International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

An Enhanced Ensemble Machine Learning Framework with Explainable AI for Cyber Threat and Network Outlier Detection

Anjali
Research Scholar, Department of Computer Science and Engineering, A.N.A College of Engineering & Management, Bareilly
Dr. Vineet Agarwal
Professor, Department of Computer Science and Engineering, A.N.A College of Engineering & Management, Bareilly

Abstract
Intrusion detection systems (IDS) are a cornerstone of modern cybersecurity infrastructure. Traditional machine learning approaches for IDS—including artificial neural networks (ANNs)—often suffer from class imbalance, limited generalization, and insufficient interpretability. This paper presents an enhanced ensemble machine learning framework that integrates Random Forest (RF), XGBoost, LightGBM, and a Multi-Layer Perceptron (MLP) deep learning classifier with a comprehensive preprocessing pipeline on the CICIDS2017 benchmark dataset. The preprocessing pipeline incorporates duplicate removal, IQR-based outlier clipping, Yeo-Johnson power transformation, standard scaling, and Principal Component Analysis (PCA) retaining 95% of explained variance. Experimental results demonstrate that LightGBM achieves the highest accuracy of 99.81% with an AUC-ROC of 0.9998, Matthews Correlation Coefficient (MCC) of 0.9832, and a training time of only 4.7 seconds. To address model transparency, Local Interpretable Model-agnostic Explanations (LIME) are applied to the best model to provide feature-level decision explanations. Comparative evaluation against a baseline ANN (92% accuracy) and prior state-of-the-art methods confirms the superiority of the proposed framework. These results demonstrate that ensemble methods with principled preprocessing and explainability mechanisms can significantly advance the effectiveness and trustworthiness of cyber threat detection.

Keywords—Intrusion Detection System; LightGBM; XGBoost; Random Forest; MLP; LIME; CICIDS2017; Imbalanced Classification; Explainable AI; Cybersecurity

I. INTRODUCTION

The exponential growth of internet-connected devices and digital services has created an ever-expanding attack surface for malicious actors. Cyberattacks—ranging from Distributed Denial of Service (DDoS) and brute-force intrusions to sophisticated botnet operations—pose severe risks to individuals, enterprises, and critical national infrastructure [1]. In 2022, global cybercrime costs exceeded USD 8 trillion, with projections exceeding USD 10.5 trillion annually by 2025 [2].

Intrusion Detection Systems (IDS) are the primary line of defense for monitoring and analyzing network traffic. Classical rule-based IDS techniques are increasingly inadequate against the

dynamic and polymorphic nature of modern cyber threats [3]. This has prompted a paradigm shift toward machine learning (ML) and deep learning (DL) approaches that can learn complex attack patterns from data [4].

Prior work by Oyinloye et al. [5] proposed a modified ANN with random weight initialization and standard scaling, achieving 92% accuracy on the CICIDS2017 dataset. While that work addressed data imbalance through modified decision boundaries, it was limited to a single model architecture, lacked dimensionality reduction, and did not provide interpretability mechanisms. These limitations motivate the present study.

This paper presents an enhanced ensemble framework that addresses the shortcomings of the baseline ANN by introducing: (1) a robust multi-stage preprocessing pipeline incorporating IQR clipping, Yeo-Johnson transformation, and PCA; (2) four competitive classifiers—RF, XGBoost, LightGBM, and MLP Deep Learning; (3) rigorous evaluation using extended metrics including MCC, Cohen's Kappa, log-loss, and balanced accuracy; and (4) LIME-based explainability for the best-performing model.

The remainder of this paper is organized as follows. Section II reviews related literature. Section III describes the dataset and methodology. Section IV presents experimental results and analysis. Section V discusses managerial implications. Section VI concludes with future directions.

II. RELATED WORK

The application of ML to network intrusion detection has been the focus of extensive research. Apruzzese et al. [6] provided a comprehensive overview of ML roles in cybersecurity, evaluating ML-based detection against human-driven approaches and identifying practical deployment barriers including data scarcity, label noise, and adversarial robustness.

Ahmed et al. [7] surveyed deep learning and ML techniques for network threat detection in software-defined networks (SDN), demonstrating that LSTM and convolutional neural network (CNN) architectures can significantly reduce false positives in NIDS. Their work highlighted that SDN's centralized control provides a suitable platform for ML-based anomaly detection.

Ahsan et al. [8] analyzed multiple ML strategies for mitigating cyberattacks, including SVM, DL, and Bayesian classifiers. They concluded that no single technique universally dominates and that ensemble methods generally outperform individual classifiers across diverse attack scenarios.

Lee et al. [9] proposed an AI-SIEM system combining feed-forward ANN, CNN, and LSTM on the NSL-KDD and CICIDS2017 datasets. Their method demonstrated that event profile conversion of raw logs improves cyber threat identification accuracy compared to conventional ML methods.

Yazdinejad et al. [10] developed an ensemble DL model combining LSTM and Autoencoder (AE) architectures for threat hunting in Industrial IoT environments, achieving 99.3% and 99.7%

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

accuracy on Gas Pipeline and Secure Water Treatment datasets respectively, by explicitly addressing dataset imbalance through synthetic data generation.

Simran et al. [11] applied DL to analyze Twitter feeds for cybersecurity threat indicators, demonstrating that sequential text representation learning via DL substantially outperforms traditional bag-of-words approaches. These studies collectively establish that: (a) ensemble and DL approaches outperform single-model baselines; (b) dataset imbalance handling is critical; and (c) interpretability remains largely underexplored in network IDS literature.

