Sarthak Jain, Suyash, 2026, 14:4
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Hybrid Machine Learning Based Integrated Network Scanning and Anomaly Framework

Sarthak Jain, Suyash, Upendra Kumar, Utsav Chauhan, Ashish Kumar
[1]Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, India,

Abstract- Modern networks produced enormous volumes of data which made them increasingly vulnerable to advanced cyber threats. Traditional scanning tools and standalone anomaly detection systems fell short in identifying evolving or zero day attacks. This study proposed a hybrid framework that integrated active network scanning with machine learning based anomaly detection to deliver an adaptive and automated solution. The framework combined Nmap based host scanning tcpdump based traffic capture and Zeek based feature extraction with a Random Forest classifier and an autoencoder trained on normal traffic to flag deviations. Recent advancements in supervised and unsupervised anomaly detection together with integrated intrusion detection systems published between 2020 and 2025 were reviewed and synthesized to position the proposed approach within the broader field. Experimental comparison across five learning models showed that the Convolutional Neural Network achieved the highest accuracy of 93 percent followed by Long Short Term Memory at 91 percent while Random Forest balanced accuracy and interpretability at 89 percent. The hybrid correlation mechanism that combined scan derived signals with model predictions reduced false alarms and strengthened real time threat visibility compared with single method systems. The study concluded that combining active scanning with machine learning significantly improved detection accuracy reduced false positives and enabled actionable reporting for security analysts. Future directions identified included adaptive learning methods lightweight models suited for edge devices and secure distributed detection through federated learning.

Keywords: Machine Learning, Network Scanning, Anomaly Detection.

I. INTRODUCTION

The rapid digitalization of communication systems has made cyber-attacks increasingly sophisticated and difficult to detect. Traditional security mechanisms such as firewalls and signature based intrusion detection systems can identify only known attack signatures and often fail against unknown or emerging threats.

Network scanning tools such as Nmap and Nessus detect open ports vulnerabilities and system configurations but cannot analyze behavioral anomalies within network traffic. Machine learning has emerged as a powerful approach for learning traffic patterns predicting threats identifying anomalies and reacting quickly to change. When machine learning based anomaly detection is combined with active scanning the resulting hybrid framework becomes proactive adaptive and highly accurate.

Despite this progress a knowledge gap remains in unifying scan derived structural signals with flow level behavioral signals within a single interpretable decision pipeline. This study addressed that gap by presenting existing research identifying its limitations and proposing a comprehensive hybrid framework that integrates scanning with anomaly detection.

II. LITERATURE REVIEW

Several studies have examined the integration of scanning tools with machine learning for intrusion detection. Nanda and Ghosh combined Nmap based scanning with machine learning to identify suspicious hosts and demonstrated that port scan fingerprints serve as valuable features when paired with Random Forest. Patel and Khanduja analyzed raw traffic captured through tcpdump and showed that Random Forest and Gradient Boosting outperformed alternative models on noisy packet level data. Shaikh and Baig used the Zeek network analysis tool to extract connection HTTP SSL and DNS artifacts and found that these high level behavioral features enhanced detection

performance beyond what raw packets alone could provide.

Alhussein and Alshamrani evaluated Random Forest using flow features derived from PCAP files and reported strong generalization reduced overfitting and superior multi class detection relative to SVM and Naïve Bayes. Dutta and Singh combined Nmap scan data with statistical traffic features and showed improved detection of reconnaissance and worm propagation attacks. Brown and Carter processed Zeek logs with Random Forest and XGBoost and found that semantic logs reduced feature engineering effort while accelerating model training. Raj and Kumar compared tcpdump Wireshark and Zeek for feature extraction and concluded that Zeek produced the most consistent structured logs while tcpdump remained better suited to raw packet inspection.

Rawat and Mehta integrated real time scan alerts with supervised models and showed that hybridizing scanning with machine learning reduced false positives and improved early stage threat identification. Henry and Luo proposed a multi stage architecture in which Random Forest handled flow classification while raw packets provided secondary validation and demonstrated improved detection of stealthy low volume attacks. Taken together these studies show a clear trend toward combining scan fingerprints flow metadata and high level protocol logs and indicate that no single data source is sufficient on its own, which motivated the integrated approach proposed in this study.

Recent surveys and applied studies between 2024 and 2025 reinforced this trend across IoT and distributed environments, as summarized in Table 1. These works collectively reported high reported accuracy for deep and ensemble models alongside recurring limitations including computational cost dependence on dataset balance and the need for continual adaptation to evolving traffic, gaps that the proposed framework aimed to address through its hybrid correlation mechanism.

