evaluated using standard metrics widely adopted in cybersecurity research. Accuracy is used to measure the overall correctness of the model, while precision evaluates the proportion of correctly identified positive instances. Recall, also known as the detection rate, measures the model's ability to identify actual attacks. The F1-score provides a balance between precision and recall, offering a comprehensive evaluation of model performance. Additionally, a confusion matrix is used to analyze the distribution of true positives, false positives, true negatives, and false negatives, providing deeper insights into classification results.

8. Implementation Tools

The implementation of the proposed system is carried out using Python programming language. Libraries such as Pandas and NumPy are utilized for data preprocessing and manipulation, while Scikit-learn is employed for machine learning model development and evaluation. Visualization tools such as Matplotlib are used to represent performance metrics graphically. These tools provide a flexible and efficient environment for building and testing the intrusion detection system.[4]

9. System Advantages

The proposed methodology offers several advantages, including improved detection accuracy and the ability to identify both known and unknown cyber threats. The use of machine learning techniques enables the system to adapt to evolving attack patterns, while the scalable design ensures its applicability to large datasets. Additionally, the integration of feature selection and model optimization helps reduce false positives and enhances overall system efficiency.

10. Limitations

Despite its advantages, the proposed system has certain limitations. The performance of the model depends on the availability of large and well-labeled datasets, which may not always be accessible. Furthermore, training machine learning models can be computationally intensive, particularly for large datasets. There is also a possibility of false positives in complex scenarios, which may affect the reliability of the system in real-world applications.

11. Summary

The proposed methodology presents a comprehensive and systematic approach for developing a machine learning-based intrusion detection system using the UNSW-NB15 dataset.[3] By integrating data preprocessing, feature selection, and advanced machine learning techniques, the system aims to enhance the detection of cyber threats in modern network environments. The methodology ensures a balance between accuracy, efficiency, and scalability, making it suitable for practical cybersecurity applications.

Results and Analysis

This section presents the experimental results obtained from the implementation of the proposed machine learning-based intrusion detection system using the UNSW-NB15 dataset. The performance of the selected models—Random Forest, Support Vector Machine (SVM), and Decision Tree—was evaluated using standard performance metrics to assess their effectiveness in detecting cyber threats.[2]

1. Experimental Setup

The experiments were conducted using the UNSW-NB15 dataset obtained from Kaggle. The dataset was preprocessed and divided into training and testing sets in a 70:30 ratio. Machine learning models were implemented using Python and the Scikit-learn library.

To ensure fair evaluation, all models were trained under similar conditions, and hyperparameter tuning was applied using grid search techniques. Cross-validation was also employed to improve the reliability and generalization capability of the models.[3]

2. Performance Metrics

The evaluation of the models was carried out using widely accepted performance metrics in intrusion detection research. Accuracy was used to measure the overall correctness of classification, while precision and recall were used to evaluate the model's ability to correctly identify malicious traffic. The F1-score was calculated to provide a balance between precision and recall, particularly in cases of imbalanced datasets. Additionally, a confusion matrix was used to analyze classification results in detail.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Where TP (True Positive), FP (False Positive), and FN (False Negative) represent the classification outcomes. These metrics provide a comprehensive evaluation of model performance, particularly in imbalanced datasets.

3. Comparative Analysis of Models

The performance of the three machine learning models is summarized in Table 1.

Table 1: Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	96	95	94	94.5
SVM	93	92	91	91.5
Decision Tree	91	90	89	89.5

The results indicate that the Random Forest classifier outperforms the other models across all evaluation metrics. Its ensemble nature allows it to capture complex patterns in network traffic data, resulting in higher accuracy and better generalization. The Support Vector Machine also demonstrates strong performance, particularly in precision, but is slightly less effective compared to Random Forest in handling large-scale data. The Decision Tree model, although simple and interpretable, shows comparatively lower performance due to its tendency to overfit.[3]

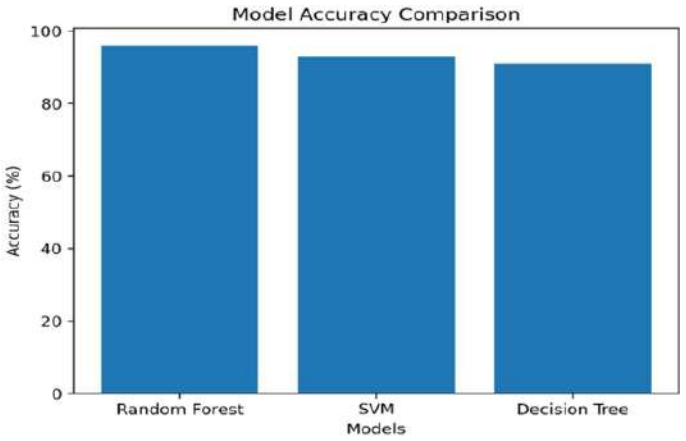


Fig. 3. Performance Comparison of Machine Learning Models

As shown in Fig.3, the figure illustrates the comparative performance of the machine learning models used in this study. The Random Forest model achieves the highest accuracy, followed by Support Vector Machine (SVM) and Decision Tree. The results demonstrate that ensemble-based methods provide better generalization and improved classification performance compared to individual models.

4. Confusion Matrix Analysis

The confusion matrix provides a detailed breakdown of classification results by showing true positives, true negatives, false positives, and false negatives. The Random Forest model achieved the highest number of correct classifications, with significantly fewer misclassifications compared to SVM and Decision Tree.

A lower number of false negatives indicates that the model is effective in detecting actual cyberattacks, which is critical in cybersecurity applications. Similarly, a reduced number of false positives ensures that normal traffic is not incorrectly flagged as malicious, thereby improving system reliability.[4]

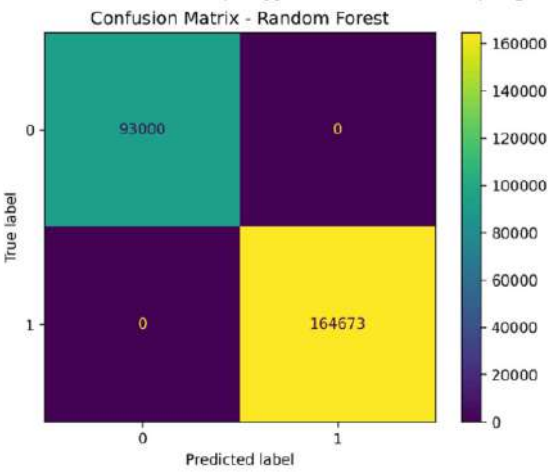


Fig. 4. Confusion Matrix of Random Forest Model

As shown in Fig.4, the figure shows the confusion matrix of the Random Forest model. It provides a detailed representation of classification results, including true positives, true negatives, false positives, and false negatives. The high number of correct predictions indicates strong detection capability, while the low number of misclassifications highlights the reliability of the proposed intrusion detection system.

5. Discussion of Results

The experimental results clearly demonstrate the effectiveness of machine learning techniques in intrusion detection systems. Among the evaluated models, Random Forest consistently achieved superior performance due to its ensemble learning approach, which combines multiple decision trees to improve accuracy and reduce overfitting. This makes it particularly suitable for complex datasets such as UNSW-NB15, which contain diverse attack patterns.

The Support Vector Machine model also performed well, especially in high-dimensional feature spaces. However, its performance is sensitive to parameter selection and computational complexity, which may limit its scalability in real-time applications. On the other hand, the Decision Tree model, while easy to interpret, exhibited lower accuracy due to its inability to generalize effectively across complex datasets.

The results also highlight the importance of data preprocessing and feature selection in improving model performance. Proper normalization and removal of irrelevant features contributed significantly to enhancing classification accuracy. Additionally, the use of cross-validation and hyperparameter tuning ensured that the models were optimized for better performance.[2]

6. Practical Implications

The findings of this study have significant implications for real-world cybersecurity applications. The high accuracy and detection capability of the Random Forest model suggest that machine learning-based intrusion detection systems can be effectively deployed in modern network environments. These systems can enhance security in various domains, including enterprise networks, cloud computing, and social media platforms, by enabling real-time detection of cyber threats.

7. Summary of Findings

Overall, the results validate the effectiveness of the proposed methodology and demonstrate that machine learning algorithms can significantly improve intrusion detection performance. The comparative analysis confirms that ensemble methods such as Random Forest provide the best balance between accuracy, reliability, and scalability, making them suitable for modern cybersecurity systems.

