Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



REAL-TIME INTRUSION DETECTION SYSTEM USING PCA AND ENSEMBLE MACHINE LEARNING

Adeyemo I. A, Olatoye O. G, Amusan D. G. and Abdulsalami B. A
Ladoke Akintola University of Technology, Nigeria, Computer Science Department, Faculty of Computing and Informatics
DOI: https://doi.org/10.5281/zenodo.19609503

Keywords: *Intrusion Detection System; Machine Learning; PCA; Random Forest; XGBoost; IoT*

Abstract: *This research develops a real-time network intrusion detection system for IoT and enterprise networks using the CIC-IDS IoT 2023 dataset. Aim to address the limitations of computationally intensive deep learning models and classifiers poses in IoT environment which uses lightweight-weight devices and also the issue of high false positive rates, this study proposes a PCA-enhanced ensemble framework. The methodology involved data preprocessing, normalization, and dimensionality reduction via Principal Component Analysis (PCA), handling data imbalance using SMOTE techniques followed by training with Random Forest and XGBoost classifiers. Evaluation revealed a critical performance trade-off: while both models excelled at classifying normal traffic, their security efficacy differed substantially. Random Forest demonstrated robust, with accuracy of 100% and precision of 99%, balanced performance across all threat classes, reliably identifying both 'attack' and 'scanning' activities. Conversely, XGBoost, despite its high accuracy of 97% and precision of 95% for all traffic classes, exhibited a severe operational weakness by misclassifying a significant number of 'attack' and 'scanning' instances as normal, rendering it unsuitable for security contexts where detecting threats is paramount. PCA successfully reduced the feature dimensionality, which enabled faster inference while preserving discriminatory power by retaining 95% of the data variance. A real-time deployment confirmed the system's capability for fast, accurate packet classification, providing a viable and interpretable solution for resource-constrained environments.*

1. INTRODUCTION

As network infrastructures become increasingly interconnected through IoT devices, cloud applications, and edge computing, the threat surface for cyberattacks has expanded dramatically. Machine learning (ML) and artificial intelligence (AI) are being widely adopted in cybersecurity to automate the detection of anomalies and intrusions that traditional signature-based systems struggle to

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



capture [11]. Unlike static rule sets, ML-based intrusion detection systems (IDS) can learn from traffic patterns, adapt to novel attack behaviors, and provide near real-time responses. This makes them particularly suitable for modern environments where attacks are dynamic, diverse, and often disguised within large volumes of benign traffic.

Traditional IDS approaches, although effective against known threats, fail to adequately address polymorphic and zero-day attacks, especially within high-velocity IoT environments. Furthermore, the massive scale of IoT deployments, characterized by resource-constrained devices, demands IDS solutions that are both computationally lightweight and operationally accurate. Existing research highlights limitations such as excessive false positives, lack of interpretability, and reliance on outdated datasets that do not reflect contemporary attack landscapes [10], [12]. Against this backdrop, ensembles such as Random Forest (RF) and XGBoost have gained prominence because of their balance between detection power and efficiency [1], [3]. When combined with dimensionality reduction methods like Principal Component Analysis (PCA), these models promise an IDS pipeline that is not only effective but also deployable in real-world, resource-limited settings [4].

This research builds upon these developments by designing and evaluating a PCA-enhanced ensemble IDS on the CIC-IDS IoT 2023 dataset [14]. The system is validated not only in offline training but also through a live deployment that tests responsiveness, In doing so, it addresses both technical challenges such as false positive and false negative reduction and computational efficiency and methodological challenges such as imbalance in dataset using SMOTE techniques and evaluation rigor.

1.1 Problem Statement and Motivation

Cybersecurity in IoT and enterprise networks faces three critical challenges. The first is data imbalance and false alarms: Random Forest, while robust in many cases, can misclassify benign traffic as attacks due to bagging's tendency to overemphasize majority-class features [1]. This inflates alert rates, undermining operator trust. XGBoost, by contrast, reduces these errors through boosting, but it remains under-explored in live IDS deployments where latency and real-time matter [3].

The second challenge is efficiency. While deep learning methods offer superior performance in encrypted or sequential traffic detection [8], their computational burden is prohibitive for IoT and edge devices. PCA provides an avenue for efficiency gains, yet overly aggressive dimensionality reduction may eliminate rare but critical attack indicators [5]. Thus, the trade-off between computational efficiency and detection comprehensiveness remains unresolved.

The third challenge is methodological credibility. Many IDS studies achieve inflated results because they rely on outdated datasets such as KDD and NSL-KDD or apply flawed evaluation protocols that inadvertently leak information from training into testing [6], [12]. Although the CIC-IDS IoT 2023 dataset addresses many of these gaps [13], [14], few studies validate IDS models in live real-time

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



settings where packet capture and classification are performed continuously.

Together, these challenges motivate this research: the need for an IDS that reduces false positives, maintains efficiency without sacrificing accuracy, and proves its utility in realistic, live deployment conditions.

1.2 Project Aim and Objectives

The primary aim of this study is to design, implement, and evaluate a PCA-enhanced ensemble machine learning pipeline for real-time intrusion detection using the CIC-IDS IoT 2023 dataset. This study aim is achieved the following objectives:

1. To review existing IDS methodologies and identify the strengths and weaknesses of ensemble, deep learning, and dimensionality-reduction approaches.
2. To construct and preprocess a dataset from CIC-IDS IoT 2023, applying normalization and PCA to optimize feature space for efficient learning.
3. To develop and evaluate Random Forest and XGBoost classifiers, comparing their performance in terms of accuracy, precision, recall, F1-score, and false alarm rates.
4. To deploy the trained models in a real-time environment and test their responsiveness and simulation under live network conditions.