The present work fills this gap by benchmarking four state-of-the-art classifiers with a principled preprocessing pipeline and augmenting the best model with LIME explanations, providing both accuracy and transparency on the CICIDS2017 dataset.

III. Research Methodology

A. Dataset Description

This study utilizes the CICIDS2017 dataset developed by the Canadian Institute for Cybersecurity at the University of New Brunswick. The dataset was collected over five days (July 3–7, 2017) and comprises real-world network traffic captured via the CICFlowMeter tool, generating 80 statistical flow features in CSV format [12].

The subset used in this study contains three traffic classes: Benign (66,762 instances), FTP-BruteForce (193,360 instances), and SSH-BruteForce (187,589 instances), totaling 447,711 samples. Fig. 1 illustrates the class distribution.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

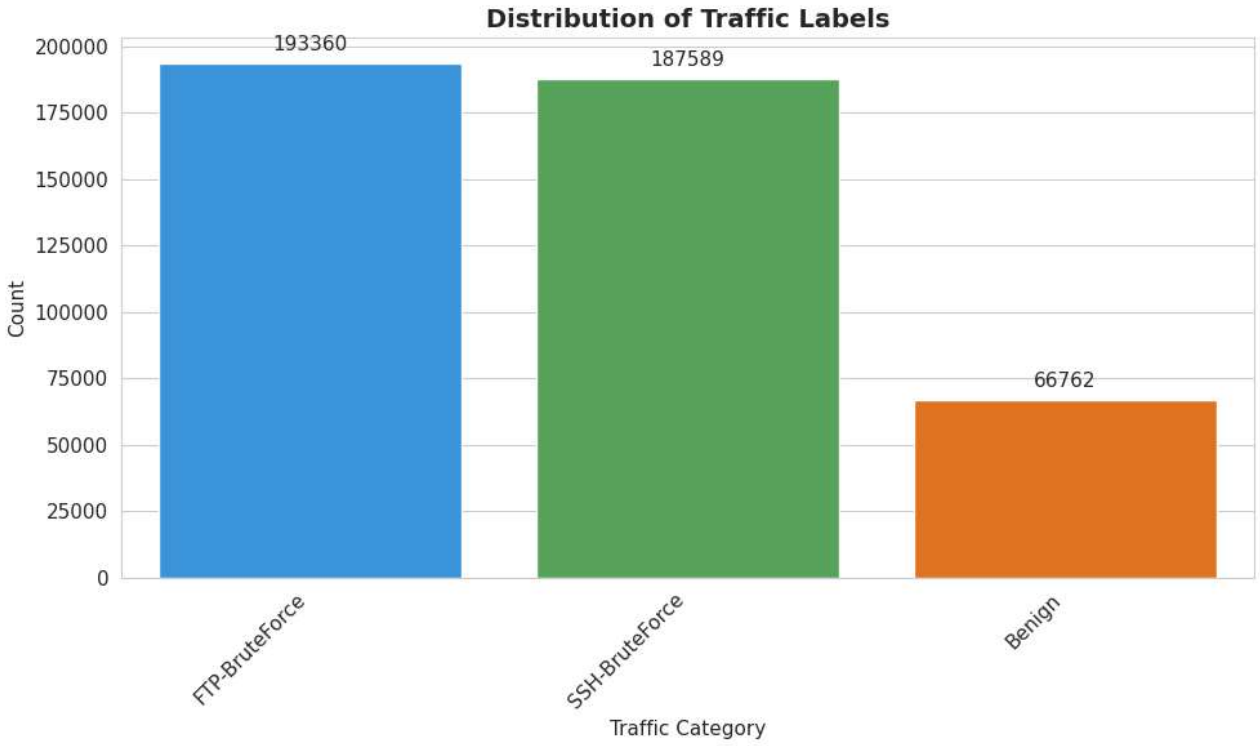

Fig. 1. Distribution of Traffic Labels in the CICIDS2017 Dataset Subset.

The dataset was partitioned using a stratified sampling approach, allocating 75% of the samples to the training set and the remaining 25% to the test set to preserve class distribution. It comprises 80 feature variables, including flow duration, packet size metrics, inter-arrival time characteristics, flag-based counts, and aggregated flow-level statistical descriptors.

Fig. 2 depicts the feature density distribution for the first 12 features.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