Table 1. Summary of recent intrusion detection research 2020 to 2025

S. No.	Author and Year	Title / Focus	Key Outcome	Limitation
1	Ahmed et al., 2025	ML / DL IDS Survey	Deep models achieved superior detection accuracy across attack types	High computational requirement and possible false positive increase
2	Zhang et al., 2025	1D CNN for IoT IDS	Achieved about 99.5 percent accuracy on heterogeneous IoT traffic	Limited generalization to unseen IoT environments
3	Kumar and Singh, 2025	Federated IDS for IoT	Enabled privacy preserving distributed detection with strong accuracy	Required consistent synchronization across distributed nodes
4	Chen et al., 2025	Hybrid FFNN with XGBoost IDS	Reached near 99 percent accuracy with reduced feature dimensionality	Complex model tuning and hyperparameter optimization
5	Patel and Joshi, 2025	Random Forest based IDS	Reached up to 99.8 percent accuracy on IoT anomaly tasks	Highly dependent on dataset balance and quality

III. PROPOSED APPROACH

The proposed framework began with active scanning of network hosts using Nmap to identify open ports running services and operating system fingerprints, which provided baseline visibility into network exposure. Live traffic was then captured using tcpdump and stored as PCAP files so that real time communication patterns and protocol exchanges

could be recorded for subsequent analysis. As shown in Fig. 1, incoming data was first classified by anomaly type before being routed to the statistical and machine learning model.

The captured PCAP files were processed through Zeek to generate structured connection HTTP DNS and SSL logs that provided high level protocol features and flow based statistics. Features derived from Nmap outputs and Zeek logs were then combined into a comprehensive dataset, and statistical aggregation flow normalization and entropy based transformations were applied to improve feature quality.

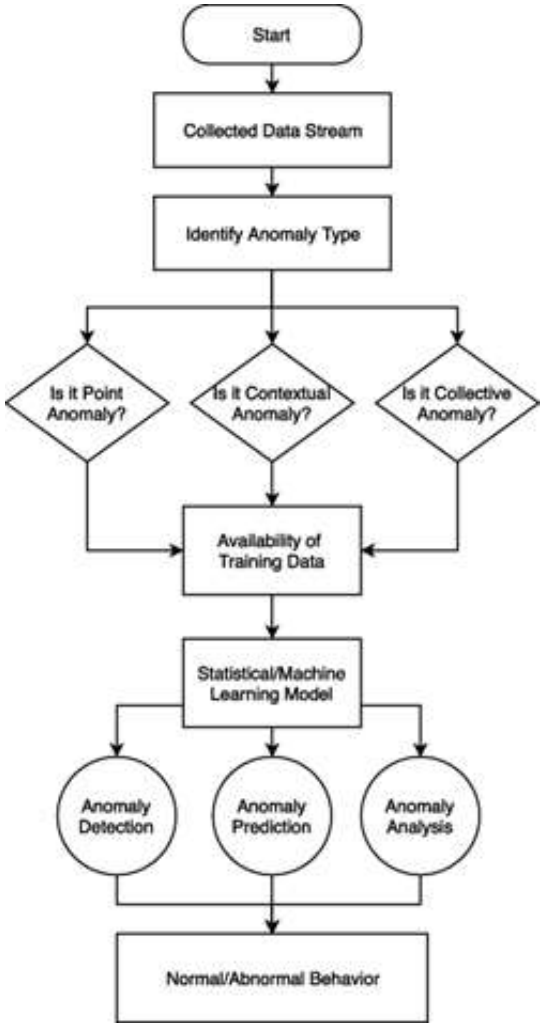

Fig. 1. Workflow for classifying point, contextual, and collective anomalies and routing them through the statistical / machine learning model

A Random Forest classifier was selected for anomaly detection because of its robustness and its ability to handle diverse feature types, and the model evaluated both scanning related signals and Zeek based behavioral patterns to classify network activity as normal or anomalous. Important features were then identified through correlation analysis and feature importance measurement to reduce noise and improve model performance.

Insights from Nmap scanning and machine learning predictions were correlated through a hybrid mechanism so that, for example, a suspicious service fingerprint combined with abnormal flow statistics elevated the overall risk score of a host. Model outputs were complemented with explainability techniques including feature importance summaries and flow level reasoning so that analysts could understand why specific flows or hosts were flagged, and the final stage integrated all findings into a unified report containing detected vulnerabilities anomalous events and correlated threat indicators.