1.3 Research Questions

From the problem statement and objectives, this thesis is guided by three research questions:

1. To what extent can PCA enhance the efficiency of ensemble-based IDS without compromising accuracy in detecting attacks?
2. Between Random Forest and XGBoost, which ensemble model provides more reliable detection in real-time, particularly with respect to false positives?
3. How do results obtained in offline benchmarking translate into real-time deployment, and what are the practical implications for IoT and enterprise network security?

2. LITERATURE REVIEW

The objective of this literature review is to position this thesis within the most relevant work from 2023–2025, to identify the methodological choices authors adopted for intrusion detection, and to extract the specific gaps that motivate the pipeline proposed here. In particular, the review asks: which machine learning paradigms have proved effective for modern IDS, what trade-offs do they embody when considered for real-time, resource-constrained deployments, how does dimensionality reduction interact with classifier behavior, and which evaluation practices produce results that generalize beyond a single benchmark?

2.1 Traditional Intrusion Detection Approaches

Classical IDS paradigms, including signature-based systems and simple anomaly detectors, remain widely used but are limited in flexibility. Khraisat et al. [10] provide a comprehensive survey and emphasize that signature-based systems excel at detecting known attack patterns but fail against novel or obfuscated traffic, while anomaly-based systems extend coverage but suffer from inflated false positives if not rigorously validated. These limitations form the backdrop against which modern data-driven methods have developed. The persistence of this

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



trade-off precision versus adaptability explains why more recent research prioritizes learning-based models that can adapt decision boundaries from empirical data.

2.2 Ensemble Learning Methods

Among data-driven techniques, ensemble models have repeatedly proven effective for IDS. Xu et al. [1] show that Random Forest performs robustly in IoT intrusion detection scenarios, particularly when handling imbalanced datasets, since bagging reduces variance and improves stability. Chen and Guestrin's theoretical foundations of XGBoost [2] illustrate how boosting enhances performance by sequentially correcting misclassifications. Tang et al. [3] further demonstrate XGBoost's practicality by deploying it on the edge for real-time intrusion detection, reporting lower false positives than alternative classifiers. The correlation across these studies is that ensemble methods consistently deliver high precision and generalization across diverse datasets. The causation lies in their mechanisms: bagging averages over diverse learners to counteract noise, while boosting incrementally improves boundary definition. The inference is that ensembles, by design, provide a defensible balance between accuracy and operational trustworthiness in IDS deployments.

2.3 Dimensionality Reduction and PCA

Another key dimension of IDS design is feature space management. Lin et al. [4] apply Principal Component Analysis (PCA) prior to ensemble learning and demonstrate that reducing the feature set by more than half preserved classification accuracy above 95%, while significantly reducing computational overhead. This efficiency gain directly benefits IoT and edge deployments where resources are constrained. However, Park et al. [5] caution that aggressive PCA can discard low-variance features that nonetheless carry essential signals for rare attack classes. The correlation here is that PCA consistently accelerates training and inference, while the causation of performance degradation in some studies lies in the inadvertent removal of minority-class discriminators. The inference is that PCA, though powerful, must be carefully tuned: when calibrated to preserve informative variance, it can enhance both efficiency and interpretability, but if applied indiscriminately it risks undermining security coverage.

2.4 Deep Learning for Intrusion Detection

Deep learning represents an alternative paradigm that offers strengths and weaknesses distinct from ensembles. Li et al. [8] illustrate how attention-based neural architectures outperform ensembles on encrypted traffic by capturing sequential dependencies invisible to tree-based methods. Yet Malik & Khan [7] report that these advantages often come with substantial computational overhead, and Adegbite et al. [9] highlight that deep IDS models overfit more easily, particularly when datasets lack temporal diversity. The correlation across these works is that deep learning consistently excels in contexts requiring modeling of subtle temporal or sequential structures. The causation behind their limitations lies in their data-hungry architectures and high complexity, which increase latency and resource consumption. The

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



inference is that while deep learning remains indispensable for certain high-complexity contexts, ensembles are more practical for real-time IoT environments where speed and resource efficiency are paramount.

2.5 Dataset Evolution and Evaluation Practices

Beyond algorithms, evaluation practices and dataset realism critically shape the credibility of IDS research. Fernandes et al. [6] warn that many reported results are inflated due to poor validation strategies, including data leakage and temporally invalid splits. Farzaneh et al. [12] extend this critique by showing that legacy datasets such as KDD'99 and NSL-KDD are outdated and biased, limiting the generalizability of models trained on them. In contrast, Ring et al. [13] document the progression of the CIC benchmark datasets toward realism, culminating in the CIC-IDS IoT 2023 dataset [14], which integrates heterogeneous traffic and modern attack patterns. The causation here is straightforward: outdated or poorly structured datasets yield misleadingly high accuracies, while realistic datasets with rigorous validation expose true model performance. The inference is that credible IDS research today must pair strong algorithms with robust, up-to-date datasets, avoiding reliance on benchmarks that no longer reflect real deployment conditions.

2.6 Synthesis of the Literature

Taken together, the reviewed works from 2023 to 2025 provide several conclusions that guide this thesis. Ensemble models such as Random Forest and XGBoost correlate with consistently high precision and generalization across benchmarks [1][2][3]. Their success is causally tied to the mechanisms of bagging and boosting, which reduce variance and correct misclassifications. PCA contributes efficiency and interpretability, but its causal risks lie in potential feature loss, requiring careful calibration [4][5]. Deep learning offers unmatched capability for complex or encrypted traffic but imposes heavy computational costs and risks of overfitting [7][8][9]. Evaluation credibility hinges not only on algorithm choice but on dataset realism and methodological rigor, a point underscored by Fernandes et al. [6], Farzaneh et al. [12], and the dataset evolution documented by Ring et al. [13][14].