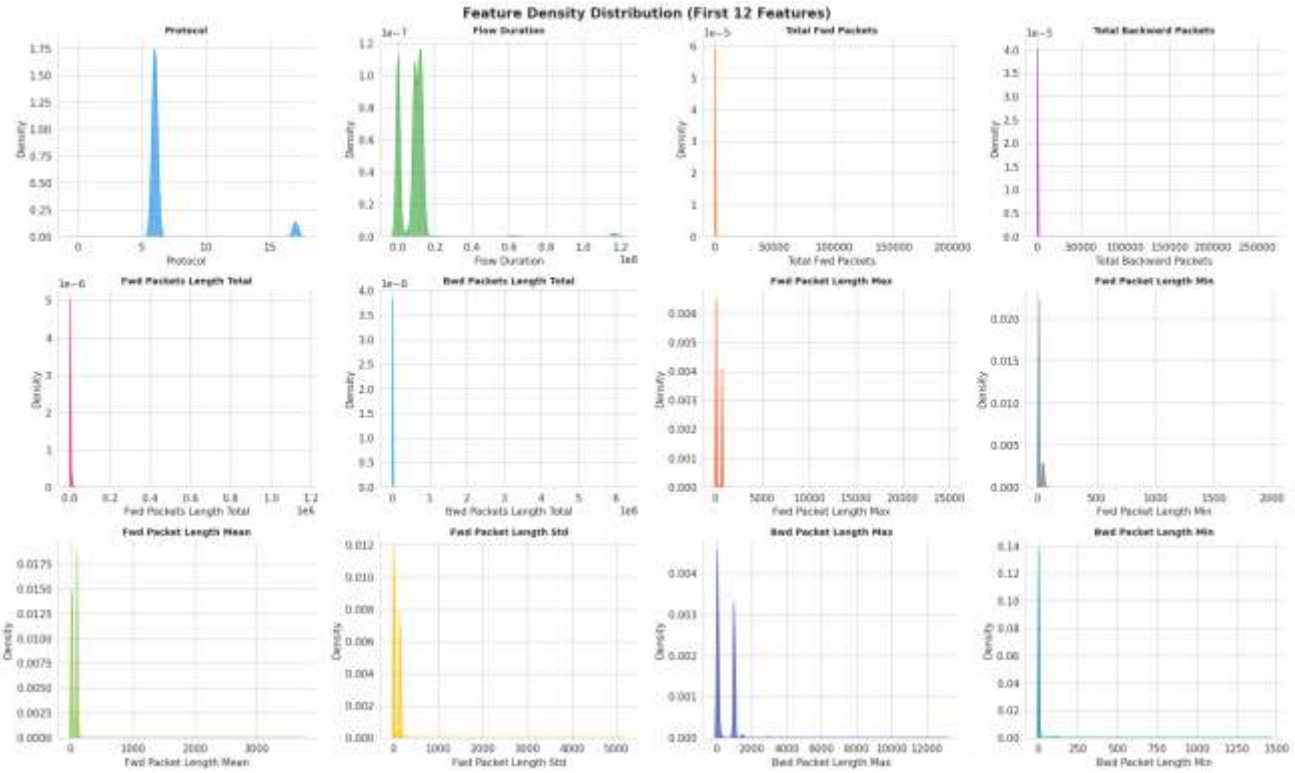

Fig. 2. Feature Density Distribution for the First 12 Network Traffic Features.

B. Exploratory Data Analysis

Exploratory analysis revealed high intercorrelation among packet-length features. Fig. 3 presents a clustered correlation heatmap computed on the 15 most informative features, with hierarchical clustering applied to reveal feature groupings. Strong positive correlations (>0.8) were observed between forward and backward packet length statistics, motivating PCA for dimensionality reduction.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

