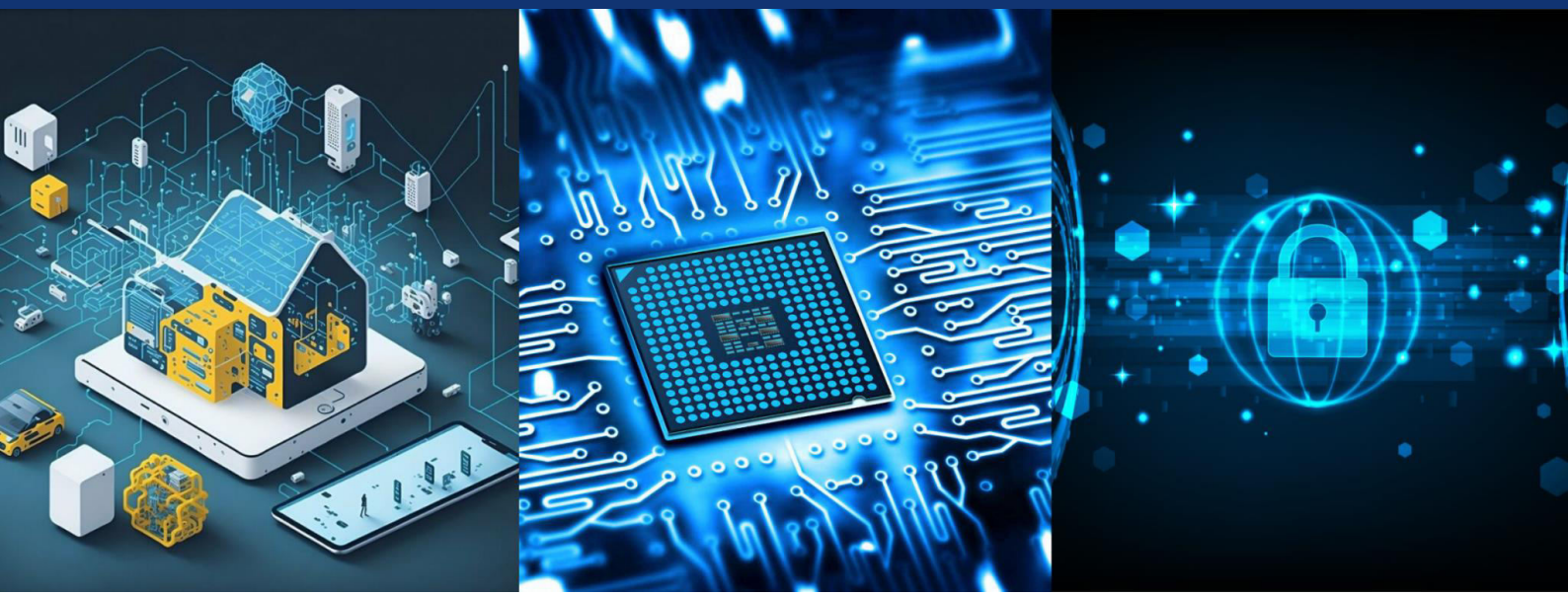
IJIRCCE

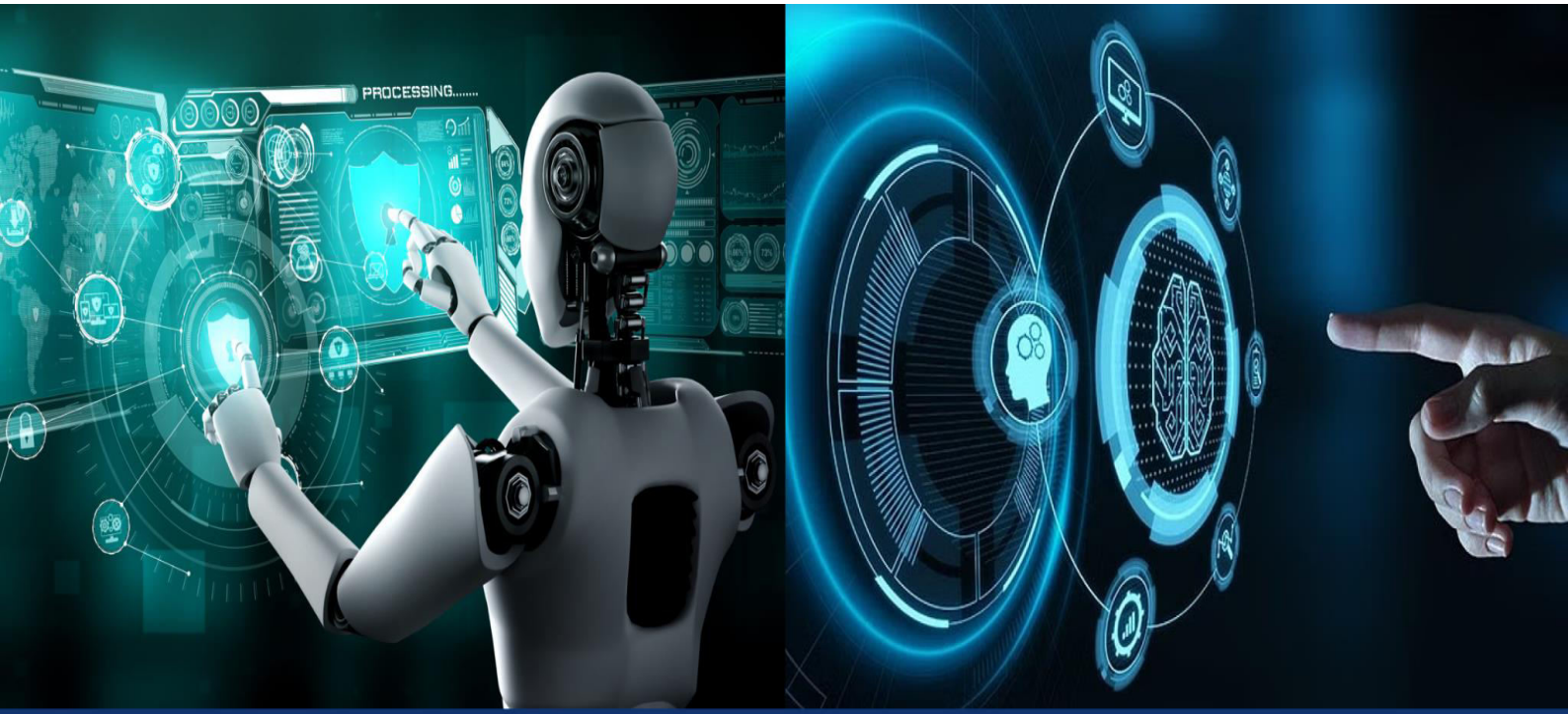
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 8, August 2026

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

XAI-driven False Positive Filtering in Anomaly-Based Intrusion Detection System

Mamidala Likhitha, Dr. M. Dhanalakshmi

Post Graduate Student, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India

Professor, Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Hyderabad, India

ABSTRACT: As dependence on online platforms grows for accessing and managing sensitive information, Intrusion Detection Systems (IDS) are now recognized as a crucial part of contemporary cybersecurity structures. The rapid advancement of machine learning, combined with access to extensive network traffic datasets, has greatly enhanced anomaly-based intrusion detection research; however, a persistent challenge in these systems is the frequent occurrence of incorrect alerts, which reduces the reliability and efficiency of security monitoring. This work introduces an enhanced method for detecting false positives by combining Explainable Artificial Intelligence (XAI) with conventional machine learning models. Rather than relying solely on prediction confidence, the proposed method analyses the relevance of input features contributing to each decision using SHAP and LIME, and trains a dedicated post-processing false positive detector on the extracted feature-importance patterns. The approach is evaluated on the LYCOS-IDS2017 dataset using Random Forest, K-Nearest Neighbors, Decision Tree, Naive Bayes, Neural Network, Voting Classifier and Stacking Classifier. Experimental results demonstrate that incorporating XAI-based feature relevance significantly improves the identification of false positives while maintaining a minimal reduction in true positives, outperforming conventional confidence-based filtering, which eliminates approximately 50% of false positives at the cost of a 0.42% loss in true positives.

KEYWORDS: Intrusion Detection System, Explainable Artificial Intelligence, Machine Learning, Network Security, Feature Importance, LYCOS-IDS2017 Dataset, Anomaly Detection, False Positive Reduction, SHAP, LIME.

I. INTRODUCTION

Intrusion Detection Systems (IDS) form a vital part of cybersecurity setups, created to observe network activity and identify unusual behaviour or breaches of security policy. IDS are generally divided into two categories: signature-based and anomaly-based. Signature-based systems recognize threats by matching traffic against established attack signatures, and while effective for familiar threats, they cannot detect new or zero-day exploits. Anomaly-based systems create a profile of typical network activity and flag deviations as possible intrusions; although this approach can uncover previously unseen attacks, it tends to generate a large number of false alarms, which lowers efficiency and places additional strain on security teams.