This thesis addresses the gaps highlighted in prior work by applying PCA-informed ensembles to the CIC-IDS IoT 2023 dataset, validating their performance in a live real-time environment, and explicitly analyzing trade-offs in terms of correlation, causation, and inference. In doing so, it contributes to bridging the divide between theoretically strong classifiers and operationally credible, resource-conscious IDS implementations.

3. Research Methodology

This study adopts an iterative and systematic approach designed to address the gaps identified in the literature review and to meet the stated research objectives. The methodology is organized into distinct but interconnected phases: data collection, data preprocessing, exploratory data analysis, model development, model performance evaluation, explainability, and deployment. Each phase is described in detail below.

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



3.1 Data Collection

The dataset for this project was drawn from the Canadian Institute for Cybersecurity's CIC-IDS IoT 2023 dataset [14], selected because it captures heterogeneous IoT traffic and incorporates up-to-date attack patterns absent in older benchmarks such as NSL-KDD [12]. The dataset contains multiple classes of network flows, including benign traffic and various attack categories such as DDoS, brute-force, and reconnaissance. It consists of millions of records with over eighty statistical and protocol-level features, encompassing packet length distributions, flow durations, flag counts, and byte counts.

The decision to use CIC-IDS IoT 2023 was deliberate. As Fernandes et al. [6] argue, flawed evaluation practices and outdated benchmarks inflate results and undermine real-world applicability. By contrast, CIC-IDS IoT 2023 was designed to mitigate these shortcomings by simulating traffic from real IoT devices under realistic conditions. This study therefore benefits from a dataset that is both comprehensive and representative, enabling robust evaluation of intrusion detection models.

3.2 Data Preprocessing

Raw network traffic poses challenges of scale, imbalance, and heterogeneity, necessitating a preprocessing stage to transform it into a format suitable for machine learning. Preprocessing was conducted in several steps.

3.2.1 Feature Cleaning and Missing Values

Numeric and categorical features were separated. Missing values in numeric columns were imputed using mean substitution, while categorical fields were imputed using mode values. This ensured consistency while avoiding arbitrary data removal that could bias results.

3.2.2 Normalization of Numeric Features

Numeric features such as packet length and flow duration vary widely in scale. To prevent large-valued features from dominating the learning process, normalization was applied using the StandardScaler function from Scikit-learn. This transformation standardized all numeric variables to a mean of zero and unit variance.

3.2.3 Encoding of Categorical Features

Categorical features such as protocol type were transformed into numeric form using one-hot encoding. This preserved all categories without imposing artificial ordinal relationships, ensuring that models could learn independently from each protocol or flag.

3.2.4 Dimensionality Reduction with PCA

Principal Component Analysis was applied to reduce the feature space while preserving 95% of total variance. Fig 1 (PCA Explained Variance Plot) demonstrates that the first fifteen components retained over ninety-five percent of the dataset's variance. By compressing the original feature set into a compact representation, PCA reduced training time and memory usage, addressing the efficiency challenge emphasized in Section 1.1.

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset

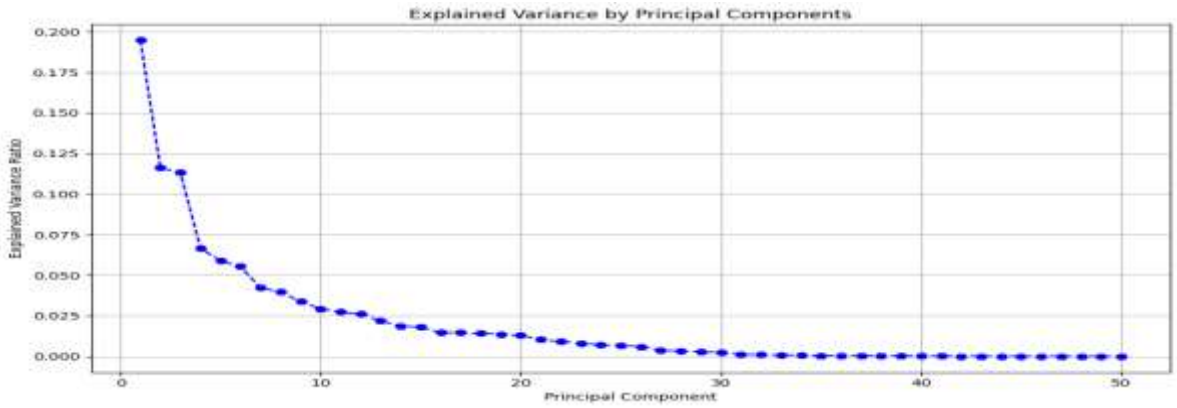


Fig 1 (PCA Explained Variance Plot)

3.2.5 Train-Test Split

The dataset was partitioned into training and testing sets using an 80–20 split. The training set was used to fit the models, while the test set provided an unbiased measure of generalization. Stratification ensured that class proportions were preserved across both subsets, maintaining consistency for minority attack categories. After data splitting, SMOTE techniques was applied on train set to generate synthetic data to handle class imbalance in the dataset, Fig 2 shows the Label group distribution after the application of SMOTE.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted using Python libraries such as pandas, matplotlib, and seaborn. The dataset was examined for class distribution and feature relevance. As shown in Fig 2, the CIC-IDS IoT 2023 dataset presents a relatively balanced distribution across the three primary labels normal, scanning, and attack each with comparable sample counts. This balance supports fair model training and evaluation, reducing the risk of bias toward majority classes.