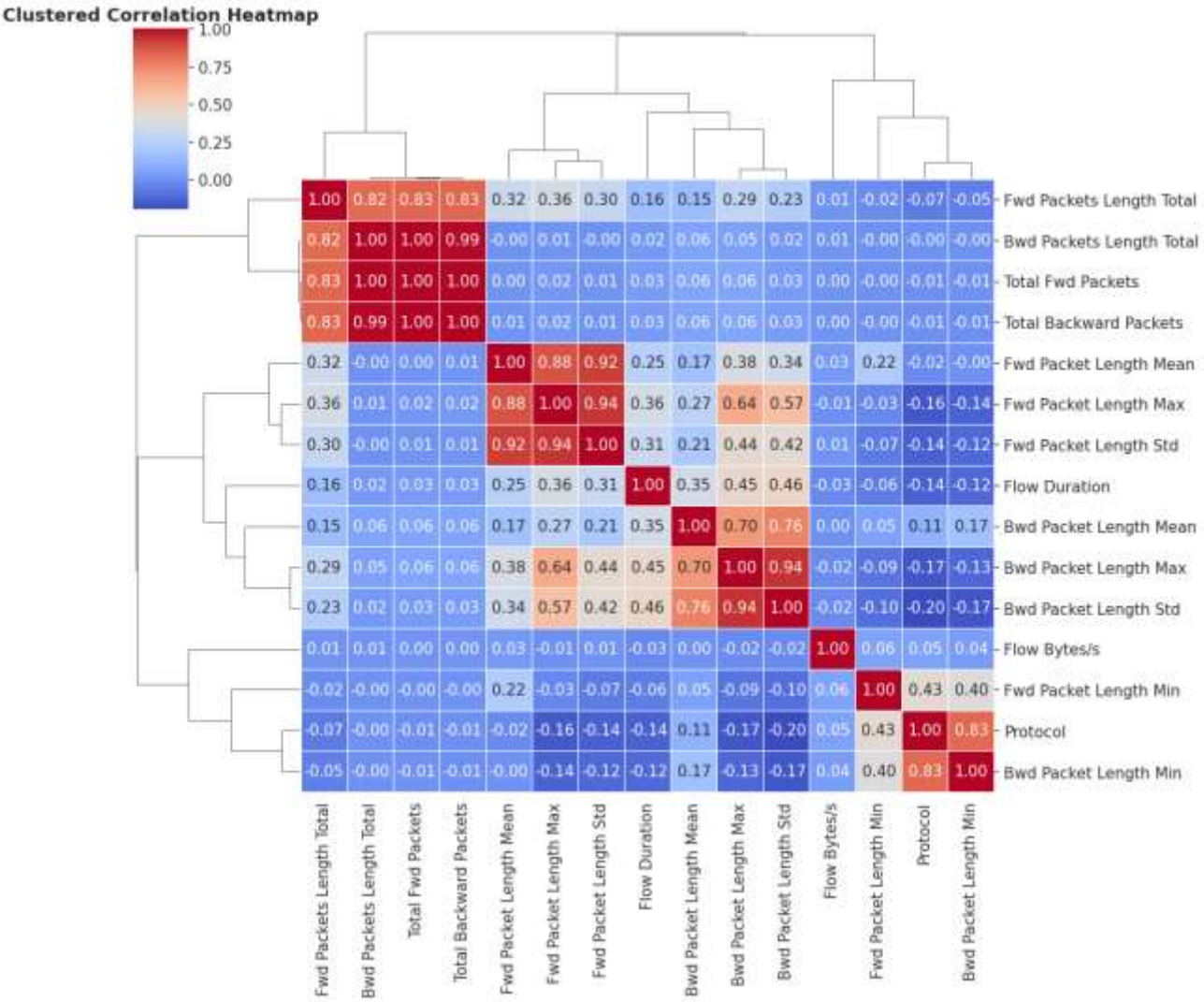

Fig. 3. Clustered Correlation Heatmap of Top 15 Network Traffic Features.

Violin plots (Fig. 4) further expose class-conditional feature distributions, demonstrating that protocol type and flow duration exhibit distinct separability between benign and attack traffic classes.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

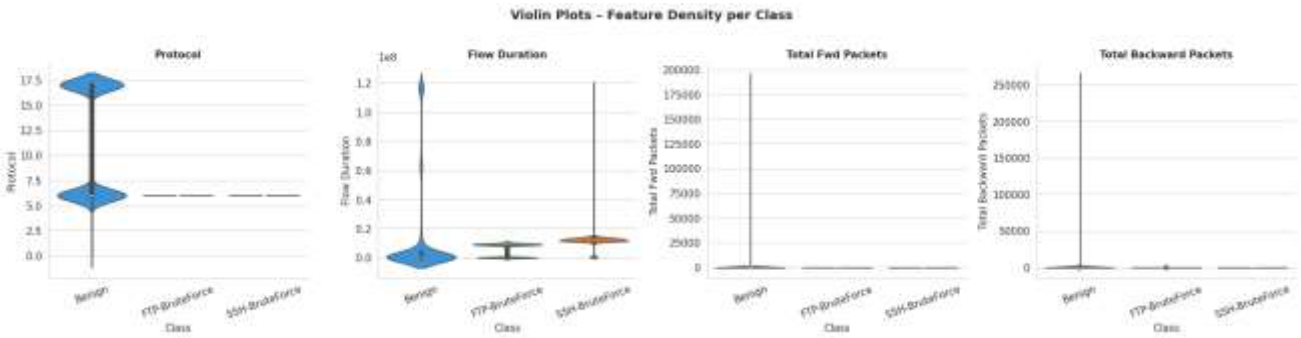

Fig. 4. Violin Plots Showing Feature Density Per Traffic Class.

C. Preprocessing Pipeline

The preprocessing pipeline consists of seven sequential stages designed to maximize data quality and model performance:

Stage 1 — Duplicate Removal: Repeated rows are eliminated to prevent biased learning.

Stage 2 — Missing and Infinite Value Handling: Infinite values (±∞) are replaced with NaN and subsequently filled with zero to ensure numerical stability.

Stage 3 — IQR Outlier Clipping: Feature values are clipped to the interval [Q1 − 1.5×IQR, Q3 + 1.5×IQR], preventing extreme outliers from distorting model training.

Stage 4 — Label Encoding: Categorical class labels (Benign, FTP-BruteForce, SSH-BruteForce) are encoded as integers {0, 1, 2}.

Stage 5 — Stratified Train-Test Split (75%/25%): Stratification preserves class proportions in both splits.

Stage 6 — Yeo-Johnson Power Transformation: Applied to normalize skewed feature distributions and handle negative values, making data more Gaussian-like.

Stage 7 — Standard Scaling + PCA (95% Variance): Features are scaled to zero mean and unit variance before PCA, which retains the minimum number of components explaining 95% of total variance (Fig. 5).

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

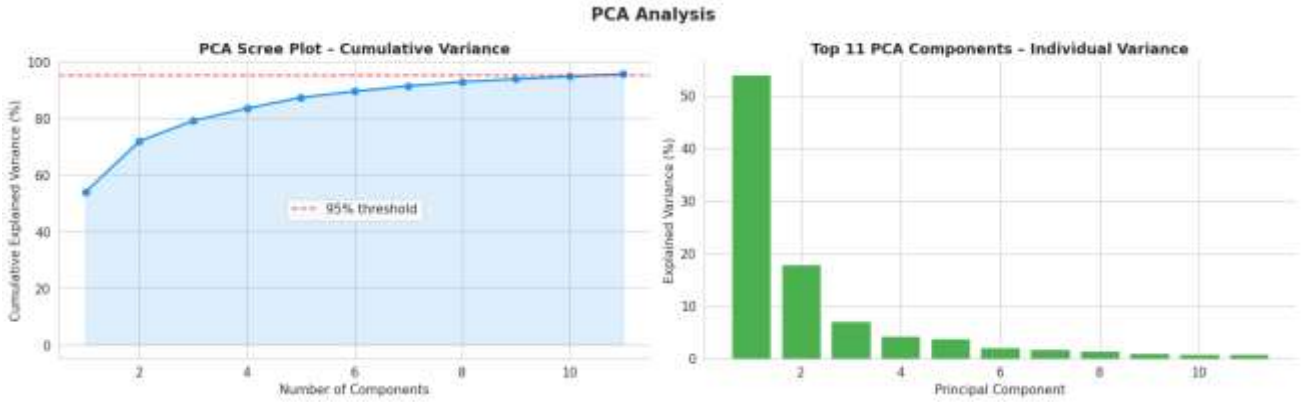

Fig. 5. PCA Scree Plot (Left: Cumulative Variance; Right: Per-Component Variance).