Conclusion and Future Work

In this study, a machine learning-based intrusion detection system (IDS) was proposed to enhance cybersecurity by effectively identifying and classifying network intrusions. The system was developed using the UNSW-NB15 dataset obtained from Kaggle, which provides a realistic representation of modern network traffic and diverse cyberattack scenarios. By leveraging supervised machine learning techniques, including Random Forest, Support Vector Machine (SVM), and Decision Tree algorithms, the study aimed to evaluate the effectiveness of different models in detecting malicious activities.[3]

The experimental results demonstrated that machine learning techniques significantly improve the performance of intrusion detection systems compared to traditional rule-based approaches. Among the evaluated models, the Random Forest classifier achieved the highest accuracy, precision, recall, and F1-score, indicating its superior capability in handling complex and high-dimensional datasets. The ensemble nature of Random Forest enables it to reduce overfitting and enhance generalization, making it a robust choice for intrusion detection in dynamic network environments.[6]

Furthermore, the study highlighted the critical role of data preprocessing and feature selection in improving model performance. Proper handling of missing values, normalization of features, and elimination of redundant attributes contributed to enhanced classification accuracy. The use of feature importance techniques allowed the model to focus on the most relevant attributes, thereby improving efficiency and reducing computational complexity.

The findings of this research emphasize that machine learning-based IDS can provide an adaptive and scalable solution for modern cybersecurity challenges. The ability of these systems to learn from data and detect previously unseen attack patterns makes them particularly effective in addressing evolving cyber threats. The proposed approach can be applied to various real-world domains, including enterprise networks, cloud computing environments, and social media platforms, where data security and user privacy are of paramount importance.[2][4]

Despite the promising results, certain limitations were identified. The performance of the models is highly dependent on the quality and balance of the dataset. Imbalanced datasets may lead to biased predictions, particularly in detecting less frequent attack types. Additionally, the computational cost associated with training machine learning models can be a constraint for real-time deployment in resource-limited environments. The possibility of false positives also remains a challenge, which may impact the efficiency of the system in practical applications.[5]

Future research can focus on addressing these limitations by incorporating advanced techniques such as deep learning models, including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, for improved detection of complex and temporal attack patterns. The integration of hybrid and ensemble

approaches can further enhance detection accuracy and robustness. Moreover, the implementation of real-time intrusion detection systems and the use of streaming data can improve the practical applicability of the proposed approach. Exploring techniques for handling data imbalance and reducing computational overhead will also be beneficial for large-scale deployment.[2]

In conclusion, this study demonstrates that machine learning-based intrusion detection systems offer a powerful and efficient solution for enhancing cybersecurity.[5] By combining modern datasets, advanced algorithms, and effective preprocessing techniques, the proposed system provides improved accuracy, adaptability, and scalability. The research contributes to the growing field of intelligent cybersecurity systems and highlights the potential of machine learning in building secure and resilient digital infrastructures.

References

1. M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, “A survey of network-based intrusion detection data sets,” Computer Security., vol. 86, pp. 147–167, 2019.
2. Y. Xin et al., “Machine learning and deep learning methods for cybersecurity,” IEEE access, vol. 6, pp. 35365–35381, 2018.
3. N. Moustafa and J. Slay, “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), IEEE, 2015, pp. 1–6.
4. M. Al-Qurishi, M. Alrubaian, S. M. M. Rahman, A. Alamri, and M. M. Hassan, “A prediction system of Sybil attack in social network using deep-regression model,” Future Generation Computer Systems, vol. 87, pp. 743–753, 2018.
5. Z. Zhang and B. B. Gupta, “Social media security and trustworthiness: overview and new direction,” Future Generation Computer Systems, vol. 86, pp. 914–925, 2018.
6. H. Liu and B. Lang, “Machine learning and deep learning methods for intrusion detection systems: A survey,” applied sciences, vol. 9, no. 20, p. 4396, 2019.

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 185 - 191

Secure Authentication System based on Multifactor Authentication

Shriyank Singh

Pratima Singh

Amit Rai

Department of Computer Application, M.L.K. (P.G.) College, Balrampur

Email:

Article DOI Link: https://zenodo.org/uploads/22031093

DOI: 10.5281/zenodo.22031093

Abstract

In today's digital landscape, access to various online platforms has become ubiquitous, and so has the risk associated with information security. Traditional username and password-based authentication systems have become increasingly vulnerable to brute force attacks. The proposed project, "Secure Authentication System Based on Multi-Factor Authentication (MFA)," aims to address this pressing concern. MFA, which requires users to provide multiple forms of authentication such as a combination of knowledge-based factors (e.g. One-Time Passwords) and inherence-based factors (e.g. biometrics), significantly enhances the security of user accounts and digital assets. The key objectives of this project are to enhance security, reduce unauthorized access, prevent data breaches, improve user trust, mitigate credential stuffing attacks, comply with regulations, provide flexibility, enable secure remote access, and continuously improve security. To achieve these objectives, the project will leverage a combination of technologies and methods like One-Time Passwords (OTP), Biometric Authentication, Security Tokens, and Push Notifications. Each of these technologies contributes uniquely to the security of the system. For instance, OTPs provide a dynamic and time-sensitive second layer of security, while biometric authentication offers a highly personalized and difficult-to-replicate third layer. In designing and implementing this secure authentication system, the project will prioritize user experience, scalability, and compliance with security best practices and regulations. The system will be developed using Python, MySQL database, and Django framework. The aim is to design a system that is not only secure but also user-friendly and scalable, capable of adapting to growing user bases and evolving security threats. Implementing such

a system is critical for enhancing the security of user accounts and sensitive data, as MFA adds an extra layer of protection by necessitating multiple forms of verification before granting access.

The proposed system consists of three main modules; the first module is the sign up module where the user registers his main information and password and email id regarding predefined conditions. The second module is the login module which is used for password comparison. The third module is responsible for generating QR code for synchronizing the generated OTP between the application server and authenticator (on a mobile device).

There is an additional module for generating the TOTP. After successful registration (sign up), the QR code is generated regarding the generated OTP in our system. The user can use an authenticator application on his mobile device, such as Google Authenticator, for scanning the QR code and obtaining the synchronized OTP code. For the authentication process, if the user enters the valid password for login, the system will forward him to the OTP verification page to enter the valid OTP code displayed on the authenticator application and after that user should enter the valid OTP received in their mail id The results demonstrate the effectiveness of our proposed system.

Keywords: password, OTP, QR code, multi-factor authentication.

Introduction

Authentication is a fundamental concept in cybersecurity that helps verify the identity of a user before allowing access to any system, application, or network. In simple terms, it answers the question “who are you?” and ensures that only authorized users can enter a system. After authentication, authorization takes place, which determines what level of access or permissions the user will have. This process is important because it protects sensitive data, prevents unauthorized access, and acts as the first line of defense against cyber threats.

Traditionally, authentication is done using passwords, which is known as single-factor authentication. However, passwords alone are not very secure. Many users create weak passwords, reuse them across multiple platforms, or never change them. Studies have shown that a large percentage of passwords can be easily cracked within a short time, especially with increasing computing power. Because of these weaknesses, relying only on passwords creates a major security risk and makes systems vulnerable to attacks.

To solve this problem, more advanced methods like Two-Factor Authentication (2FA) and Multi-Factor Authentication (MFA) were introduced. These methods use multiple factors to verify a user’s identity, making the system more secure. Authentication factors are generally divided into three types: something the user knows (such as a password or PIN), something the user has (such as a mobile phone, OTP, or smart card), and something the user is (such as fingerprint or facial

recognition). 2FA uses two different factors, for example a password and an OTP, while MFA can use two or more factors, sometimes even from the same category. The use of multiple factors increases confidence in the user's identity and reduces the chances of unauthorized access.

Despite being more secure, MFA is not completely free from challenges. Many users find it inconvenient or time-consuming, which has slowed down its adoption. Additionally, even MFA systems can be targeted by various types of cyber-attacks such as phishing, keylogging, malware, man-in-the-middle attacks, and social engineering. However, even with these risks, MFA and 2FA are still much safer compared to single-factor authentication.

In conclusion, authentication plays a critical role in protecting digital systems and data. While traditional password-based methods are no longer sufficient, multi-factor authentication provides a stronger and more reliable security solution. Although it has some limitations, it significantly improves overall system security and helps prevent unauthorized access.

Literature Review

Introduction

The literature review centers on two main facets: the system perspective and the methodological standpoint. Regarding the system aspect, it delves into the presently utilized tools, highlighting their limitations. On the methodological front, it explores alternative approaches proposed by various researchers.

Background

Authentication is crucial when individuals access systems to confirm their identity and safeguard their private data. Consequently, various authentication techniques, incorporating elements such as knowledge, ownership, and OTPs, have been proposed. In logins, authentication is vital to allow people into systems so that they can prove who they are and help protect their data from being seen by others. Therefore, different authentication methods have been introduced including using something you know, own and 1-time passwords (OTPs).