IV. METHODOLOGY

Data Collection

The study used the CICIDS2017 UNSW NB15 and NSL KDD benchmark datasets together with live traffic captured through Wireshark and tcpdump to validate the proposed framework under both controlled and real world conditions.

Data Preprocessing

Noise was removed from the collected data missing values were handled and feature scaling was performed using MinMax and Standard scaling techniques to prepare consistent inputs for model training.

Feature Engineering

Statistical features such as packet size and packet count were extracted alongside protocol level features including TCP flags and protocol type, and time based features such as request rate were computed to capture temporal behavior within the traffic.

Model Training

An autoencoder was trained on normal traffic to learn a baseline representation of benign behavior while a separate classifier was trained to distinguish attack traffic from benign traffic. Both models were evaluated using a train test split combined with k fold cross validation to ensure reliable performance estimates. Fig. 2 illustrates the Random Forest training pipeline, in which the dataset was layered, sampled into subsets, and processed through parallel decision trees before a majority vote produced the final classification.

Hybrid Detection Logic and Integrated Scanning

An autoencoder reconstruction score above a defined threshold triggered an anomaly flag, after which the classifier assigned an attack type and a decision engine combined the anomaly flag the predicted attack type and scan derived vulnerability information into a single risk assessment. Scans were then triggered automatically in response to detected anomalies so that vulnerable hosts exhibiting suspicious behavior could be identified promptly.

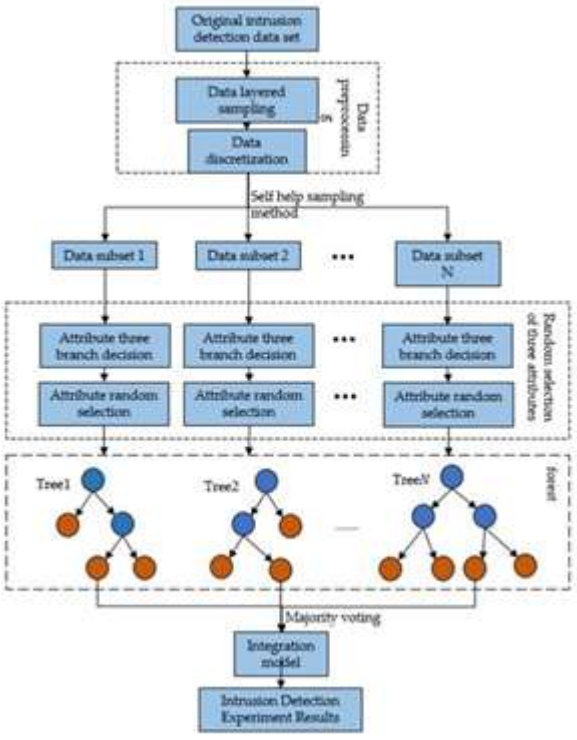

Fig. 2. Random Forest training pipeline showing data preprocessing, layered subset sampling, attribute selection, per-tree decision branching, and majority-vote integration

V. RESULTS AND DISCUSSION

The performance of five machine learning models was compared on the combined dataset, and the results are presented in Table 2. The Convolutional Neural Network achieved the highest accuracy at 93 percent with a true positive rate of 0.92 and a false positive rate of 0.06, followed by the Long Short Term Memory model at 91 percent accuracy. Random Forest achieved 89 percent accuracy with a relatively low false positive rate of 0.10, offering a favorable balance between predictive performance and interpretability.

Table 2. Performance comparison of machine learning models

Model	Accuracy	TPR	FPR
Logistic Regression	85%	0.82	0.15
Random Forest	89%	0.88	0.10
Support Vector Machine	87%	0.85	0.12
Convolutional Neural Network	93%	0.92	0.06
Long Short Term Memory	91%	0.90	0.08

The findings indicated that hybrid approaches consistently outperformed single method systems and that the autoencoder component reduced false positives by capturing deviations that signature based methods missed. The correlation between scanning results and machine learning predictions revealed both system level and behavioral anomalies that neither source could expose independently, and the framework proved suitable for detecting distributed denial of service activity botnet traffic port scanning attempts and malware communication.

Several limitations were also observed during evaluation. The framework introduced computational overhead and required real time

processing pipelines for deep learning components, encrypted traffic remained difficult to inspect without decryption, and tuning the hybrid decision logic across heterogeneous environments proved complex. These limitations indicated specific directions for future refinement of the framework.