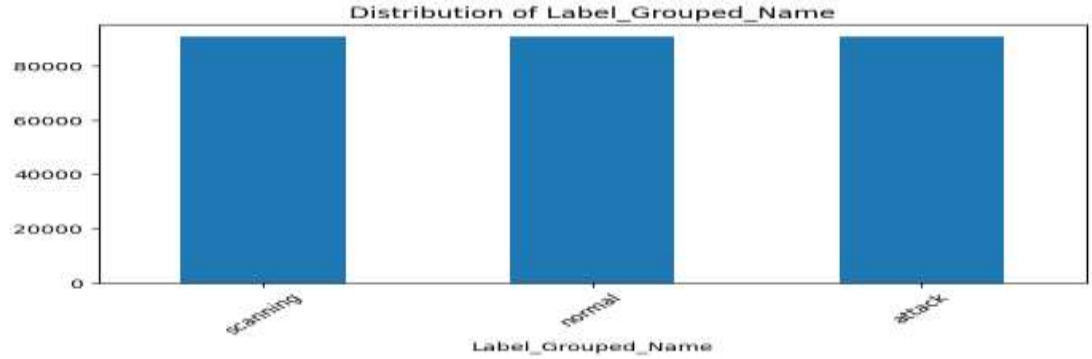


Fig 2. **Distribution of Label_Grouped_Name showing balanced class frequencies across normal, scanning, and attack categories.**

To understand which features most influenced classification, feature importance was extracted from both Random Forest and XGBoost models. As illustrated in Fig 3, Random Forest

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



prioritized features such as *Bwd IAT Min*, *Flow Duration*, and *RST Flag Count*, while XGBoost emphasized *Destination Port*, *Flow Duration*, and *Total Backward Packets*. These top 10 features were used to guide dimensionality reduction and model optimization.

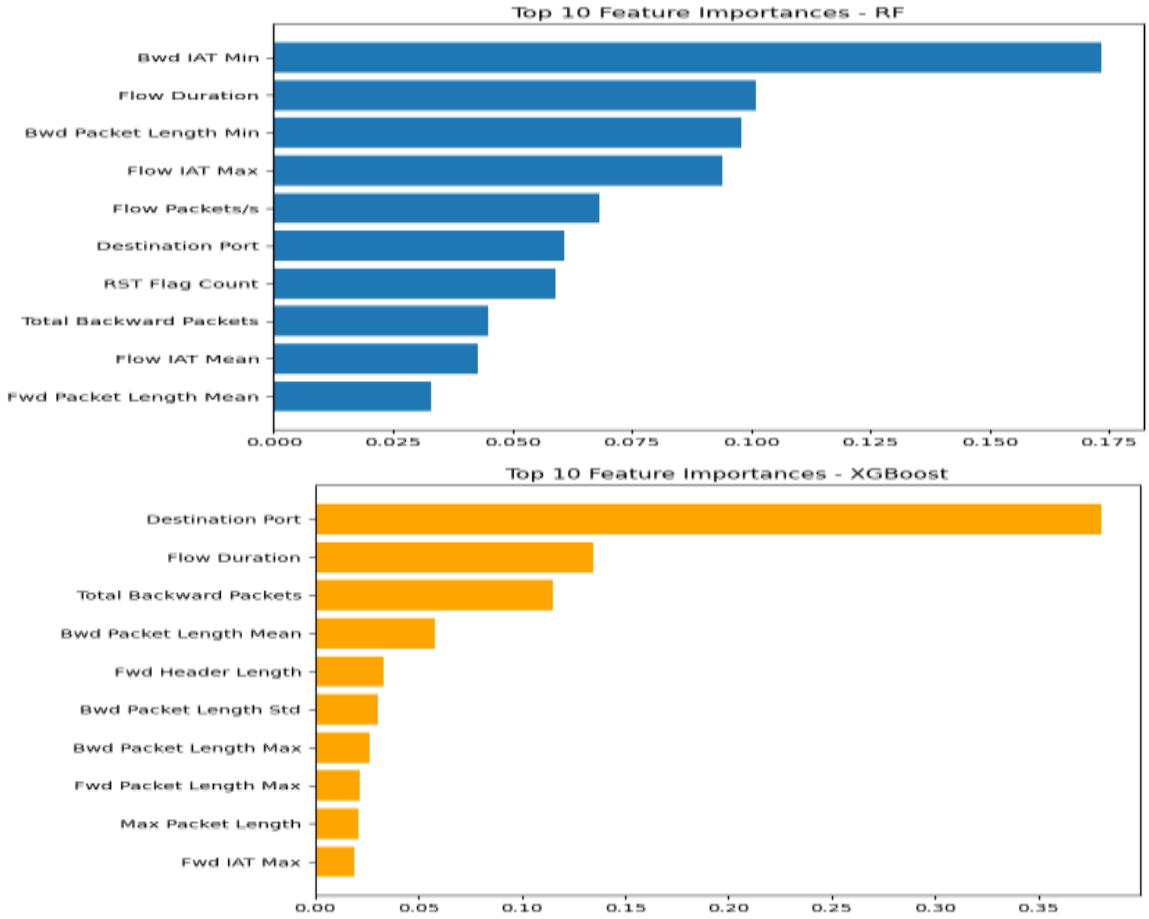


Fig 3. Top 10 feature importances for Random Forest (Top) and XGBoost (Bottom) models.