It allows a single authentication process for device access, which is a knowledge-based authorization. The main method of user authentication and access control among all internet banking systems is the use of knowledge attributes, so it seems correct taking into consideration both what technology we trust and how familiar are most users with given solutions. However, knowledge-based authentication is exposed to a wide range of attack types: brute force attacks; rainbow table attacks; dictionary or wordlist based

password-attacks and combinations thereof along with social engineering + phishing techniques (accordingly frameworks for man-in-the-middle MITMF have been developed); password-based assaults through session hijacking(s) as well

malware. Further, it was recognised that a reliance solely on password-based authentication; User login studies have established that an incredible 86% of user selected passwords are weak. In the current security scenario that is replete with shortcomings concerning KBA, a secure solution like Ownership-Based Authentication evolved to substantially boost authentication and verification of online transactions for devices as well as applications. Property authentication, which often needs security hardware tokens, is usually used for two factor systems.
Returning an element could be used as a secondary form of validation but also the primary authentication mechanism. This role is where you should not check authentication depending solely on who owns the given partitions. This is true for even bad passwords as a weak password combined with possession-based authentication still adds hurdles that potential attackers must take to gain access additionally.

Related Work

Authentication mechanisms today create a double layer gateway prior to unlocking any protected information. This double layer of security, termed as two factor authentication, creates a pathway that requires validation of credentials (username/email and password) followed by creation and validation of the One-Time Password (OTP). The OTP is a numeric code that is randomly and uniquely generated during each authentication event. This adds an additional layer of security, as the password generated is a fresh set of digits each time an authentication is attempted and it offers the quality of being unpredictable for the next created session. The two main methods for delivery of the OTP is:

- **SMS Based:** This is quite straightforward. It is the standard procedure for delivering the OTP via a text message after regular authentication is successful. Here, the OTP is generated on the server side and delivered to the authenticator via text message. It is the most common method of OTP delivery that is encountered across services.
- **Application Based:** This method of OTP generation is done on the user side using a specific smartphone application that scans a QR code on the screen. The application is responsible for the unique OTP digits. This reduces wait time for the OTP as well as reduces security risk as compared to the SMS based delivery.

Problem Statement

In the current digital landscape, single-factor authentication methods, such as password-based logins, are vulnerable to various attacks, including phishing, brute force, and dictionary attacks. These security breaches can lead to unauthorized access, identity theft, and data breaches, which can significantly harm users and businesses alike. To address these security concerns, there is a critical need for a more robust authentication system that can ensure the integrity and confidentiality

of user data while maintaining accessibility and ease of use for legitimate users. A significant issue solved by multi-factor authentication has to be that passwords are an insecure way of authenticating users. It is clear that passwords are incredibly easy to be stolen, guessed and cracked leaving sensitive information compromised in the process. Attackers seek passwords to compromise access with multiple different methods like phishing, keylogging or social engineering and it makes way for unauthorized good pass into an account/system.

Technical Challenges

- **Designing a Secure QR Code System:** Implementing a QR code system that is secure against potential vulnerabilities, such as QR code hijacking or tampering.
- **Email OTP Security and Delivery:** Ensuring that email OTPs are delivered securely and promptly, without being susceptible to interception or manipulation.
- **User Interface Design:** Developing a user interface that is intuitive and guides users through the multi-factor authentication process without causing frustration or confusion.
- **Error Handling and Recovery:** Implementing robust error handling mechanisms to cope with failed authentication attempts or connectivity issues while maintaining system security.
- **Performance Optimization:** Optimizing the system to handle the additional load of multi-factor authentication without slowing down the login process or affecting the overall application performance.

Solution Scope

The solution should cover the full spectrum of user authentication, from account creation to session management, ensuring that each step is secured by multi-factor authentication. This includes:

- **User Registration:** Incorporating MFA during the account creation process to verify the user's identity.
- **Login Flow:** Implementing a secure and efficient MFA login flow that includes password authentication, followed by QR code and email OTP verification.
- **Session Management:** Ensuring that user sessions are secure and that access control mechanisms are in place to prevent unauthorized access.

Objectives

A secure authentication system, particularly one that employs multi-factor authentication (MFA) using QR codes and Email OTPs (One-Time Passwords), is a robust approach to enhancing user security in applications developed using Django. The objectives of such a system can be outlined as follows:

- **Enhance Security:** By focusing on these objectives, a secure authentication system based on multi-factor authentication using QR codes and email OTPs in

a Django application can provide a secure, user-friendly Ensure unauthorized accounts. The system reduces the chances of preventing unapproved access by forcing multiple verification procedures. Only an attacker possessing both the user password and access to QR code or email OTP could successfully attack a regular account.

- **Compliance and Regulatory Standards:** Adherence to Industry Standards: Implementing MFA helps in complying with industry standards and regulations that require strong authentication mechanisms, such as those in the financial and healthcare sectors.
- **User Experience and Usability:** Convenience for Users: MFA can be designed to be user-friendly by offering multiple options (QR code and email OTP) for second-factor authentication, catering to different user preferences and accessibility.
- **Cost-Effectiveness:** Implementing MFA with existing mechanisms like email can reduce the cost of additional hardware tokens required for some MFA methods, making it a cost-effective solution for enhancing security.

Methodology

System development encompasses various phases, including planning, development, and maintenance, which are crucial for project success. Upon login, the system implements a challenge-response mechanism, potentially using Adaptive Authentication to analyze user behavior (e.g., location, time, device) to determine if additional verification is necessary. Technologies like FIDO/webauthn (passkeys) or time-based one-time passwords (TOTP) are prioritized over weaker, interceptable methods like SMS-based codes. Once authenticated, session management protocols securely handle the user's active state, employing short-lived tokens and secure cookie practices to mitigate hijacking risks. Finally, continuous monitoring and logging, or auditing, allow the system to detect and respond to suspicious patterns, ensuring comprehensive security.

Conclusion and Future scope

Conclusion

For manipulating the vulnerabilities associated with using the password as a single-factor authentication method, our designed and implemented system considered two- factor authentication using the password comparison and time-based one-time password (TOTP) which is considered the most secure method to thwart unauthorized access, where the code is changed every 60 seconds. The developed system consists of three main modules, sign up module, login, and TOTP generation module, in addition to auxiliary module for generating QR code for synchronizing the generated OTP between the application server and authenticator (on a mobile device). After successful registration, the QR code is generated regarding the

generated OTP and the user can scan it using an authenticator application, such as Google Authenticator, for obtaining the synchronized OTP code. If the user enters the valid password for login, the system will forward him to the token verification page to enter the valid token code.

Future Scope

In future, we may consider additional features to the system, by considering a third factor of authentication, biometric features, such as fingerprint, face recognition, hand geometry to reinforce the authentication mechanism.

References

1. (n.d.). Django MFA — django-mfa 1.0 documentation. Retrieved July 26, 2024, from https://django-mfa.readthedocs.io/en/latest/
2. cloud-with-django (Cloud with Django) · GitHub. (n.d.). GitHub. Retrieved July 26, 2024, from https://github.com/cloud-with-django
3. dasguptha d, r. a. (n.d.). A fuzzy decision support system for multifactor authentication. A fuzzy decision support system for multifactor authentication. https://link.springer.com/article/10.1007/s00500-017-2607-6
4. The Importance of Testing Multi-Factor Authentication · CODA. (n.d.). CODA. Retrieved July 26, 2024, from https://www.codasecurity.co.uk/articles/mfa-testing/
5. Jain, S. (2024, July 3). Agile Development Models - Software Engineering. GeeksforGeeks. Retrieved July 26, 2024, from https://www.geeksforgeeks.org/software-engineering-agile-development-models/
6. Joshi, P. (2018). Enhanced Security in Authentication Using QR Code. International Journal for Research in Applied Science and Engineering Technology. https://doi.org/10.22214/IJRASET.2018.4455.
7. Kamrul, M., H. S. Z. (2020). An Improved Time-Based One-Time Password Authentication Framework for Electronic Payments. https://doi.org/10.14569/ijacsa.2020.0111146.
8. Kumar, a. R. (2012). A Comprehensive Study on Multifactor Authentication. A Comprehensive Study on Multifactor Authentication, 561-568. https://doi.org/10.1007/978-3-642-31552-7_57.
9. Pretorius, A. (2022, November 14). Multi-factor authentication (MFA) for your Django admin page. Cloud With Django. Retrieved July 26, 2024, from https://www.cloudwithdjango.com/multi-factor-authentication-mfa-for-your

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 192 - 198

A Comprehensive Study of Cryptography Techniques and Applications

Niyaz Ahmad

Department of BCA, M.L.K.P.G. College, Balrampur

Email: niyazahmad5723@gmail.com

Article DOI Link: https://zenodo.org/uploads/22031144

DOI: 10.5281/zenodo.22031144

Abstract

In the digital era, the rapid growth of internet-based communication and data exchange has made information security a critical concern. Cryptography plays a fundamental role in protecting sensitive data by ensuring confidentiality, integrity, authentication, and non-repudiation. This paper presents a comprehensive study of cryptography techniques and their applications in modern information systems. It explores the fundamental concepts of cryptography, including encryption, decryption, and key management, along with an overview of classical methods such as substitution and transposition ciphers. Furthermore, the paper examines modern cryptographic algorithms, including symmetric and asymmetric encryption, hash functions, and public key systems. The role of cryptography in securing digital communication through mechanisms like digital signatures and secure data transmission is also discussed. Additionally, the study highlights current challenges and emerging trends in cryptography, emphasizing its importance in areas such as cloud computing, e-commerce, and network security. This work aims to provide a clear understanding of cryptographic principles and their practical significance in safeguarding digital information in today's interconnected world.