A false alarm happens when normal network behaviour is mistakenly flagged as harmful. This not only increases unnecessary alert generation but also reduces trust in the IDS, leading to alert fatigue among security analysts; in real-world scenarios, excessive false alarms can cause critical alerts to be ignored, potentially allowing real attacks to go unnoticed. Traditional machine learning-based IDS mostly depend on prediction confidence scores to assess whether an instance is malicious, but confidence by itself is not always enough to differentiate between accurate and inaccurate forecasts.

This paper proposes an improved method that combines machine learning models with Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME. The central idea is that the model's decision-making process can be better understood by examining feature relevance or importance values, so that patterns linked to false positives can be identified by evaluating the contribution of each feature to a prediction. This encourages the creation of a post-processing system that eliminates erroneous detections while preserving genuine security risks, moving anomaly-based IDS away

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

from a black-box design toward a system that is intelligent, comprehensible and reliable enough for practical cybersecurity deployment.

II. RELATED WORK

Early foundational work on intrusion detection evaluation was carried out using the DARPA/Lincoln Laboratory Intrusion Detection Evaluation framework, which provided one of the first standardized environments for benchmarking IDS performance and laid the groundwork for modern IDS research [1]. Building on this, several studies examined alert quality and post-processing: Spathoulas and Katsikas investigated methods for reducing false positives in IDS and emphasized that excessive false alarms degrade system trust and operational efficiency [2], while Kim et al. proposed a cost-effective mechanism for evaluating AI prediction reliability using explainable artificial intelligence and statistical analysis, introducing a reliability indicator that helps analysts prioritise which alerts require immediate attention [3].

The limitations of conventional machine learning-based IDS have also been widely studied. Sommer and Paxson argued that many existing approaches rely on unrealistic assumptions about data distributions and training environments, causing poor generalisation in open-world conditions [4]. Nisioti et al. surveyed unsupervised intrusion detection and attacker-attribution methods, highlighting the importance of feature engineering for deeper forensic analysis [5], while Viegas et al. showed that careful dataset design and feature selection are critical for effective real-world IDS deployment [6]. Ring et al. surveyed network-based intrusion detection datasets and identified significant inconsistencies that directly impact model reliability, reinforcing the importance of using well-structured datasets such as LYCOS-IDS2017 [7].

Recent advances in explainable AI have significantly influenced intrusion detection research. Lundberg and Lee introduced SHAP, a unified, game-theoretic framework for feature attribution that has become a standard for measuring each feature's contribution to a model's prediction [8], building on Shapley's foundational work on cooperative game theory [9]. Ribeiro et al. proposed LIME, a model-agnostic explanation technique that enables local interpretability of individual predictions, making complex models easier to trust in security applications [10]. Marino et al. demonstrated how interpretability can be incorporated into deep learning-based security models to increase resilience against adversarial attacks [11], while Shrikumar et al. and Bach et al. introduced complementary feature-attribution techniques (DeepLIFT and layer-wise relevance propagation) for neural network interpretability [12], [13].

False positive reduction remains a central focus of IDS research. Pitre et al. proposed an IDS specifically designed to reduce false positives in zero-day attack scenarios, showing that careful model design and feature selection can significantly improve detection precision [14], and Alshammari et al. introduced a neuro-fuzzy method to lower false-positive notifications, demonstrating that hybrid intelligent systems can improve accuracy while minimising unnecessary alarms [15]. Existing approaches, however, either overlook the relevance of individual attributes in the detection process or do not explicitly combine confidence scores with feature-importance analysis, which can leave false alarms inadequately filtered. The proposed work addresses this gap by directly evaluating XAI-derived feature importance to train a dedicated false-positive detector.

III. PROPOSED ALGORITHM

A. System Architecture:

Fig. 1 illustrates a three-step framework for detecting false positives (FP) in an anomaly-based IDS using explainable AI. The pipeline is fed with positive samples from the dataset together with the anomaly-based IDS outputs. In Step 1 (Threshold Setup), predictions are categorized: high-confidence instances above a defined threshold are classified strictly as True Positives (TP), while low-confidence instances below the threshold represent a mixture of False Positives and True Positives (TP + FP). Step 2 receives this low-confidence pool and applies SHAP and LIME to extract feature-importance values, saving them into an XAI-attributes database. In Step 3 (Model Building), the newly extracted XAI attributes train a dedicated machine learning FP detector, which evaluates incoming low-confidence traffic to differentiate genuine security breaches from benign anomalies, significantly reducing alarm fatigue.