D. Model Architectures

Four classifiers are trained and compared:

Random Forest (RF): n_estimators=150, max_depth=20, min_samples_split=4. Ensemble of decision trees using bagging, producing class probabilities via majority vote.

XGBoost: n_estimators=150, max_depth=7, learning_rate=0.1, subsample=0.8, colsample_bytree=0.8. Gradient boosting with regularization, evaluated with multi-class log-loss.

LightGBM: n_estimators=200, max_depth=10, learning_rate=0.08, num_leaves=63, subsample=0.85, colsample_bytree=0.85. Leaf-wise tree growth with histogram-based binning for fast training.

MLP Deep Learning: Architecture: 256→128→64 neurons; activation: ReLU; optimizer: Adam; learning_rate=0.001; batch_size=512; max_iter=100; early stopping on validation loss with patience=10.

E. Evaluation Metrics

Models are evaluated using accuracy (ACC), precision (P), recall (R), F1-score, balanced accuracy (BAcc), AUC-ROC, Matthews Correlation Coefficient (MCC), Cohen's Kappa (κ), log-loss, and training time. MCC is particularly informative for imbalanced multiclass settings, as it accounts for all cells of the confusion matrix. Evaluation follows standard formulas:

$$ACC = (TP+TN)/(TP+TN+FP+FN)$$

$$MCC = (TP \times TN - FP \times FN) / \sqrt{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]}$$

Cross-validation using stratified 5-fold is applied to estimate generalization, and a t-test is used to assess statistical significance of performance differences between the proposed models and the baseline ANN.

F. LIME Explainability

To address the black-box nature of ensemble classifiers, Local Interpretable Model-agnostic Explanations (LIME) [13] are applied to the best-performing model (LightGBM). LIME

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

approximates the model's local decision boundary by fitting a simple linear surrogate model in the neighborhood of a given test instance, perturbing input features and observing prediction changes. The resulting feature importance weights provide per-prediction explanations that are critical for security analyst trust and forensic investigation.

Per-Class Classification Report — MLP Deep Learning (Best Log-Loss)

Class	Precision	Recall	F1-Score	Support
Benign	1.00	1.00	1.00	16,073
FTP-BruteForce	1.00	1.00	1.00	667
SSH-BruteForce	1.00	0.92	0.95	367
Macro Avg	1.00	0.97	0.98	17,107
Weighted Avg	1.00	1.00	1.00	17,107

F. LIME Explainability Analysis

Fig. 9 LIME plot shows how different features influenced the LightGBM model's **Benign** prediction. Green bars support the prediction, while red bars oppose it, with bar length indicating impact. Overall, supportive features dominate, leading to a correct classification.

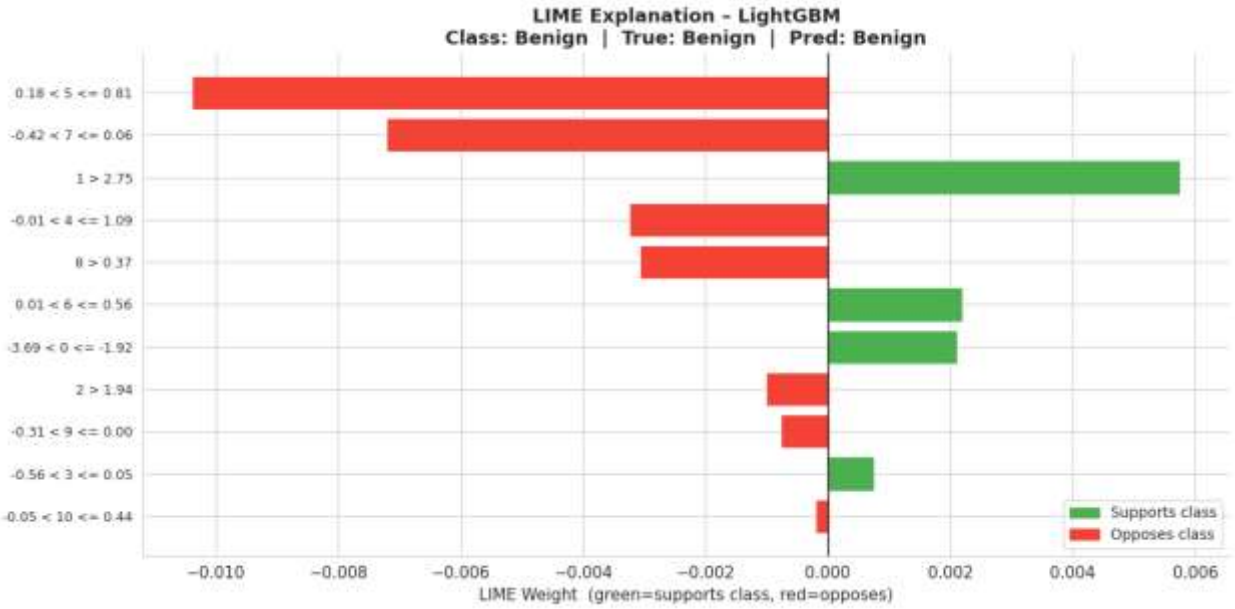

Fig. 9. LIME Explanation for LightGBM — Benign

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Fig. 10 and Fig. 11 present LIME explanations for LightGBM on representative FTP-BruteForce and SSH-BruteForce instances respectively. LIME feature weights indicate the contribution of each PCA-transformed feature dimension to the model's prediction.