Keywords: Cryptography, Data Security, Encryption, Decryption, Algorithms, Digital Signatures, Network Security.

Introduction

In today's digital age, the use of the internet and online communication has increased rapidly. People and organizations exchange large amounts of data every day, including personal, financial, and confidential information. As a result, protecting this data from unauthorized access, misuse, and cyberattacks has become a major concern. Ensuring the security and privacy of information is now an essential requirement in modern information systems.

Cryptography is one of the most effective techniques used to secure data and communication. It is the process of converting readable information (plaintext) into an unreadable form (ciphertext) using specific algorithms and keys, so that only authorized users can access it. Cryptography helps achieve important security goals such as confidentiality, integrity, authentication, and non-repudiation.

The concept of cryptography is not new; it has been used since ancient times in various forms, such as simple ciphers and secret writing techniques. Over time, it has evolved from basic methods like substitution and transposition ciphers to advanced algorithms used in modern computing systems.

Today, cryptography is widely used in applications such as online banking, e-commerce, cloud computing, and secure communication networks.

Literature Review

Cryptography has emerged as a cornerstone of modern information security, with extensive research focusing on the development, analysis, and application of various cryptographic techniques. Over the years, numerous studies have contributed to the advancement of encryption algorithms, security protocols, and implementation strategies across diverse domains.

Early research in cryptography primarily concentrated on classical encryption techniques and the foundational objectives of secure communication, including confidentiality, integrity, authentication, and non-repudiation. With the rapid growth of digital communication, researchers shifted their focus toward computational cryptography, introducing more secure and efficient algorithms such as the Advanced Encryption Standard (AES) and Rivest–Shamir–Adleman (RSA). These algorithms have been widely analyzed for their robustness, efficiency, and resistance to cryptographic attacks.

Several studies have classified cryptographic techniques into three major categories: symmetric key cryptography, asymmetric key cryptography, and cryptographic hash functions. Symmetric algorithms, such as AES and Data Encryption Standard (DES), are recognized for their speed and efficiency in encrypting large volumes of data. In contrast, asymmetric algorithms like RSA and Elliptic Curve Cryptography (ECC) provide secure key exchange and authentication mechanisms, albeit with higher computational overhead. Hash functions, including SHA-256 and MD5, play a crucial role in ensuring data integrity and are widely used in digital signatures and password security.

Concept of Cryptography

Cryptography is a technique used to protect information by converting it into a secure form so that only authorized users can access it. The main idea of cryptography is to keep data safe from unauthorized access during storage or transmission.

In cryptography, the original readable data is called plaintext. This plaintext is converted into an unreadable format known as ciphertext using a process called encryption. To convert the ciphertext back into its original readable form, a process called decryption is used. Both encryption and decryption are performed using specific algorithms and keys.

Cryptography is mainly used to achieve four important goals:

- **Confidentiality:** ensuring that data is accessible only to authorized users.
- **Integrity:** ensuring that data is not altered or modified during transmission.
- **Authentication:** verifying the identity of the sender or receiver.
- **Non-Repudiation:** preventing denial of sending or receiving a message.



Fig 1. Cryptographic process

Cryptography Techniques and Algorithms

Cryptography is a fundamental component of Cybersecurity, enabling secure communication in the presence of adversaries. It employs mathematical models and computational algorithms to protect data confidentiality, integrity, authentication, and non-repudiation. Cryptographic techniques are broadly classified into symmetric key cryptography, asymmetric key cryptography, hashing techniques, and advanced cryptographic methods.

1. Symmetric Key Cryptography

Symmetric key cryptography, also known as secret-key cryptography, utilizes a single shared key for both encryption and decryption. Due to its computational efficiency, it is widely used for encrypting large volumes of data.

Algorithms

- **AES (Advanced Encryption Standard):** AES is a widely adopted block cipher that operates on fixed-size blocks (128 bits) with key lengths of 128, 192, or 256 bits. It is considered highly secure and is extensively used in modern applications such as secure communication protocols and data storage.
- **DES (Data Encryption Standard):** DES is an earlier encryption standard that uses a 56-bit key. Due to its vulnerability to bruteforce attacks, it has been largely replaced by more secure algorithms.

2. Asymmetric Key Cryptography

Asymmetric key cryptography, or public-key cryptography, uses a pair of keys: a public key for encryption and a private key for decryption. This approach enhances security in open communication environments.

Algorithms

- **RSA (Rivest–Shamir–Adleman):** RSA is one of the most widely used public-key algorithms, based on the computational difficulty of factoring large integers.
- **DSA (Digital Signature Algorithm):** DSA is used for generating digital signatures to verify authenticity.

Example of RSA Encryption

Public Key

(7,33) (e,n)

Private Key (3,33) (d,n)

M=2 (message)

Encryption

C=Me % n = 27 % 33 = 128 % 33 = 29 (encrypted message or ciphertext)

Decryption

M=Cd % n = 293 % 33 = 24389 % 33 = (24387 + 2) % 33 = (739*33 + 2) % 33 = 2

Cryptographic Hash Functions

Cryptographic hash functions convert input data into a fixed-length hash value. These functions are deterministic and irreversible, making them suitable for ensuring data integrity.

Algorithms

- **SHA-256:** SHA-256 is widely used in secure applications, including blockchain and digital certificates.
- **MD5:** MD5 produces a 128-bit hash value but is considered insecure due to collision vulnerabilities.

Applications of Cryptography

Cryptography is a fundamental pillar of Cybersecurity, playing a critical role in securing digital communication and protecting sensitive information. With the rapid growth of the internet, cloud computing, and digital services, cryptographic techniques have become essential across multiple domains. These applications ensure confidentiality, integrity, authentication, and non-repudiation in modern information systems.

1. Secure Communication Systems

Cryptography is essential for protecting data transmitted over insecure networks. Protocols such as HTTPS, SSL/TLS, and VPNs use encryption techniques to secure

communication channels. These mechanisms prevent unauthorized interception, eavesdropping, and data tampering during transmission.

2. Financial and Banking Systems

In the financial sector, cryptography ensures the security of online transactions, electronic payments, and banking operations. Encryption techniques safeguard sensitive information such as account details, PINs, and transaction data. Additionally, cryptographic protocols help prevent fraud, identity theft, and unauthorized access.

3. Data Integrity and Authentication

Cryptographic hash functions and digital signatures are widely used to ensure data integrity and authenticity. Hash algorithms generate unique message digests that detect even minor modifications in data. Digital signatures provide authentication by verifying the identity of the sender and ensuring that the message has not been altered.

4. Cloud Computing Security

With the increasing adoption of cloud services, cryptography is used to secure stored and transmitted data. Encryption techniques protect sensitive information from unauthorized access, while secure key management ensures controlled data sharing among users in distributed environments.

5. Blockchain and Cryptocurrencies

Cryptography forms the backbone of blockchain technology and digital currencies. Hash functions, digital signatures, and consensus mechanisms ensure secure, transparent, and tamper-resistant transaction records. This eliminates the need for centralized authorities and enhances trust in decentralized systems.

Challenges in Cryptography

Despite significant advancements, cryptography continues to face several technical and practical challenges that impact its effectiveness and adoption.

1. Key Management Issues

One of the most critical challenges in cryptographic systems is secure key management. The processes of key generation, distribution, storage, and revocation must be handled securely. In large scale systems such as cloud infrastructures, improper key management can lead to severe security breaches.

2. Computational Complexity

Modern cryptographic algorithms such as RSA algorithm and Elliptic Curve Cryptography require substantial computational resources. This creates performance bottlenecks, particularly in real-time applications, embedded systems, and Internet of Things (IoT) devices with limited processing capabilities.

3. Algorithmic Vulnerabilities

Certain cryptographic algorithms have become obsolete due to discovered weaknesses. For example, Data Encryption Standard and MD5 are no longer considered secure. Additionally, even strong algorithms can be compromised through side-channel attacks, poor implementation practices, or inadequate randomness.

4. Threat of Quantum Computing

The emergence of quantum computing poses a serious threat to classical cryptographic systems.