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

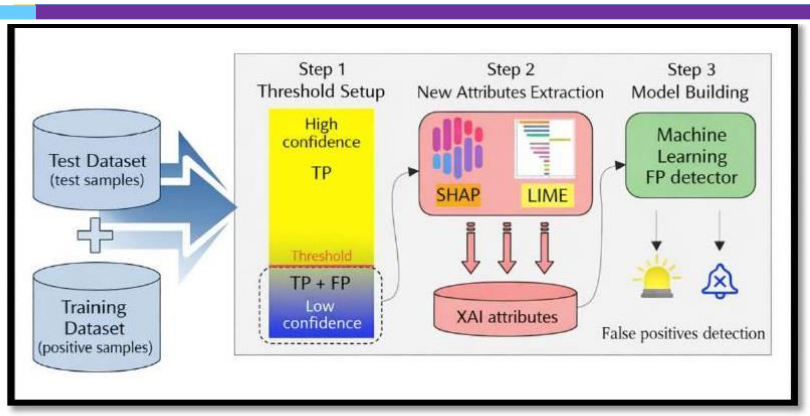


Fig. 1. Proposed System Architecture

B. Design Considerations:

- The false-positive detection model is trained and evaluated on the LYCOS-IDS2017 dataset, which offers a wide variety of benign and malicious network traffic patterns.
- Data pre-processing removes missing values, irrelevant attributes and inconsistent records, and converts categorical features into numerical form.
- Multiple machine learning classifiers - Random Forest, KNN, Decision Tree, Naive Bayes and Neural Network - are trained for baseline classification performance.
- Ensemble learning methods, namely a Voting Classifier (Random Forest + AdaBoost) and a Stacking Classifier (Random Forest, MLP and LightGBM), are used to increase detection accuracy and resilience.
- SHAP and LIME are used to compute feature-importance values for low-confidence predictions, which are stored as XAI attributes.
- A dedicated post-processing FP detector is trained on the XAI attributes to distinguish genuine attacks from benign anomalies.
- The system is deployed through a Flask web interface offering real-time prediction and interactive SHAP/LIME explanations.

C. Description of the Proposed Algorithm:

The suggested technique aims to precisely identify false positives in anomaly-based intrusion detection while preserving genuine attack detections. There are five main steps in the proposed algorithm.

Step 1: Data Pre-processing.

The LYCOS-IDS2017 dataset is loaded and pre-processed; missing values, redundant records and irrelevant features are removed, categorical attributes are encoded, and the data is normalised. Let D = {x1, x2, x3, ..., xn} represent the pre-processed network traffic dataset consisting of n traffic records.

Step 2: Model Training and Confidence Scoring.

Relevant traffic features are used to train the classification models. Each trained model M produces a predicted class and an associated confidence score:

Confidence = P(y = attack | x) eq. (1)

Instances with confidence above a defined threshold T are marked as True Positives; the remaining low-confidence instances form the candidate pool for XAI-based analysis:

Candidate Pool = {x : Confidence(x) < T} eq. (2)

Step 3: XAI-Based Feature Attribution.

For every instance in the candidate pool, SHAP and LIME are applied to compute the contribution of each input feature to the model's decision:

Importance(x) = {SHAP(x), LIME(x)} eq. (3)

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Step 4: False Positive Detector Training.
The extracted feature-importance vectors, together with the original confidence scores, are used as input features to train a dedicated machine learning FP detector F:
FP Status = F(Importance(x), Confidence(x)) eq. (4)

Step 5: Alert Refinement.
If FP Status = 1, the alert is suppressed (muted notification); if FP Status = 0, the alert is retained as a genuine attack (active alarm). This adaptive filtering stage reduces alarm fatigue while maintaining a high true-positive detection rate