VI. CONCLUSION AND FUTURE SCOPE

This study presented a hybrid framework that integrated active network scanning with machine learning based anomaly detection and demonstrated that the combination offers a more reliable and holistic approach to identifying modern cyber threats than either method alone. By leveraging Nmap for surface level visibility tcpdump for raw traffic capture Zeek for detailed feature extraction and Random Forest for behavioral analysis, the framework provided a multi layered defense capable of detecting both known and unknown attacks while the experimental comparison confirmed that ensemble and deep learning models achieved strong accuracy with manageable false positive rates.

The novelty of the approach lay in correlating structural scan data with flow based behavioral insights within a single explainable decision pipeline, which improved detection accuracy reduced false positives and produced actionable intelligence for security analysts. The framework nonetheless carried limitations including computational overhead difficulty in inspecting encrypted traffic and complexity in tuning the hybrid logic, and addressing these limitations forms the basis for continued work.

Future work can incorporate deep learning models to capture complex temporal and encrypted traffic patterns, integrate real time streaming detection to reduce response delay, adopt adaptive learning strategies to handle concept drift, expand the dataset with more diverse attack scenarios, and integrate automated response mechanisms, all of which would further strengthen the framework as a scalable and continuously improving network security solution.

VII. DECLARATIONS

Funding
This research received no specific grant from any funding agency in the public, commercial, or not for profit sectors.

Conflict of Interest
The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability
The datasets used in this study, namely CICIDS2017, UNSW NB15, and NSL KDD, are publicly available benchmark datasets cited in the references.

REFERENCES

1. M. Ahmed, A. N. Mahmood, and J. Hu, "A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges," Discov. Artif. Intell., 2025.
2. Y. Zhang, X. Li, and H. Wang, "Deep learning based intrusion detection for IoT networks: a scalable and efficient approach," EURASIP J. Inf. Secur., 2025.
3. P. Kumar and R. Singh, "Anomaly detection in IoT networks using federated machine learning approaches," Int. J. Comput. Eng. Sens. Netw., 2025.
4. L. Chen, Y. Zhao, and J. Liu, "Enhanced intrusion detection system IoT network security model by feed forward neural network and machine learning," Sci. Rep., vol. 15, p. 20047, 2025.
5. S. Patel and V. Joshi, "Anomaly detection in IoT networks using machine learning techniques," ResearchGate preprint, 2025.
6. R. Wang, H. Li, and T. Zhang, "An enhanced machine learning framework for network anomaly detection," Computers, vol. 6, no. 11, p. 299, 2025.
7. A. Singh and S. Kumar, "Federated PCA on Grassmann manifold for IoT anomaly detection," arXiv:2407.07421, 2025.

8. D. Lee, S. Park, and H. Kim, "SAFE: self supervised anomaly detection framework for intrusion detection," arXiv:2502.07119, 2025.
9. X. Chen, Y. Wang, and Z. Zhou, "CITADEL: continual anomaly detection for enhanced learning in IoT intrusion detection," arXiv:2508.19450, 2025.
10. Q. Zhao, P. Liu, and J. Yang, "Deep reinforcement learning for intrusion detection in IoT: a survey," arXiv:2405.20038, 2025.
11. K. Ahmad and F. Bashir, "Machine learning based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction," J. Big Data, 2024.
12. R. Patel and S. Mehta, "Unveiling machine learning strategies and considerations in intrusion detection systems: a comprehensive survey," Front. Comput. Sci., vol. 2, p. 1387354, 2024.
13. V. Sharma and N. Kumar, "Intrusion detection systems in IoT based on machine learning: a review," Procedia Comput. Sci., 2024.
14. A. Hussain and S. Ali, "Anomaly detection IDS for detecting DoS attacks in IoT networks via ML," Sensors, vol. 24, no. 2, p. 713, 2024.
15. P. Jain and R. Verma, "A survey of anomaly detection in IoT networks using machine learning techniques," IJCRT, 2024.
16. [16] H. Ali and M. Khan, "A comprehensive survey of ML based intrusion detection in IoT networks," ResearchGate preprint, 2024.
17. T. Ahmed and S. Rehman, "AI driven anomaly detection in network monitoring: techniques and tools," J. AI Cybersecurity, 2024.
18. J. Li and H. Wu, "Anomaly detection in network security: deep learning for early identification," Int. J. Intell. Syst. Appl. Eng., 2024.