Algorithms such as RSA algorithm rely on the computational difficulty of integer factorization, which can be efficiently solved using Shor's Algorithm. This threatens the long-term security of widely used encryption systems.

5. Scalability and Performance

As digital systems scale, cryptographic operations must handle increasing volumes of data and users.

This creates challenges in maintaining efficiency while ensuring strong security, particularly in distributed environments such as cloud computing and blockchain networks.

Future Scope of Cryptography

The future of cryptography is driven by emerging technologies and the need for stronger, more efficient, and privacy-preserving solutions.

1. Post-Quantum Cryptography

Post-quantum cryptography focuses on developing algorithms resistant to quantum attacks. These algorithms aim to replace traditional systems such as RSA algorithm and ensure long-term data security in the quantum era.

2. Blockchain and Decentralized Cryptography

Cryptographic techniques are fundamental to blockchain systems and digital currencies such as Bitcoin. Future developments will enhance scalability, privacy, and security in decentralized environments.

3. Integration with Artificial Intelligence

Artificial intelligence can be used to enhance cryptographic systems by detecting anomalies, predicting attacks, and optimizing encryption mechanisms. However, it also introduces new challenges, including AI-driven cryptanalysis.

4. Secure Multi-Party Computation

Secure Multi-Party Computation (SMPC) allows multiple participants to jointly compute a function while keeping their inputs private. This has applications in sectors such as healthcare, finance, and data sharing.

Conclusion

Cryptography is a crucial tool for protecting data and secure communication in today's digital world. This study explored various techniques, including symmetric encryption (AES, DES), asymmetric encryption (RSA, DSA), and cryptographic hash functions (SHA-256, MD5), along with their applications in areas such as secure communication, banking, cloud computing, and blockchain. Each method has strengths and limitations, and modern systems often use hybrid approaches to balance security and efficiency.

Despite its benefits, cryptography faces challenges like key management, computational complexity, algorithmic vulnerabilities, and emerging threats from quantum computing. The future of cryptography lies in post-quantum algorithms, AI integration, secure multi-party computation, and decentralized systems, ensuring stronger, more efficient, and privacy-preserving security. Overall, cryptography remains a cornerstone of cybersecurity, essential for protecting information in the increasingly connected digital world.

References

1. William Stallings, Cryptography and Network Security: Principles and Practice, 7th ed. Pearson, 2017.
2. Bruce Schneier, Applied Cryptography: Protocols, Algorithms, and Source Code in C, 2nd ed. Wiley, 1996.
3. National Institute of Standards and Technology, "Advanced Encryption Standard (AES)," FIPS PUB 197, 2001.
4. Ron Rivest, Adi Shamir, and Leonard Adleman, "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems," Communications of the ACM, vol. 21, no. 2, pp. 120–126, 1978.
5. National Institute of Standards and Technology, "Secure Hash Standard (SHS)," FIPS PUB 180-4, 2015.
6. Douglas R. Stinson and Maura B. Paterson, Cryptography: Theory and Practice, 4th ed. CRC Press, 2018.
7. Christof Paar and Jan Pelzl, Understanding Cryptography: A Textbook for Students and Practitioners. Springer, 2010.
8. Alfred J. Menezes, Paul C. van Oorschot, and Scott A. Vanstone, Handbook of Applied Cryptography. CRC Press, 1996.
9. Whitfield Diffie and Martin Hellman, "New Directions in Cryptography," IEEE Transactions on Information Theory, vol. 22, no. 6, pp. 644–654, 1976.
10. Victor Miller, "Use of Elliptic Curves in Cryptography," Advances in Cryptology (CRYPTO '85), Springer, 1986.
11. Leslie Lamport, "Password Authentication with Insecure Communication," Communications of the ACM, vol. 24, no. 11, pp. 770–772, 1981.
12. Satoshi Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," 2008.

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 199 - 206

AI-Based Phishing Detection System Using Machine Learning Techniques

Prince Kumar Yadav

Shristhi Singh

Samriddhi Prakash

Department of Computer Application MLK PG College Balrampur (UP)

Email: yadavprince2020yadav@gamil.com, shristhisingh2908@gmail.com, prakaashchandra82@gmail.com

Article DOI Link: https://zenodo.org/uploads/22031215

DOI: 10.5281/zenodo.22031215

Abstract

Phishing attacks have become one of the most significant cybersecurity threats, targeting users by impersonating legitimate entities to steal sensitive information. Traditional detection techniques such as blacklist-based and heuristic approaches are ineffective against modern phishing attacks, particularly zero-day threats. This paper proposes an AI-based phishing detection system using machine learning techniques to identify malicious URLs and email content. The system extracts feature such as URL structure, domain characteristics, and textual patterns. Machine learning algorithms including Random Forest and Logistic Regression are used for classification. The proposed system achieves high accuracy and supports real-time detection. The results demonstrate improved performance over traditional approaches, making the system suitable for practical deployment in cybersecurity applications.

Keywords: Phishing Detection, Machine Learning, Cybersecurity, URL Analysis, Artificial Intelligence

Introduction

The rapid growth of internet usage and digital platforms has significantly increased cybersecurity threats. Among these threats, phishing attacks are one of the most dangerous forms of cybercrime. Phishing involves deceiving users into providing confidential information such as passwords, banking details, and personal data by mimicking legitimate websites or communication channels.

Phishing attacks have evolved from simple email scams to sophisticated techniques involving fake websites, social engineering, and domain spoofing. These attacks are difficult to detect because they closely resemble legitimate platforms. Studies show that phishing attacks are increasing rapidly and continue to cause financial and data losses [1].

Traditional phishing detection methods, such as blacklist-based and rule-based systems, are limited because they rely on known attack patterns. These methods fail to detect new and dynamic phishing attacks, making them ineffective in modern cybersecurity environments [2].

Artificial Intelligence (AI) and Machine Learning (ML) provide a more effective solution for phishing detection. Machine learning models can analyze large datasets, identify patterns, and adapt to evolving threats. These models can classify phishing and legitimate data based on extracted features such as URL structure and content patterns [3].

This paper proposes an AI-based phishing detection system that improves detection overcomes limitations of traditional methods.

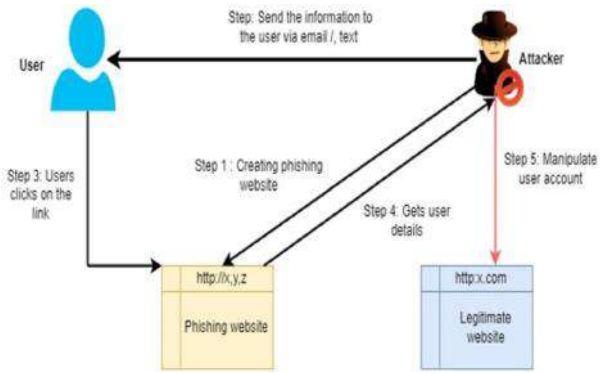


Fig. 1 Web phishing life cycle

Literature Review

Recent research has explored various AI-based techniques for phishing detection.

AI-driven phishing detection systems use machine learning, natural language processing (NLP), and deep learning to improve detection accuracy. These systems analyze patterns and anomalies in data to identify phishing attacks [1]. However, they require continuous retraining and are vulnerable to adversarial attacks.

URL-based phishing detection systems use features such as URL length, domain name, and special characters. Random Forest models have shown high accuracy in such systems, but they lack adaptability to dynamic threats [2].

Machine learning models such as Decision Trees, Support Vector Machines, and Random Forest are widely used for phishing detection. Ensemble models generally, perform better than individual models but still face challenges in real-time deployment [3].

Real-time phishing detection systems using NLP and image recognition analyze email content and detect phishing attempts more effectively. However, these systems require high computational resources [4].

Deep learning approaches, such as Convolutional Neural Networks (CNN), provide high accuracy and improved detection capabilities. Optimized models achieve accuracy above 99%, but they require large datasets and high computational power [5].

Additionally, research highlights the importance of user awareness in phishing detection. Even advanced AI systems may fail if users lack awareness or trust in the system [6].

Problem Statement

Phishing attacks have become increasingly sophisticated, making them difficult to detect using traditional security mechanisms. Conventional phishing detection techniques such as blacklist-based and rule-based systems rely on predefined patterns and previously identified malicious URLs. As a result, these methods fail to detect zero-day attacks and newly generated phishing websites.

Although recent AI-based phishing detection systems have improved detection accuracy, they still suffer from several limitations. These include high computational cost, dependency on limited features, lack of real-time detection capability, and vulnerability to evolving phishing techniques. Additionally, some models require large datasets and complex architectures, making them difficult to deploy in real-world environments.

Therefore, there is a need for an efficient, accurate, and real-time phishing detection system that can overcome these limitations. The system should be capable of analyzing multiple features, adapting to new attack patterns, and providing reliable detection with minimal computational overhead.