IV. PSEUDOCODE

Step 1: Load the LYCOS-IDS2017 dataset containing benign and attack network traffic records.
Step 2: Perform data pre-processing - remove missing/duplicate values, eliminate irrelevant features, encode categorical attributes and normalise the data.
Step 3: Split the dataset into input features (X) and target labels (Y), then into training and testing subsets.
Step 4: Train Random Forest, KNN, Decision Tree, Naive Bayes, Neural Network, Voting Classifier and Stacking Classifier models.
Step 5: Compute prediction confidence for each test instance.
Step 6: IF Confidence >= Threshold THEN classify as True Positive; ELSE add instance to the low-confidence candidate pool.
Step 7: Apply SHAP and LIME on the candidate pool to compute feature-importance values for each instance.
Step 8: Store the extracted feature-importance values in the XAI-attributes database.
Step 9: Train the false-positive detector using the XAI attributes and confidence scores.
Step 10: Pass each low-confidence instance through the trained FP detector.
Step 11: IF FP Status = 1 THEN suppress the alert (mute notification); ELSE raise the alert (active alarm).
Step 12: Compute performance metrics - Accuracy, Precision, Recall and F1-score.
Step 13: Display predictions and SHAP/LIME explanations on the Flask web interface.
Step 14: Continue monitoring incoming network traffic for new predictions.
Step 15: End.

V. EXPERIMENTAL RESULTS

A. Table and Graphs:
The proposed framework was implemented in Python using the LYCOS-IDS2017 dataset and evaluated with Accuracy, Precision, Recall and F1-score. Table 1 summarises the performance of all evaluated classifiers.

ML Model	Accuracy	F1-Score	Recall	Precision
Random Forest	0.871	0.868	0.871	0.897
KNN	0.871	0.868	0.871	0.897
Decision Tree	0.871	0.868	0.871	0.897
Naive Bayes	0.578	0.605	0.578	0.857
Neural Network	0.871	0.868	0.871	0.897
Stacking Classifier	0.871	0.868	0.871	0.897
Voting Classifier	1.000	1.000	1.000	1.000