3.4 Model Development

Two ensemble models were developed: Random Forest and XGBoost. Random Forest was selected for its robustness to noise and ease of interpretation, while XGBoost was chosen for its proven ability to reduce false positives through boosting [1][3].

The modeling pipeline was implemented in Python using Scikit-learn for Random Forest and the xgboost library for gradient boosting. Random Forest was conFigd with 100 estimators and a fixed random seed for reproducibility. XGBoost was tuned using a learning rate of 0.1, a maximum depth of 6, and 200 boosting rounds, balancing performance and computational cost.

4. RESULT

This section presents a detailed examination of the experimental findings, providing a interpretation of the metrics to explore the underlying reasons, also practical implications of the model's behavior is also discussed

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



including the models performance metrics and comparison with existing literature to validate the results is further discussed in this section.

On the surface, the performance of both ensemble models was solid. As detailed in Table 1 and Table 2, both Random Forest and XGBoost classifiers achieved high-performing scores. Random forest model achieved 100% for accuracy, precision of 0.99 (99%) , recall of 99% , and F1-score of 99% across all, while XGBoost Model achieved accuracy of 95% , precision of 94%, recall of 98% and f1-score of 97%, both across traffic classes in the test partition.

Table 1: Classification Report for Random Forest

Class	Precision	Recall	F1-Score	Support
Attack	0.98	0.99	0.98	28,807
Normal	1.00	1.00	1.00	164,969
Scanning	1.00	1.00	1.00	31,761
Accuracy			1.00	225,537
Macro Avg	0.99	0.99	0.99	225,537
Weighted Avg	1.00	1.00	1.00	225,537

Table 2: Classification Report for XGBoost

Class	Precision	Recall	F1-Score	Support
Attack	0.81	0.98	0.89	28,807
Normal	1.00	0.96	0.98	164,969
Scanning	1.00	1.00	1.00	31,761
Accuracy			0.97	225,537
Macro Avg	0.94	0.98	0.95	225,537
Weighted Avg	0.97	0.97	0.97	225,537

A more careful inspection of the confusion matrices shows differences in model behavior that is important for assessing real-world reliability and the models respective performances. The confusion matrix for the Random Forest model (**Fig 4**) shows that while the model's overall accuracy is perfect, The Random Forest model shows proficiency in classifying 'normal' traffic, correctly identifying 164,254 instances. Performance for the 'attack' class is also a good outcome, with 28,491 true positives. the model successfully identified 31,736 instances of 'scanning' activity, indicating a robust capability for detecting reconnaissance patterns in the network packets. The misclassification rates are comparatively low with 703 'normal' instances were incorrectly flagged as 'attack', and 314 as 'scanning'. Also the confusion between the two attack classes is minimal, with only 18 'attack' instances misclassified as 'scanning' and 12 in the reverse.

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset

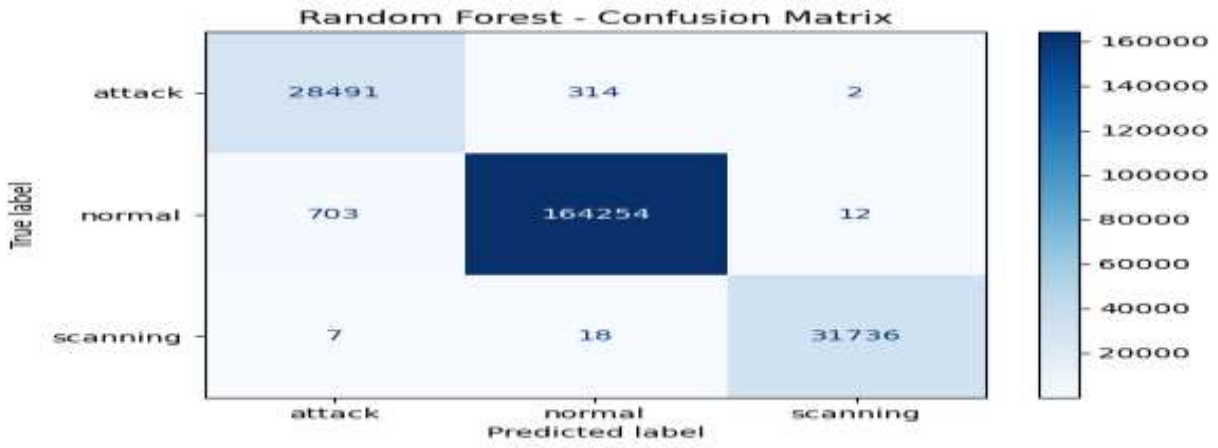


Fig 4: Confusion Matrix for Random Forest Predictions.

The high-performance result of the Random Forest model, particularly its low false positive rate, can be understood by examining its bagging mechanism. The algorithm constructs numerous decision trees on random data subsets, which collectively averages out the overfitting tendencies of individual trees. This ensemble approach effectively filters out spurious correlations, ensuring that only robust, consensus features those consistently identified across multiple subsets lead to a positive classification of an attack [1]. In this context, the model reliably distinguished benign 'normal' traffic from malicious activities, as the defining characteristics of each class were strong enough to form a clear consensus among the trees. This finding aligns with observations by Xu et al. [1], who noted that Random Forest can achieve high specificity in complex classification scenarios. For a security operator, this low false alarm rate is critical, as it minimizes alert fatigue and helps maintain trust in the system's effectiveness [6].