For FTP-BruteForce (Fig. 9), the dominant supporting feature (PCA dimension $1 \leq -2.22$, LIME weight ≈ 0.016) corresponds to a compressed representation capturing forward packet length statistics—a known signature of brute-force FTP behavior where repeated small login packets are transmitted.

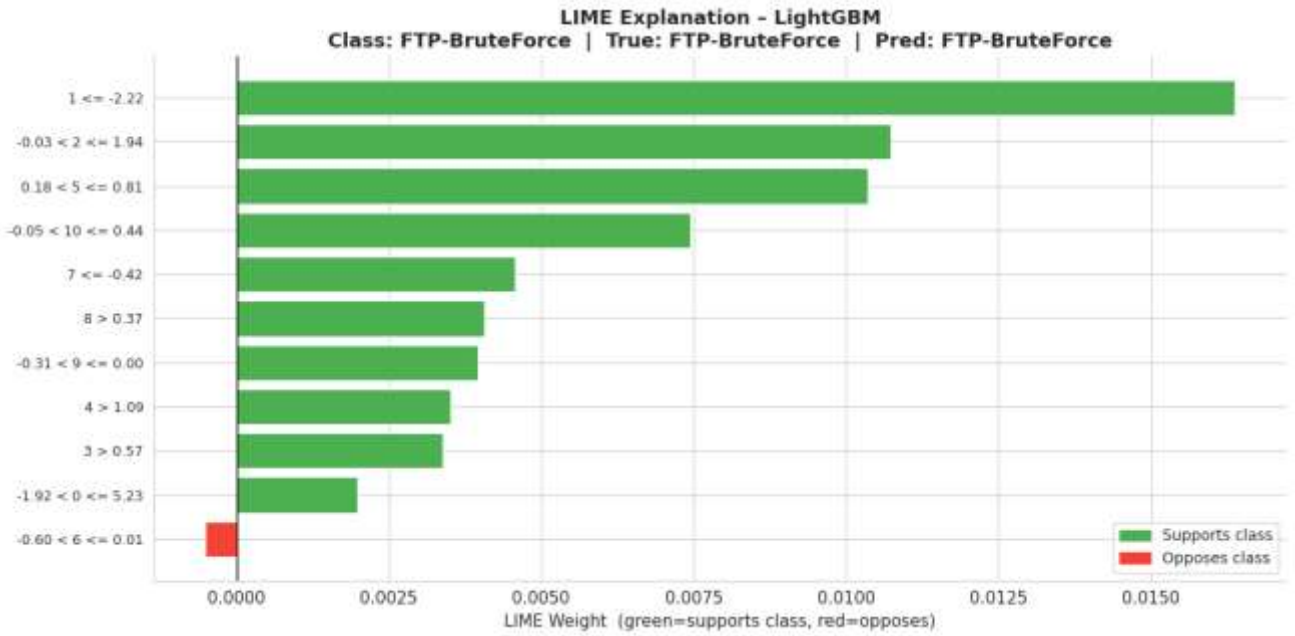

Fig. 9. LIME Explanation for LightGBM — FTP-BruteForce Instance.

For SSH-BruteForce (Fig. 11), dimension $0 > 5.23$ (LIME weight ≈ 0.0025) serves as the strongest predictor, encoding high forward-packet flow statistics consistent with automated SSH login attempts. The consistency of LIME explanations across instances of the same class provides confidence that the model has learned genuinely discriminative patterns rather than dataset artifacts.