Table 1. Assessment of Performance

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

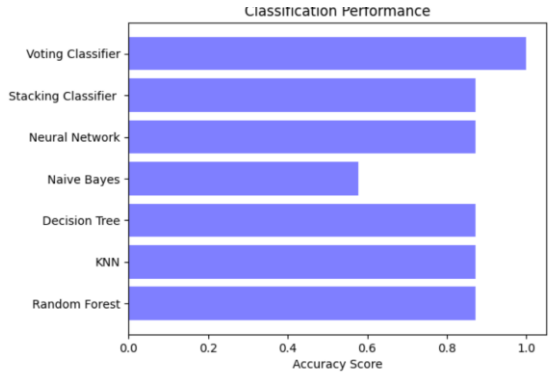


Fig. 2(a). Accuracy Comparison Graph

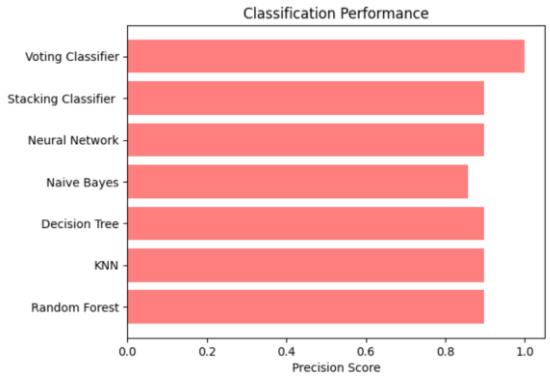


Fig. 2(b). Precision Comparison Graph

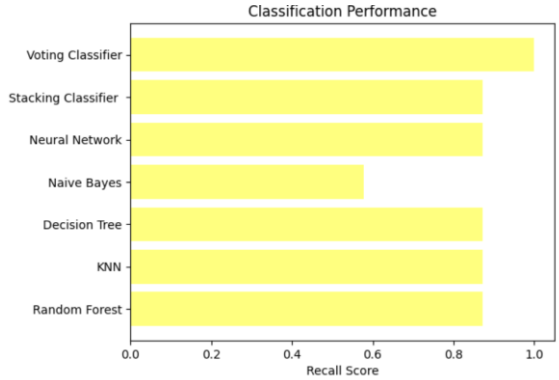


Fig. 3(a). Recall Comparison Graph

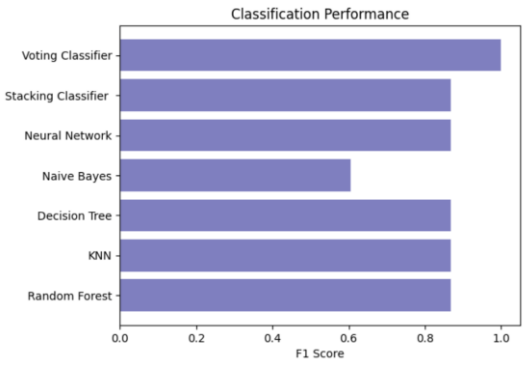


Fig. 3(b). F1-Score Comparison Graph

Across all seven models, the Voting Classifier (Random Forest + AdaBoost) achieved the highest scores (Accuracy = Precision = Recall = F1 = 1.000), followed closely by Random Forest, KNN, Decision Tree, Neural Network and the Stacking Classifier, each with an accuracy of 0.871. Naive Bayes recorded the lowest accuracy (0.578), reflecting its independence assumption being a poor fit for the correlated network-traffic features in LYCOS-IDS2017. These results confirm that ensemble methods combined with XAI-driven feature relevance provide the most reliable false-positive filtering.

B. Metrics for Evaluation:

Let TP, TN, FP and FN denote True Positives, True Negatives, False Positives and False Negatives, respectively. The evaluation metrics are defined as:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad \text{eq. (5)}$$

$$\text{Precision} = TP / (TP + FP) \quad \text{eq. (6)}$$

$$\text{Recall} = TP / (TP + FN) \quad \text{eq. (7)}$$

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad \text{eq. (8)}$$

Compared with conventional confidence-based filtering, which removes approximately 50% of false positives at a 0.42% loss in true positives, the proposed XAI-driven post-processing stage achieves a more favourable balance between false-positive reduction and true-positive retention, while SHAP and LIME explanations additionally give security analysts a transparent, per-prediction justification for every retained or suppressed alert.

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. System Screenshots and Explainability Visualizations:

The proposed framework was deployed through a Flask web interface providing real-time prediction and interactive SHAP/LIME explanations. The following screenshots illustrate the front-end workflow and the explainability visualizations generated for individual predictions.