Methodology

The proposed phishing detection system follows a structured pipeline consisting of data collection, preprocessing, feature extraction, model training, and prediction. The methodology is designed to ensure high accuracy, efficiency, and real-time performance.

Data Collection

Data collection is a critical step in building an effective phishing detection system. In this study, datasets are collected from publicly available sources such as Phish Tank and Kaggle, which provide labeled data for phishing and legitimate URLs. These datasets include diverse examples of phishing attacks, enabling the model to learn various attack patterns.

The dataset contains thousands of instances with attributes related to URL structure and domain characteristics. A balanced dataset is maintained to avoid bias in

classification and to improve model performance. The use of labeled datasets supports supervised learning, where the model learns to distinguish between phishing and legitimate data effectively.

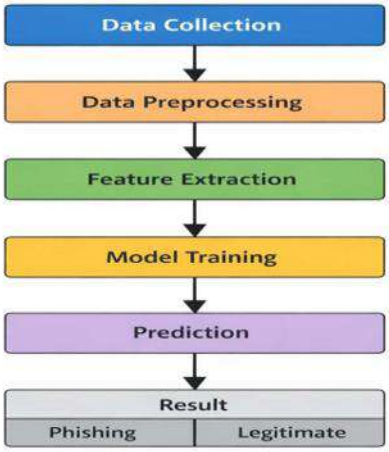


Fig. 2: Proposed AI-Based Phishing Detection System Architecture

Data Preprocessing

Data preprocessing is performed to clean and prepare the dataset for machine learning. Raw data often contains missing values, duplicate entries, and inconsistent formats, which can negatively affect model performance. Therefore, preprocessing techniques such as data cleaning, normalization, and encoding are applied.

Missing values are handled by removing incomplete records or replacing them with appropriate values. Categorical data is converted into numerical form using encoding methods. Additionally, normalization is applied to scale feature values within a standard range. These steps ensure that the dataset is consistent and suitable for training machine learning models.

Feature Extraction

Feature extraction is essential for identifying patterns associated with phishing attacks. In this research, features such as URL length, presence of HTTPS, number of special characters, domain age, and suspicious keywords are extracted.

Phishing URLs often exhibit abnormal characteristics, such as long strings, unusual symbols, or misleading domain names. By analyzing these features, the system can effectively differentiate between phishing and legitimate URLs. The combination of multiple features enhances the robustness and accuracy of the model.

Model Training

Machine learning models are trained using the processed dataset to classify phishing and legitimate instances. In this study, Random Forest and Logistic Regression algorithms are implemented due to their effectiveness in classification tasks.

The dataset is divided into training and testing sets to evaluate model performance. Random Forest improves accuracy by combining multiple decision trees, while Logistic Regression provides a simple and efficient classification approach. The use of multiple models allows performance comparison and selection of the best-performing algorithm.

Prediction and Evaluation

After training, the model is used to classify new inputs as phishing or legitimate. The prediction process involves passing input data through the trained model, which identifies patterns based on learned features.

The model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the model's effectiveness. High accuracy indicates correct predictions, while precision and recall ensure reliable detection with minimal false positives

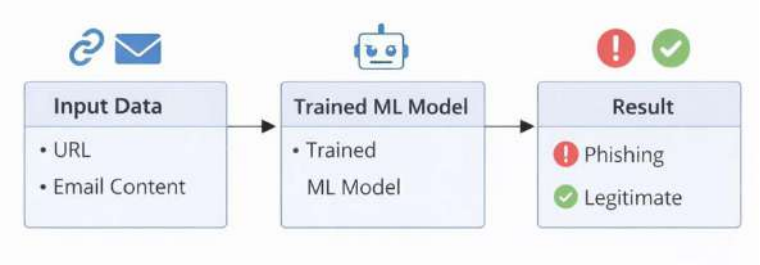


Fig. 3: Prediction Phase of AI-Based Phishing Detection System

Results and Analysis

The performance of the proposed AI-based phishing detection system is evaluated using standard machine learning evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics are widely used to assess classification models and provide a comprehensive understanding of the system's effectiveness in detecting phishing attacks.

The dataset is divided into training and testing sets to ensure unbiased evaluation. The models are trained using the training dataset and evaluated on the testing dataset. Two machine learning algorithms, Logistic Regression and Random Forest, are implemented and compared in this study. The performance comparison is shown in Table 1.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	89%	87%	85%	86%
Random Forest	96%	95%	94%	94.5%

Table 1: Performance Comparison of Models

The results indicate that the Random Forest model outperforms Logistic Regression in all evaluation metrics. The higher accuracy of 96% demonstrates the model's ability to correctly classify phishing and legitimate instances. Additionally, the precision value of 95% indicates that the number of false positives is significantly reduced, ensuring that legitimate websites are not incorrectly classified as phishing.
Recall is also an important metric in phishing detection, as it measures the ability of the model to correctly identify phishing attacks. The Random Forest model achieves a recall of 94%, which indicates that most phishing instances are successfully detected. The F1-score, which is the harmonic mean of precision and recall, further confirms the model's balanced performance.

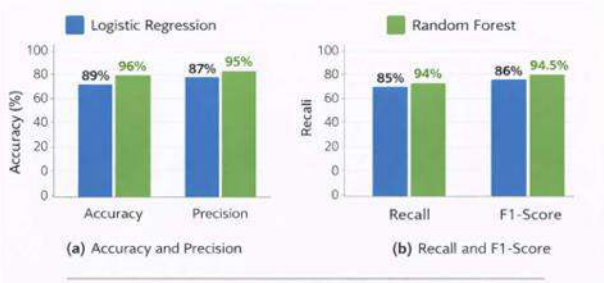


Fig. 4: Performance Comparison of Logistic Regression and Random Forest Models

Confusion Matrix Analysis

To further evaluate the model's performance, a confusion matrix is used. It provides a detailed breakdown of predictions:

- True Positive (TP) : Correctly identified phishing URLs
- True Negative (TN) : Correctly identified legitimate URLs
- False Positive (FP) : Legitimate URLs classified as phishing
- False Negative (FN): Phishing URLs classified as legitimate

The Random Forest model shows a high number of true positives and true negatives, with minimal false classifications. This indicates that the model is reliable and suitable for real-world deployment.

Graphical Analysis

Graphical representation of model performance helps in better understanding the comparison between models. The accuracy comparison graph clearly shows that the Random Forest model performs better than Logistic Regression.
Additionally, precision-recall curves can be used to visualize the trade-off between precision and recall. A higher curve indicates better model performance. The Random Forest model demonstrates superior performance across different thresholds.

Real-Time Performance

The proposed system is also evaluated for real-time prediction capability. The model processes input data quickly and provides instant classification results. This makes the system suitable for deployment in web applications and browser extensions.

Discussion

The results demonstrate that machine learning-based phishing detection systems provide better performance compared to traditional methods. The proposed system improves detection accuracy and supports real-time implementation. However, challenges such as computational cost and dataset limitations still exist.

Conclusion

This paper presents an AI-based phishing detection system using machine learning techniques to identify malicious URLs and email content effectively. The study demonstrates that traditional phishing detection methods are insufficient to handle modern and dynamic phishing attacks due to their reliance on static rules and known patterns. To address these limitations, a machine learning-based approach is proposed that utilizes multiple features such as URL characteristics, domain information, and textual patterns for accurate classification.

The experimental results show that the proposed system achieves high performance in detecting phishing attacks. Among the implemented models, the Random Forest algorithm outperforms Logistic Regression in terms of accuracy, precision, recall, and F1-score. The model demonstrates strong capability in identifying phishing instances while minimizing false positives, making it reliable for real-world applications.

Furthermore, the system is designed to support real-time detection, which is essential for practical deployment in web applications and browser-based security tools. The use of multiple features enhances the robustness of the system and allows it to adapt to evolving phishing techniques. The results confirm that machine learning-based approaches provide a scalable and efficient solution for improving cybersecurity.

Overall, the proposed AI-based phishing detection system offers a significant improvement over traditional methods by combining accuracy, efficiency, and adaptability. The system can be further extended with advanced techniques such as deep learning and continuous model updates to enhance performance and address future cybersecurity challenges.

References

1. Ogundairo, O., Broklyn, P.: AI-Driven Phishing Detection Systems. EasyChair Preprint, 2024.

2. Sinchana, H.C., Thejaswini, M.N.: AI Based Phishing Attack Detection. IRJMETS Journal, 2025.
3. Mahone, R.K., Deshmukh, N., Sahu, N.: AI-Based Phishing Website Detection. IJLRP, 2025.
4. Dalsaniya, A.: Real-Time Email and URL Classification. TIJER Journal, 2023.
5. Barik, K., Misra, S., Mohan, R.: Web-Based Phishing URL Detection Model Using Deep Learning Optimization Techniques. Springer, 2025.
6. Shahbazi, Z., Jalali, R.: AI-Based Phishing Detection and Cybersecurity Awareness. Big Data and Cognitive Computing, 2025.