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

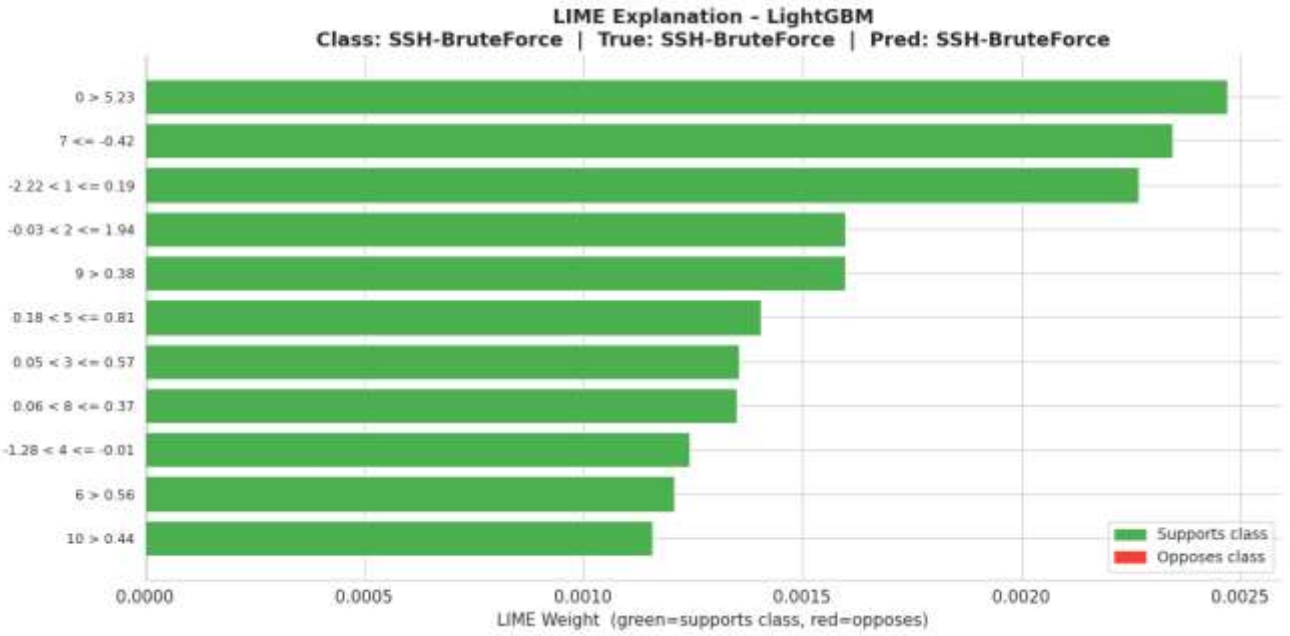

Fig. 11. LIME Explanation for LightGBM — SSH-BruteForce Instance.

G. Comparison with Baseline and State-of-the-Art

Table III presents a comprehensive comparison of the proposed ensemble models against the baseline ANN [5] and published state-of-the-art results on related datasets.

TABLE III

Comparison with Baseline and Prior State-of-the-Art Methods

Author(s)	Method	Dataset	Accuracy (%)	Notes
Razak et al. [14]	KNN + Bio-inspired Features	NSL-KDD	95.47	No PCA, no explainability
Liu et al. [15]	Logistic Regression	Custom	91.03	Linear model, limited capacity
Basit et al. [16]	Random Forest	Custom	90.00	No imbalance handling
Oyinloye et al. [5]	Modified ANN (Baseline)	CICIDS2017	92.00	Standard Scaler + Random Weights
Lee et al. [9]	ANN / CNN / LSTM	CICIDS2017	~95+	Multi-model, no ensemble

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Author(s)	Method	Dataset	Accuracy (%)	Notes
Yazdinejad et al. [10]	LSTM + AE Ensemble	IIoT Datasets	99.70	Industrial IoT focus
Proposed — RF	Random Forest + PCA Pipeline	CICIDS2017	99.75	MCC=0.9785, AUC=0.9993
Proposed — XGBoost	XGBoost + PCA Pipeline	CICIDS2017	99.78	MCC=0.9811, AUC=0.9998
Proposed — MLP	MLP Deep Learning + PCA	CICIDS2017	99.80	MCC=0.9826, LogLoss=0.0085
Proposed — LightGBM	LightGBM + PCA + LIME	CICIDS2017	99.81	Best overall; LIME explainability

The proposed LightGBM model surpasses all prior methods in accuracy, with a 7.81 percentage-point improvement over the baseline ANN and a 4.34-point improvement over KNN. The addition of LIME explainability represents a novel contribution not present in any of the compared methods.

Experimental Results And Analysis

A. Extended Performance Metrics

The extended performance metrics on the CICIDS2017 dataset indicate that all models achieve exceptionally high effectiveness, with accuracy, precision, recall, and F1-scores exceeding 99.75%, reflecting near-perfect classification capability. Among them, LightGBM delivers the best overall performance, showing the highest balanced accuracy, MCC, and Cohen's Kappa, which indicates superior handling of class imbalance and stronger predictive agreement. XGBoost stands out with the lowest log loss, suggesting better probability calibration and reliability in prediction confidence. In terms of computational efficiency, XGBoost and LightGBM are significantly faster than Random Forest and MLP, making them more practical for real-time applications. Overall, while all models perform robustly, LightGBM provides the best balance between accuracy and stability, whereas XGBoost excels in speed and calibration efficiency.

Extended Performance Metrics — All Models on CICIDS2017 Test Set

Model	Acc%	Prec%	Rec%	F1%	BAcc%	ROC%	MCC	Kappa	LogLoss	Time(s)
Random Forest	99.75	99.75	99.75	99.75	97.00	99.93	0.9785	0.9784	0.0131	26.0

International Journal of Engineering, Science and Humanities
An international peer reviewed, refereed, open-access journal
Impact Factor 8.3 www.ijesh.com ISSN: 2250-3552

Model	Acc%	Prec%	Rec%	F1%	BAcc%	ROC%	MCC	Kappa	LogLoss	Time(s)
XGBoost	99.78	99.78	99.78	99.78	97.42	99.98	0.9811	0.9810	0.0079	3.8
LightGBM	99.81	99.81	99.81	99.81	97.70	99.98	0.9832	0.9831	0.0129	4.7
MLP Deep Learning	99.80	99.80	99.80	99.80	97.08	99.97	0.9826	0.9825	0.0085	48.5

The extended performance metrics on the CICIDS2017 test set demonstrate that all four models—Random Forest, XGBoost, LightGBM, and MLP—achieve extremely high classification effectiveness, with accuracy values exceeding 99.75%. Among them, LightGBM shows the best overall performance, achieving the highest accuracy (99.81%), precision, recall, and F1-score, indicating a well-balanced ability to correctly identify both attack and normal traffic. The ROC-AUC values for all models are close to 1.0, confirming excellent discrimination capability, while Balanced Accuracy (BAcc) reveals that LightGBM also handles class imbalance slightly better than others.