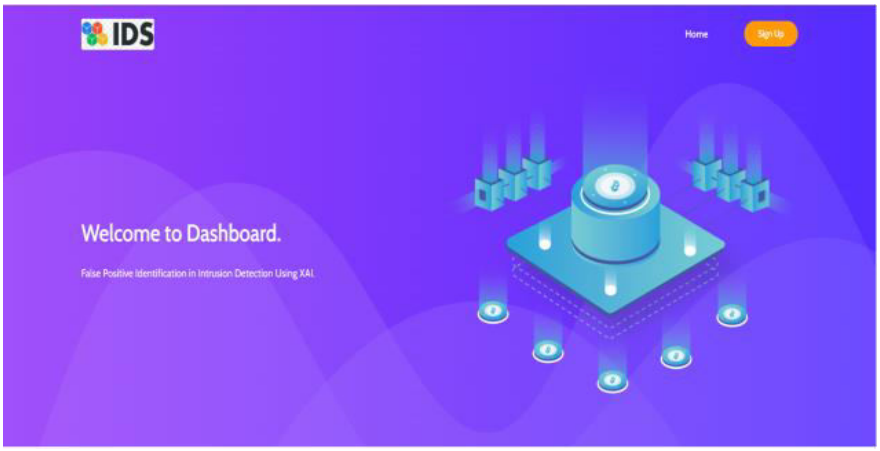


Fig. 4. Home Page

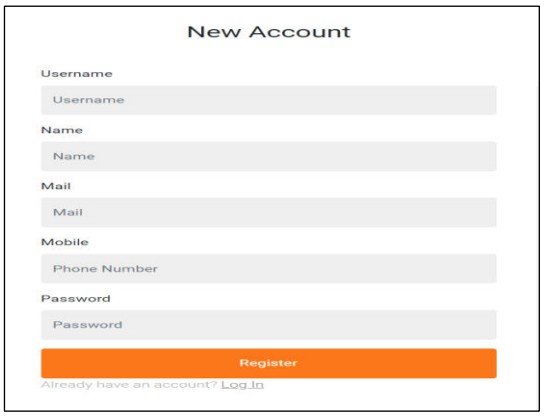


Fig. 5. Signup Page

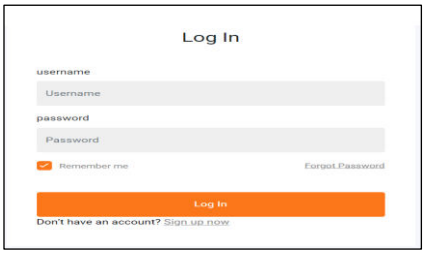


Fig. 6. Login Page

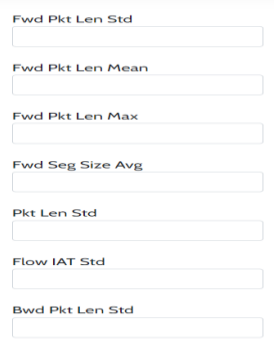


Fig. 7(a). Upload Input

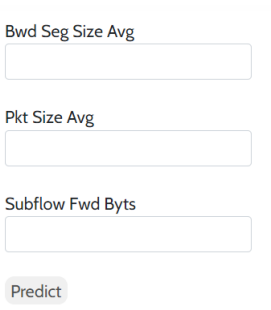


Fig. 7(b). Upload Input (continued)

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Fig. 7(c). Predict Result (DDOS)

Fig. 7(d). Predict Result (Normal)

Fig. 8(a). LIME Interactive Explanation for DDoS Traffic Detection

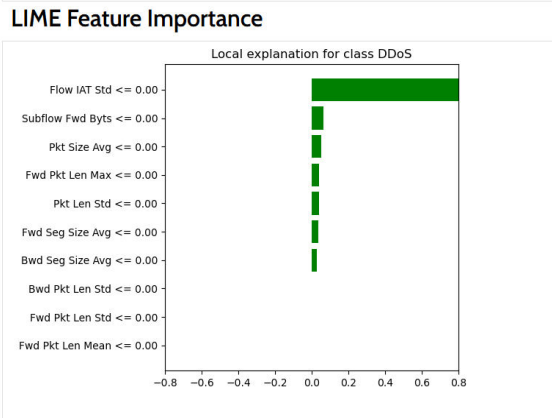


Fig. 8(b). LIME Interactive Explanation for DDoS Traffic Detection (feature view)

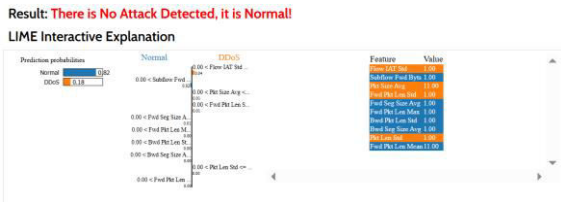


Fig. 8(c). LIME Interactive Explanation for Normal Traffic Detection

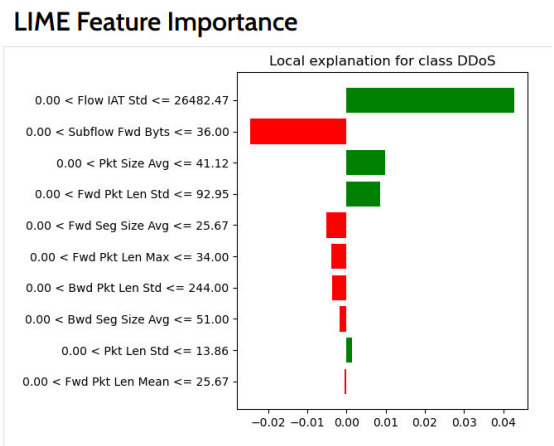


Fig. 8(d). LIME Feature Importance for Normal Classification

© 2026 IJIRCCE | Volume 14, Issue 8, August 2026 DOI: 10.15680/IJIRCCE.2026.1408010