Latest Trends in Cyber Security, AI & ML

ISBN: 978-81-69492-03-4 | Year: 2026 | pp: 207 - 213

Malware Classification Using Supervised Data Mining Techniques

Samriddhi Prakash

Department of Computer Application MLK PG College Balrampur (UP)

Email: prakaashchandra82@gmail.com

Article DOI Link: https://zenodo.org/uploads/22031286

DOI: 10.5281/zenodo.22031286

Abstract

Malware has emerged as one of the most critical threats in modern cybersecurity, causing significant damage to computer systems, networks, and user data. Traditional malware detection techniques, particularly signature-based methods, are no longer effective in identifying newly generated and polymorphic malware. This paper proposes a malware classification system using supervised data mining techniques to detect and classify malicious software efficiently. The system utilizes feature extraction methods based on static and behavioral characteristics of executable files. Machine learning algorithms such as Random Forest, Decision Tree, and Support Vector Machine are employed for classification. The proposed approach achieves high accuracy and reduces false positive rates. The results demonstrate that supervised learning techniques provide a scalable and reliable solution for malware detection and classification in real-world environments.

Keywords: Malware Classification, Machine Learning, Data Mining, Cybersecurity, Supervised Learning

Introduction

The rapid expansion of digital technologies and internet usage has significantly increased the risk of cyber threats, among which malware remains one of the most dangerous. Malware refers to malicious software designed to disrupt, damage, or gain unauthorized access to computer systems. Over time, malware has evolved from simple viruses to complex forms such as ransomware, spyware, and rootkits.

Traditional malware detection systems rely heavily on signature-based techniques, which identify malware based on known patterns. However, these methods are ineffective against zero-day attacks and polymorphic malware, as such threats continuously change their structure to evade detection [1].

To overcome these limitations, machine learning and data mining techniques have been introduced in the field of cybersecurity. These approaches enable systems to learn patterns from data and detect unknown malware effectively. Supervised learning techniques, in particular, have shown promising results in malware classification by using labeled datasets to distinguish between malicious and benign files [2].

This paper presents a malware classification system based on supervised data mining techniques, aiming to improve detection accuracy and provide a reliable solution for modern cybersecurity challenges.

Literature Review

Recent studies have explored the use of machine learning techniques for malware detection and classification. Data mining approaches have been widely used to analyze malware behavior and classify malicious files based on extracted features. Researchers have demonstrated that classification algorithms such as Decision Trees, Naïve Bayes, and Support Vector Machines can effectively detect malware by analyzing patterns in executable files [3].

Several studies have highlighted the limitations of traditional signature-based detection methods, emphasizing their inability to detect new and evolving malware variants. To address this issue, hybrid approaches combining static and dynamic analysis have been proposed, which improve detection accuracy by analyzing both code structure and runtime behavior [4].

Moreover, recent advancements in machine learning, including ensemble methods such as Random Forest, have shown significant improvements in classification performance. These models reduce overfitting and enhance prediction accuracy by combining multiple classifiers [5].

Despite these advancements, existing systems still face challenges such as high computational cost, feature selection complexity, and difficulty in detecting highly obfuscated malware.

Problem Statement

The increasing volume and complexity of malware present a significant challenge for existing cybersecurity systems. Traditional detection techniques, particularly signature-based methods, rely on predefined patterns and are unable to detect newly generated or modified malware. As a result, these systems fail to identify zero-day attacks and polymorphic malware, leading to reduced detection accuracy and increased security risks.

Although machine learning-based approaches have improved malware detection, they still suffer from several limitations, including dependency on limited features, high computational overhead, and difficulty in achieving real-time detection. Additionally, many models struggle with balancing accuracy and false positive rates, which affects their reliability in practical applications.

Therefore, there is a need for an efficient malware classification system that utilizes supervised data mining techniques to accurately detect and classify malware while minimizing false positives and computational complexity. The system should be capable of adapting to evolving malware patterns and providing reliable performance in real-world environments.

Methodology

1. Data Collection

Data collection is a fundamental step in developing an effective malware classification system, as the quality and diversity of the dataset directly influence the performance of machine learning models. In this research, malware and benign datasets are collected from publicly available sources such as Kaggle repositories and malware analysis platforms. The dataset consists of executable files and their extracted features, including both malicious and legitimate samples.

To ensure robustness, the dataset includes different categories of malware such as ransomware, spyware, trojans, and worms. This diversity allows the model to learn various behavioral and structural patterns associated with different malware families. Additionally, care is taken to maintain a balanced dataset to prevent bias toward a particular class. A balanced dataset improves the model's ability to generalize and reduces the risk of overfitting.

Furthermore, the dataset is labeled, which is essential for supervised learning. Each sample is marked as either malicious or benign, enabling the model to learn classification boundaries effectively. The use of real-world datasets enhances the reliability and applicability of the proposed system in practical cybersecurity environments.

2. Data Preprocessing

Data preprocessing is a crucial stage that prepares raw data for effective model training. The collected dataset often contains inconsistencies such as missing values, duplicate entries, noise, and irrelevant attributes, which can negatively affect the performance of machine learning algorithms. Therefore, several preprocessing techniques are applied to clean and standardize the data.

Initially, missing values are handled either by removing incomplete records or by replacing them with appropriate statistical values such as mean or median. Duplicate entries are identified and removed to avoid redundancy in training. Noise in the data is minimized through filtering techniques, ensuring that only meaningful information is retained.

In addition, categorical data is converted into numerical form using encoding techniques, as machine learning algorithms require numerical input. Feature scaling and normalization are also applied to bring all feature values into a similar range, preventing bias toward features with larger magnitudes. These preprocessing steps

ensure that the dataset is consistent, clean, and suitable for training high-performance models.

3. Feature Extraction

Feature extraction plays a vital role in malware classification, as it determines how effectively the model can differentiate between malicious and benign software. In this research, both static and behavioral features are considered to capture comprehensive information about executable files.

Static features are extracted without executing the program and include attributes such as file size, header information, opcode sequences, and imported libraries. These features provide insights into the structure and composition of the file. On the other hand, behavioral features are obtained by analyzing the runtime behavior of the program, including API calls, system interactions, network activity, and registry modifications.

The combination of static and dynamic features enhances the model's ability to detect complex and obfuscated malware. Feature selection techniques are also applied to reduce dimensionality and eliminate irrelevant attributes. This not only improves model performance but also reduces computational cost. By selecting the most relevant features, the system becomes more efficient and accurate in identifying malicious patterns.

4. Model Training

Model training is the process in which machine learning algorithms learn patterns from the prepared dataset. In this study, supervised learning techniques are used, where the model is trained using labeled data. The dataset is divided into training and testing sets, typically in a ratio such as 80:20, to evaluate model performance effectively.

Three major classification algorithms are implemented: Decision Tree, Random Forest, and Support Vector Machine (SVM). The Decision Tree algorithm creates a hierarchical structure of decisions based on feature values, making it easy to interpret. Random Forest, an ensemble learning method, combines multiple decision trees to improve accuracy and reduce overfitting. It is particularly effective in handling large datasets with complex relationships.

Support Vector Machine is used for its ability to handle high-dimensional data and create optimal decision boundaries between classes. During training, hyperparameters of these models are tuned to achieve optimal performance. The training process involves iterative learning, where the model adjusts its parameters to minimize classification errors and improve prediction accuracy.

5. Prediction and Evaluation

After the training phase, the developed model is used to classify new and unseen data samples. The prediction process involves feeding input features into the trained

model, which then determines whether the file is malicious or benign based on learned patterns.

To evaluate the effectiveness of the model, several performance metrics are used, including accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of the model, while precision indicates how many predicted malware instances are actually malicious. Recall evaluates the model's ability to detect all actual malware samples, and F1-score provides a balance between precision and recall.

In addition to these metrics, confusion matrix analysis is performed to understand the distribution of true positives, true negatives, false positives, and false negatives. This helps in identifying potential weaknesses in the model. The evaluation results demonstrate that the proposed system achieves high accuracy with minimal false positives, making it suitable for real-world deployment.

Furthermore, the system is capable of real-time prediction, enabling it to detect malware efficiently in practical applications such as antivirus software and network security systems.