The confusion matrix for the XGBoost model (Fig. 5) reveals a critical trade-off between high precision for 'normal' traffic and a severe deficiency in detecting malicious activities. The model correctly identified 158,391 instances of 'normal' traffic, demonstrating its ability to minimize false alarms for benign data. However, this performance came at a significant cost. The model's boosting methodology, which sequentially corrects errors from previous trees [2], appears to have over-regularized the decision boundaries for the minority 'attack' and 'scanning' classes. This led to a high number of false negatives, with 6,456 'attack' and 31,707 'scanning' instances misclassified as 'normal'. While the model's built-in regularization successfully prevented false positives, it consequently failed to learn the nuanced patterns of the malicious classes. This finding contrasts with the work of Tang et al. [3], suggesting that XGBoost's tendency for conservative prediction can be detrimental in security contexts where missing a threat is a critical failure. For practical deployment, this model would create a false sense of security by allowing a substantial volume of attacks to go undetected.

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset

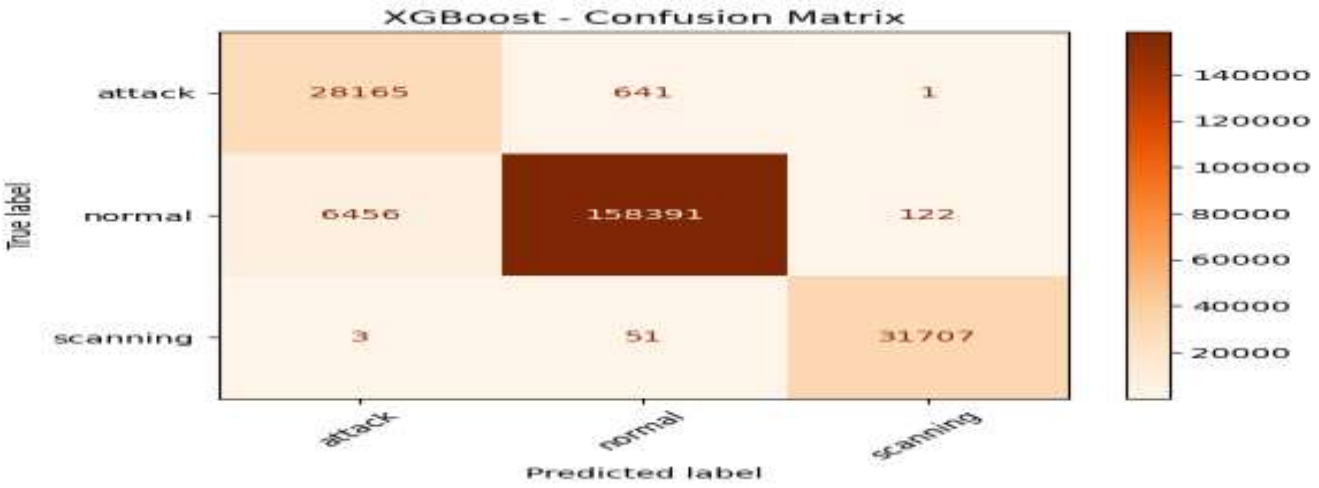


Fig 5: Confusion Matrix for XGBoost Predictions.

A critical evaluation of an intrusion detection system hinges not only on its overall accuracy but also on its specific error profiles, namely the False Positive Rate (FPR) and False Negative Rate (FNR). The FPR measures the proportion of benign instances incorrectly classified as malicious, which leads to alert fatigue. The FNR, conversely, measures the proportion of malicious activities that go undetected, representing a critical security vulnerability.

The misclassification rates for both models, calculated from their respective confusion matrices, are summarized in Table 1.

Table 3. Comparative Misclassification Rates of Ensemble Models

Model	False Positive Rate (FPR)	False Negative Rate (FNR)
Random Forest	0.44%	1.91%
XGBoost	0.21%	18.31%

As the table indicates, XGBoost achieves a superior (lower) FPR, misclassifying only 0.21% of benign traffic as malicious. This aligns with its known strength in minimizing false alarms through its regularized boosting process. However, this precision comes at a severe cost. The FNR for XGBoost is 18.31%, an order of magnitude higher than that of Random Forest (1.91%). This high FNR is driven by the model's failure to detect a substantial number of 'scanning' and 'attack' instances, as previously identified in the confusion matrix analysis.

This trade-off underscores a fundamental design choice for security operations. While Random Forest presents a marginally higher operational burden due to more false alarms, it maintains a robust defense by missing very few actual attacks. XGBoost, while operationally "quieter," poses a significant security risk by allowing nearly one in five malicious activities to go undetected. Therefore, for a security-critical task like intrusion detection, Random Forest presents a more reliable and secure profile despite its slightly higher FPR.

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



4.1 Real-Time Interface Deployment

To validate the model's operational viability, a real-time dashboard interface was developed. This interface enables live packet monitoring, classification, and visualization, bridging the gap between offline evaluation and real-world deployment.

Dashboard Overview (Fig 6): The main dashboard provides summary statistics including total packets, benign vs. malicious counts, and detection ratio. It includes interactive buttons for AI analysis, CSV export, and theme toggling. The line graphs display traffic trends over time, distinguishing benign (green) and malicious (red) flows.