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

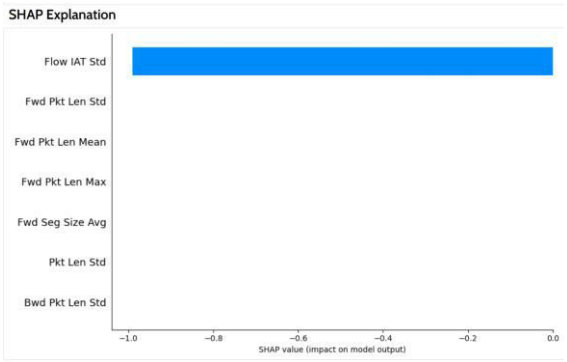


Fig. 9. SHAP Explanation for DDoS Traffic Detection

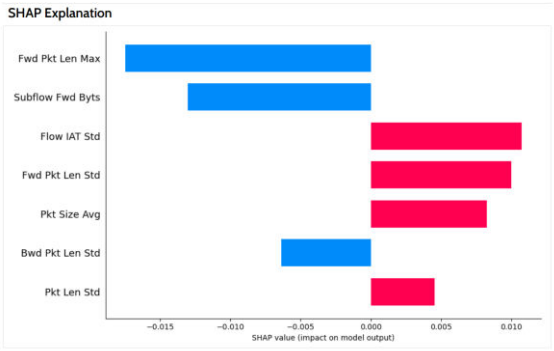


Fig. 10. SHAP Explanation for Normal Traffic Detection

These visualizations demonstrate how the system exposes per-prediction feature attributions to security analysts, allowing both DDoS and normal-traffic classifications to be inspected and trusted before an alert is retained or suppressed.

VI. CONCLUSION AND FUTURE SCOPE

The growing intricacy of cyber threats necessitates sophisticated and flexible security systems that can accurately detect illicit activity. Although intrusion detection systems are essential for protecting network environments, they frequently have a high false-positive rate, which lowers operational effectiveness and adds to the workload of security analysts. The proposed framework combines explainable AI with machine learning and ensemble classifiers, evaluated on the LYCOS-IDS2017 dataset, to gain deeper insight into model decision-making and more effectively discriminate between true and false alerts. Experimental results show that ensemble methods provide improved stability and performance, and that incorporating SHAP/LIME-based feature relevance strengthens trust in automated decisions while filtering unreliable predictions without significantly affecting true-positive detection rates.

Future work will focus on extending the framework to real-time and streaming network environments with minimal latency, integrating live enterprise or cloud data sources, and applying online learning techniques so the system can adapt to changing attack patterns. Using sophisticated deep learning architectures such as transformer-based models and graph neural networks may further improve detection accuracy for stealthy attacks, while enhanced SHAP-based global and local interpretability analysis can provide deeper insight into model behaviour and further improve analyst trust.

REFERENCES

[1] Lippmann, R., et al., “The 1999 DARPA Off-Line Intrusion Detection Evaluation,” Computer Networks, vol. 34, no. 4, pp. 579–595, 2000.
[2] Spathoulas, G. P., and S. K. Katsikas, “Reducing False Positives in Intrusion Detection Systems,” Computers & Security, vol. 29, no. 1, pp. 35–44, 2010.
[3] Kim, D., et al., “A Cost-Effective Method for Evaluating AI Prediction Reliability in Intrusion Detection Using Explainable AI,” IEEE Access, 2022.
[4] Sommer, R., and V. Paxson, “Outside the Closed World: On Using Machine Learning for Network Intrusion Detection,” IEEE Symposium on Security and Privacy, pp. 305–316, 2010.
[5] Nisioti, A., A. Mylonas, P. D. Yoo, and V. Katos, “From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods,” IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 3369–3388, 2018.
[6] Viegas, E., A. Santin, and L. Oliveira, “Toward a Reliable Anomaly-Based Intrusion Detection in Real-World Environments,” Computer Networks, vol. 127, pp. 200–216, 2017.
[7] Ring, M., S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, “A Survey of Network-Based Intrusion Detection Data Sets,” Computers & Security, vol. 86, pp. 147–167, 2019.
[8] Lundberg, S. M., and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[9] Shapley, L. S., "A Value for n-Person Games," Contributions to the Theory of Games, vol. 2, no. 28, pp. 307–317, 1953.
[10] Ribeiro, M. T., S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, 2016.
[11] Marino, D. L., C. S. Wickramasinghe, and M. Manic, "An Adversarial Approach for Explainable AI in Intrusion Detection Systems," Proceedings of IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society, pp. 3237–3243, 2018.
[12] Shrikumar, A., P. Greenside, A. Shcherbina, and A. Kundaje, "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences," arXiv:1605.01713, 2016.
[13] Bach, S., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," PLOS ONE, vol. 10, no. 7, 2015.
[14] Pitre, A., K. Gandhi, V. Konde, P. Adhao, and V. Pachghare, "A Machine Learning Based Intrusion Detection System for Zero-Day Attack Detection with Reduced False Positives," International Journal of Computer Applications, 2021.
[15] Alshammari, A., A. Sonamthiang, M. Teimouri, and D. Riordan, "Using Neuro-Fuzzy Approach to Reduce False Positive Alerts," Proceedings of the 5th Annual Conference on Communication Networks and Services Research, pp. 345–349, 2007.
[16] Breiman, L., "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details