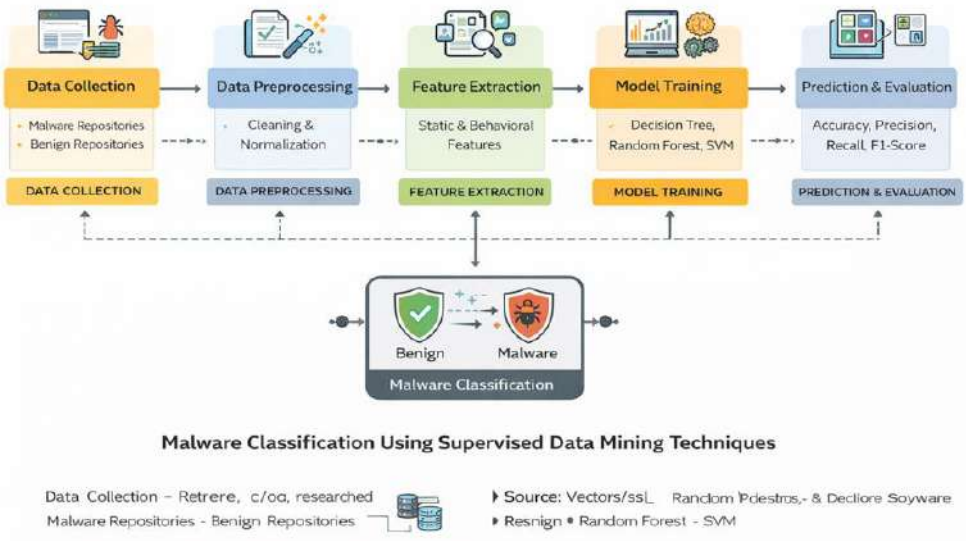


Fig. 1: Architecture of Proposed Malware Classification System

Results and Analysis

The performance of the proposed malware classification system is evaluated using standard machine learning metrics. The dataset is divided into training and testing sets to ensure unbiased evaluation. The models are trained using the training dataset and tested on unseen data.

The experimental results indicate that the Random Forest algorithm outperforms other models in terms of accuracy and reliability. The model achieves an accuracy of approximately 99%, demonstrating its effectiveness in correctly classifying

malware and benign files. Precision and recall values are also high, indicating that the system minimizes false positives and successfully detects most malware instances.

The results confirm that supervised data mining techniques significantly improve malware detection performance compared to traditional methods. The use of ensemble learning further enhances model accuracy and stability.

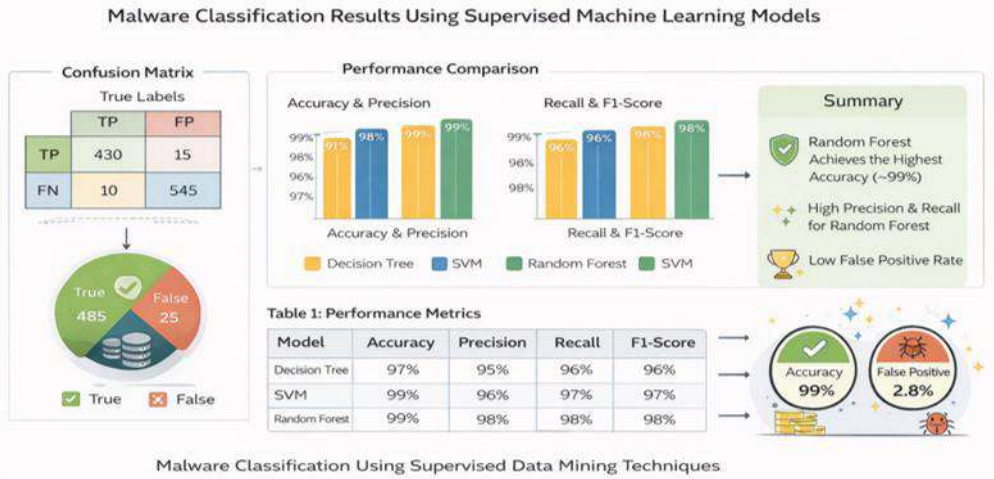


Fig. 2: Comparative Analysis of Machine Learning Models Based on Accuracy, Precision, Recall, and F1-Score

Discussion

The findings of this study highlight the effectiveness of machine learning techniques in malware classification. The proposed system successfully addresses the limitations of traditional detection methods by utilizing feature-based classification and supervised learning algorithms. The use of multiple features improves detection accuracy, while ensemble models provide better generalization.

However, challenges such as computational complexity and the need for large datasets still exist. Additionally, highly obfuscated malware remains difficult to detect, which requires further research in advanced techniques such as deep learning and behavioral analysis.

Conclusion

This paper presents a malware classification system using supervised data mining techniques to address the limitations of traditional detection methods. The study demonstrates that machine learning algorithms, particularly Random Forest, provide high accuracy and reliable performance in detecting malware. The system effectively classifies malicious and benign files by analyzing extracted features and learning patterns from data.

The results indicate that supervised learning techniques offer a scalable and efficient solution for malware detection in modern cybersecurity environments. The

proposed system can be further enhanced by integrating advanced techniques and real-time detection capabilities.

References

1. Ayeni, O. A., Owolafe, O., & Akinsola, O. (2021). Malware Detection Using Machine Learning. Proceedings of the International Conference for Internet Technology and Secured Transactions (ICITST). https://doi.org/10.20533/ICITST.2021.0008
2. Pokhrel, S. (2022). ML-Based Malware Classification Techniques. International Journal of Computer Applications.
3. Norouzi, M., Souri, A., & Samad, M. S. (2016). A Data Mining Classification Approach for Malware Detection in Android. Journal of Information Security and Applications.
4. Najari, S., Hamdan, M., & Al-Zoubi, A. M. (2020). Malware Detection Using Data Mining Techniques: A Review. International Journal of Advanced Computer Science and Applications.
5. Anil Kumar, S., & Shravani, K. (2021). An ML Approach for Malware Detection and Classification. International Journal of Recent Technology and Engineering.

ABOUT THE EDITORS



Dr. Sadguru Prakash

He is an experienced academician, researcher, author, editor, and science educator who has been serving in the Department of Zoology, M.L.K.P.G. College, Balrampur, Uttar Pradesh, since 2003. With more than two decades of teaching and research experience, he has made significant contributions to Zoology, biodiversity, environmental science, aquatic biology, and sustainable development. Dr. Prakash has published more than 160 research papers in reputed national and international journals and 45 research papers in proceedings and edited books. He has authored 05 books and edited 13 books. He has participated in and organized more than 40 national and international conferences, seminars, and symposia and has delivered invited lectures as a Resource Person on various academic and scientific topics. He has been honoured with fellowships from several professional organizations, including the Blue Planet Society (FBPS), Asian Biological Research Foundation (FABRF), Glocal Environment & Social Association (FGESA), Nepal Aquaculture Society (NEAQUAS), Zoological Society of India (ZSI), Global Academic Society (GAS), International Institute of Organized Research (I2OR), and International Association of Zoologists (FIAZ). His major awards and recognitions include the Dr. A.P.J. Abdul Kalam Green Environment Promotion Award (GESA), Innovative Educationist Award (GESA), Innovative Zoologist Award (ABRF), Lal Bahadur Shastri Glocal Award for Biodiversity (GESA), Biodiversity Conservation Awareness Award (IFEE), Best Teacher Award for Zoological Innovations (ABRF), Global Eminent Educator Award (GAS), International Scientist Awards (VDGOOD, 2021), ABRF Excellence Award for Zoological Research (2022), Prof. S.S. Khanna Gold Medal Award (NEAQUAS, 2022), and the Indo Global International Excellence Award (ECRD Research and Education Council, Hisar, Haryana, 2025). Dr. Prakash has supervised two Ph.D. scholars, one M.Phil. scholar, and more than 20 M.Sc. dissertation projects. He has also contributed extensively to academic administration and institutional development. His academic responsibilities have included serving as Examination In-charge, Member of the Proctorial Board, Assistant Coordinator of IGNOU, and Member of the IQAC of the college. He has also served as a Member of the Board of Studies in Zoology at Siddharth University, Kapilvastu, Siddharthnagar; Maa Pateswari University, Balrampur; and KDC Autonomous College, Bahraich. Through his sustained contributions to teaching, research, academic leadership, science communication, and biodiversity and environmental awareness, Dr. Prakash has established himself as a dedicated educator and active researcher in the field of Zoology and environmental sciences.



Mr. Avinash Singh

Avinash Singh working as a Assistant Professor and Head, Department of Computer Application at Maharani Lal Kunwari (P.G.) College Balrampur . His research domains include Blockchain and Fin-tech security, Machine Learning, and Mobile Ad-Hoc Networks. Mr. Singh has various publications in Scopus indexed journals, ESCI journals and having book chapters in different reputed book publisher including Tylor and Francis and many more.

Nature Light Publications

309 West 11, Manjari VSI Road, Manjari Bk.,
Haveli, Pune- 412 307.
Website: www.naturelightpublications.com
Email: naturelightpublications@gmail.com
Contact No: +919822489040 / 9922489040

Nature Light Publications

Nature- 130

ISBN: 978-81-69492-03-4
9 788169 492034
Price- 750/-