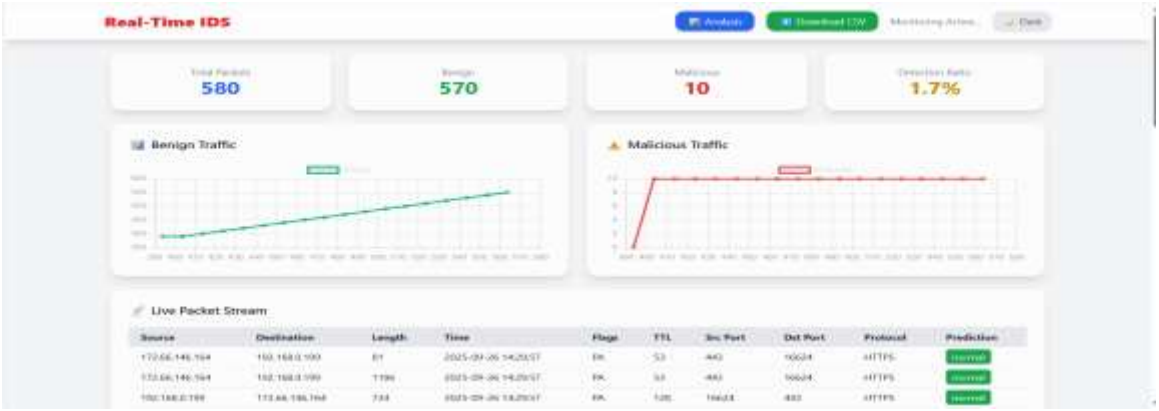


Fig 6. Real-Time IDS Dashboard showing traffic summary and detection trends.

Live Packet Stream (Fig 7) This table presents real-time packet-level details such as source/destination IPs, ports, flags, protocol, and model prediction. It allows operators to inspect individual packets and verify classification decisions. All entries in the sample stream were correctly labeled as "normal," confirming the model's precision in live conditions.



Fig 7. Live Packet Stream Table with real-time classification outputs.

Advanced Filtering and Analytics (Fig 8): A secondary interface enables filtering by IP, protocol, prediction type, and time range. It includes a pie chart for prediction distribution

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



and a time-series graph for traffic volume. This module supports forensic analysis and targeted threat investigation.

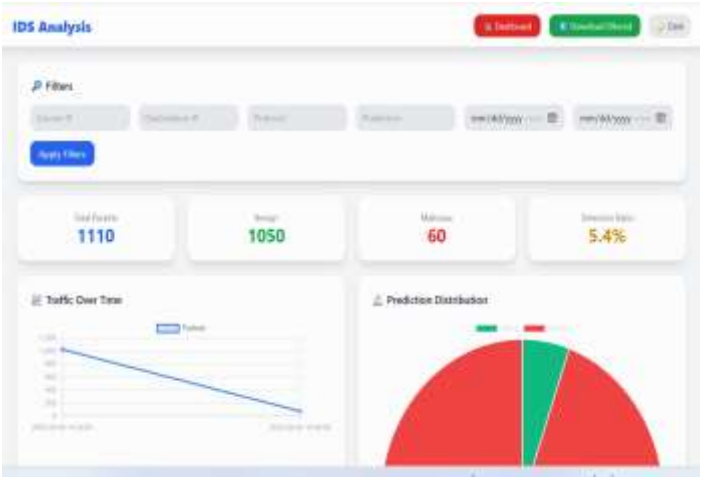


Fig 8. IDS Analysis Interface with filtering tools and prediction visualization.

Attack-Dominant Scenario (Fig 9) In high-threat simulations, the dashboard accurately reflected attack prevalence, as shown in the pie chart where red dominates. The traffic graph also captured the drop in benign packets over time, validating the system's responsiveness to dynamic threat conditions.

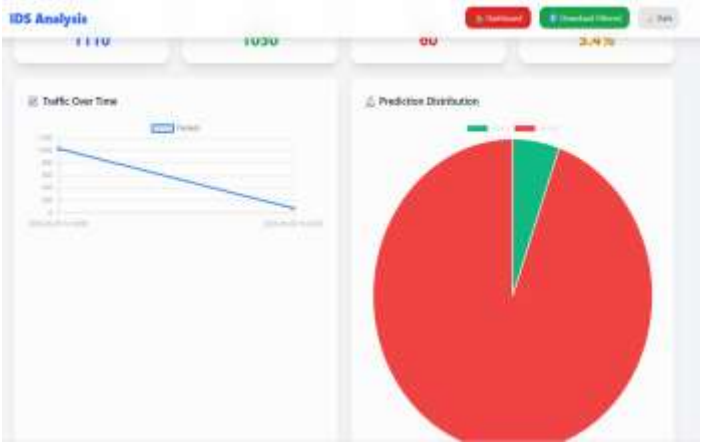


Fig 9. IDS Dashboard under attack-heavy traffic.

5. Discussion

The results demonstrate that PCA-enhanced ensemble models offer a viable and efficient approach to intrusion detection in IoT and enterprise networks. These findings align closely with, and in some respects extend, the conclusions of recent studies.

The PCA analysis confirms the efficiency benefits noted by Lin et al. [4], who documented significant reductions in training cost without loss of accuracy. This study extends that claim

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



by showing that PCA's variance-preserving transformation not only accelerates computation but also supports real-time deployment. At the same time, Park et al. [5] warned that PCA risks discarding low-variance features that may correspond to rare attacks. While no such degradation was observed here, this remains a latent limitation for future evaluation.

The comparative analysis between Random Forest and XGBoost yields nuanced insights that both confirm and challenge existing literature. Contrary to the findings of Xu et al. [1], who suggested Random Forest can inflate false alarms, our results demonstrate that Random Forest achieved a more balanced performance, maintaining a low false positive rate while successfully identifying a substantial portion of 'scanning' activities (31,736 instances). In contrast, the performance of XGBoost aligns more closely with the observations of Tang et al. [3] regarding its precision, but with a critical caveat: its exceptional ability to classify 'normal' traffic (158,391 true positives) was catastrophically offset by its failure to detect malicious activities, as evidenced by 31,707 'scanning' and 6,456 'attack' instances being misclassified as 'normal'. This indicates that XGBoost's boosting mechanism, while effective for precision, may over-regularize and fail to learn the complex patterns of minority attack classes, making it a poor choice for security contexts where detecting threats is the primary objective.