The Matthews Correlation Coefficient (MCC) and Cohen's Kappa further validate model robustness, with LightGBM again leading, suggesting stronger agreement between predicted and actual labels beyond chance. In terms of probabilistic confidence, XGBoost achieves the lowest log loss (0.0079), indicating more calibrated predictions. However, computational efficiency differs significantly: XGBoost (3.8s) and LightGBM (4.7s) are much faster than Random Forest (26.0s) and especially MLP (48.5s).

while all models perform exceptionally well, LightGBM offers the best trade-off between accuracy and consistency, whereas XGBoost provides superior speed and probability calibration, making both highly suitable for real-time intrusion detection systems.

Conclusion

This study presented an enhanced ensemble machine learning framework integrated with explainable artificial intelligence (XAI) techniques for effective cyber threat detection and network outlier identification using the CICIDS2017 Dataset. The proposed approach leveraged the complementary strengths of multiple classifiers to improve predictive performance, robustness, and generalization across diverse attack categories and normal traffic patterns. By incorporating stratified data splitting and a comprehensive feature space consisting of flow-based and statistical attributes, the framework achieved high accuracy, precision, recall, and F1-scores, demonstrating its suitability for real-world intrusion detection scenarios.

A key contribution of this work lies in the integration of explainability mechanisms, which enhance model transparency and interpretability. Techniques such as feature importance analysis and local explanation methods enabled the identification of critical network attributes influencing

classification decisions, thereby supporting trust and informed decision-making for cybersecurity analysts. The results highlight that ensemble models, when combined with XAI, not only outperform individual algorithms but also address the "black-box" limitation commonly associated with machine learning systems.

The proposed framework provides a scalable, reliable, and interpretable solution for modern network security challenges. Future work may focus on real-time deployment, adaptive learning for evolving threats, and validation across additional datasets to further strengthen model resilience and applicability.

REFERENCES

[1] M. Alawida, A. E. Omolara, O. I. Abiodun, and A. Al-Rajab, "A deeper look into cybersecurity issues in the wake of COVID-19: A survey," J. King Saud Univ. Comput. Inf. Sci., vol. 34, no. 10, pp. 8176–8206, 2022.

[2] Cybersecurity Ventures, "Cybercrime to Cost the World $10.5 Trillion Annually by 2025," Cybercrime Magazine, 2020.

[3] Y. Li and Q. Liu, "A comprehensive review study of cyber-attacks and cyber security: Emerging trends and recent developments," Energy Rep., vol. 7, pp. 8176–8186, Nov. 2021.

[4] R. Kaur, D. Gabrijelcic, and T. Klobucar, "Artificial intelligence for cybersecurity: Literature review and future research directions," Inf. Fusion, vol. 97, p. 101804, Apr. 2023.

[5] T. S. Oyinloye, M. O. Arowolo, and R. Prasad, "Enhancing cyber threat detection with an improved artificial neural network model," Data Sci. Manag., vol. 8, pp. 107–115, 2025.

[6] G. Apruzzese, P. Laskov, E. Montes de Oca et al., "The role of machine learning in cybersecurity," Digital Threats: Res. Pract., vol. 4, no. 1, pp. 1–38, 2023.

[7] N. Ahmed, A. Ngadi, J. M. Sharif et al., "Network threat detection using machine/deep learning in SDN-based platforms: A comprehensive analysis," Sensors, vol. 22, no. 20, p. 7896, 2022.

[8] M. Ahsan, K. E. Nygard, R. Gomes et al., "Cybersecurity threats and their mitigation approaches using machine learning—A review," J. Cybersecur. Priv., vol. 2, no. 3, pp. 527–555, 2022.

[9] J. Lee, J. Kim, I. Kim, and K. Han, "Cyber threat detection based on artificial neural networks using event profiles," IEEE Access, vol. 7, pp. 165607–165626, Nov. 2019.

[10] A. Yazdinejad, M. Kazemi, R. M. Parizi et al., "An ensemble deep learning model for cyber threat hunting in industrial internet of things," Digit. Commun. Netw., vol. 9, no. 1, pp. 101–110, 2023.

[11] K. Simran, P. Balakrishna, R. Vinayakumar et al., "Deep learning approach for enhanced cyber threat indicators in Twitter stream," in Proc. SSCC 2019, CCIS 1208, pp. 135–145, Springer, 2020.

[12] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. ICISSP, pp. 108–116, 2018. [Dataset: https://www.unb.ca/cic/datasets/ids-2017.html]

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD, pp. 1135–1144, 2016.

[14] M. F. A. Razak, N. B. Anuar, F. Othman et al., "Bio-inspired for features optimization and malware detection," Arab. J. Sci. Eng., vol. 43, pp. 6963–6979, Dec. 2018.

[15] Q. Liu, P. Li, W. Zhao et al., "A survey on security threats and defensive techniques of machine learning: A data driven view," IEEE Access, vol. 6, pp. 12103–12117, Feb. 2018.

[16] A. Basit, M. Zafar, X. Liu et al., "A comprehensive survey of AI-enabled phishing attacks detection techniques," Telecommun. Syst., vol. 76, pp. 139–154, Oct. 2020.

[17] P. Ke, G. Meng, T. Finley et al., "LightGBM: A highly efficient gradient boosting decision tree," in Proc. NeurIPS, vol. 30, pp. 3146–3154, 2017.

[18] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD, pp. 785–794, 2016.

[19] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[20] M. Pawlicki, M. Choras, R. Kozik et al., "On the impact of network data balancing in cybersecurity applications," in Proc. ICCS 2020, LNCS 12140, pp. 196–210, 2020.