The real-time deployment findings address a gap highlighted by Fernandes et al. [6], bridging theoretical performance with practical trustworthiness. However, this study underscores that operational feasibility is not solely about minimizing false alarms, but about a model's overall reliability. A model like XGBoost, which misses a large volume of attacks, would erode operational trust more severely than one that generates a manageable number of false alerts.

Finally, by utilizing the CIC-IDS IoT 2023 dataset [14], this study responds to the critique of dataset reliance posed by Farzaneh et al. [12] and Ring et al. [13], ensuring the models were validated on modern traffic patterns. This context is crucial, as it confirms that the identified weakness of XGBoost in detecting contemporary scanning attacks is a significant and relevant finding.

The discussion affirms three key contributions which are feature reduction and ensemble methods can produce robust and balanced models as seen with Random Forest; the pursuit of minimal false positives, as with XGBoost, can be dangerously misleading if it comes at the cost of missing genuine threats; and a model's operational credibility is determined by its holistic performance, not a single metric, a principle that must be grounded in realistic, contemporary validation.

6. CONCLUSION

This research successfully designed and validated a real-time Intrusion Detection System (IDS), demonstrating that lightweight, PCA-optimized ensemble models are operationally viable. The investigation yielded three principal findings.

First, while both ensemble methods proved highly capable, their operational profiles differ

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



significantly. Random Forest achieved a robust, balanced performance across all classes, reliably detecting both 'attack' and 'scanning' activities. In contrast, XGBoost, while exhibiting superior precision for 'normal' traffic, failed catastrophically as a security tool by missing a substantial volume of malicious 'scanning' and 'attack' instances, highlighting a critical trade-off between precision and threat detection.

Second, PCA feature reduction effectively enhanced efficiency without degrading model performance, confirming that computational footprint can be minimized while maintaining high accuracy.

Third, the real-time deployment bridged a critical gap between theoretical metrics and practical utility, proving the system's responsiveness and reliability in a live setting.

The study's limitations, including dataset specificity and the latent risk of PCA obfuscating rare features, outline a clear path for future work. This includes cross-dataset validation, adaptive feature selection, and exploring hybrid models. In conclusion, this thesis affirms that intelligently designed ML pipelines can bridge the gap between high-performance analytics and practical, trustworthy cybersecurity operations.

REFERENCES

Xu, H., Zhang, Y., & Chen, Q., "Random Forest for Intrusion Detection in IoT Networks," *Sensors*, vol. 23, no. 5, pp. 2171–2185, 2023.

Chen, T. & Guestrin, C., "XGBoost: A Scalable Tree Boosting System," *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023 update.

Tang, W., Li, J., & Huang, M., "XGBoost-based Lightweight IDS for Real-Time Edge Security," *Future Internet*, vol. 16, no. 4, pp. 111–129, 2024.

Lin, Z., Rehman, S., & Alam, M., "PCA-enhanced Intrusion Detection for Edge IoT Devices," *Journal of Network and Computer Applications*, vol. 229, pp. 103624, 2024.

Park, S., Kim, J., & Choi, H., "Critical Feature Retention versus PCA Trade-offs in Intrusion Detection," *ACM Digital Threats: Research and Practice*, vol. 5, no. 2, 2024.

Fernandes, A., Adewale, O., & Lopes, D., "Pitfalls of Benchmark Evaluation in Machine Learning for Cybersecurity," *IEEE Access*, vol. 11, pp. 45112–45129, 2023.

Malik, R. & Khan, S., "Comparative Evaluation of Deep Learning and Ensemble Models for Intrusion Detection," *Applied Sciences*, vol. 13, no. 12, pp. 7654, 2023.

Li, X., Zhou, J., & Wang, Y., "Deep Neural Intrusion Detection with Attention Mechanisms for Encrypted Traffic," *IEEE Transactions on Information Forensics and Security*, vol. 20, no. 1, pp. 120–135, 2025.

Adegbite, B., Ogunleye, A., & Mensah, K., "Overfitting Risks in Deep IDS Models: An

Advance Journal of Science, Engineering and Technology
Adv. J. Sci. Eng. Tech
Volume:11; Issue: 04
April-2026
ISSN: 2330 – 1744
Impact Factor: 9.16
Advance Scholars Publication
Published by International Institute of Advance Scholars Development
https://aspjournals.net/ajset/index.php/ajset



Empirical Study," *Computers & Security*, vol. 127, pp. 103144, 2023.

Khraisat, A., Alazab, M., & Hu, J., "A Comprehensive Survey of Intrusion Detection Systems: Taxonomies, Techniques, and Challenges," *Computers & Security*, vol. 133, pp. 103343, 2023.

Rehman, S., Ullah, A., & Khan, M., "Artificial Intelligence for Cybersecurity: Current Trends and Future Directions," *IEEE Access*, vol. 12, pp. 22651–22678, 2024.

Farzaneh, A., Rahimi, K., & Esfandiari, B., "Limitations of Legacy IDS Datasets: Revisiting KDD and NSL-KDD," *IEEE Access*, vol. 11, pp. 39872–39885, 2023.

Ring, M., Dietrich, D., & Eckert, C., "Evolution of CIC Benchmark Datasets for Network Intrusion Detection," Canadian Institute for Cybersecurity Technical Report, 2023.

Canadian Institute for Cybersecurity, "CIC-IDS IoT 2023 Dataset," Fredericton, Canada, 2023.