

Original Article

An Optimized Ensemble Intrusion Detection Framework with Hybrid Feature Selection and Multi-Dataset Validation

Pooja Rani¹, Parvinder Singh²

^{1,2}Computer Science and Engineering, Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Sonapat, Haryana, India.

¹Corresponding Author: poojaresearch.26@gmail.com

Received: 08 June 2026

Revised: 08 September 2026

Accepted: 14 October 2026

Published: 28 October 2026

Abstract - Network Intrusion Detection Systems (NIDS) play a vital part in protecting modern communication networks against many classy cyberattacks. But the existing intrusion detection methods suffer from problems including data imbalance, redundant features, and poor performance in various network environments. Hence, this work presents a optimized ensemble Intrusion Detection System (IDS) based on comprehensive data preprocessing, hybrid feature selection, and ensemble learning techniques to improve intrusion detection performance. To enhance the data quality and class balance, the proposed framework includes missing value handling, outlier removal, noise filtering and Synthetic Minority Oversampling Technique (SMOTE). A hybrid feature selection method, which combines the Mutual Information (MI) and Recursive Feature Elimination (RFE), is used to identify the most useful network traffic features. Then, an attack classification ensemble model is constructed using Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM). This framework is tested on the CICIDS2017 and UNSW-NB15 datasets via 5-fold cross-validation, and its performance is measured by standard classification metrics. Additionally, explainability using SHapley Additive exPlanations (SHAP) is integrated to enhance the interpretability of the model prediction. The experimental results show that the proposed framework provides a strong and stable intrusion detection performance in both of the benchmark datasets.

Keywords - Ensemble learning, Explainable Artificial Intelligence (XAI), Hybrid feature selection, Mutual Information, Recursive Feature Elimination, Network Intrusion Detection System (NIDS).

1. Introduction

Modern network infrastructures are becoming increasingly complex due to the development of cloud computing, digital communication technologies, Internet of Things (IoT) devices and interconnected information systems. As these advances are beneficial, they also open the door to complex cyber-attacks. This has made NIDS an integral part of cybersecurity strategies, helping to detect and neutralize threats. Known attacks can be handled by traditional signature-based IDS, but are not effective against novel and evolving attacks. This has made attention to Machine Learning (ML)-based intrusion detection techniques as they can adapt to the complex traffic patterns and can be able to detect new attacks.

Recently, ML and Deep Learning (DL) techniques are effective in improving the IDSs performance on benchmark datasets like CICIDS2017 and UNSW-NB15. The use of many of the existing systems remains constrained by issues such as high-dimensional feature space, class imbalance,

noisy network traffic, and model interpretability. These three areas, namely, data preprocessing, feature optimization, and explainability, have been identified as the pivotal research areas in intrusion detection research from recent reviews [1, 2]. There have been multiple research studies that have studied ensemble learning strategies to enhance attack detection performance. Ensemble-based models and hybrid learning [3] have proven better performance in classification than single ML models, especially in the context of heterogeneous network traffic patterns. Recent research has shown that ensemble and hybrid architectures can achieve high detection rates and performance for benchmark intrusion detection datasets.

One of the major difficulties is redundant and irrelevant features included in network traffic datasets. High-dimensional feature spaces can lead to a high computational cost and poor generalization of the model. In order to overcome such a problem, feature selection technique has been broadly employed to determine the features that carry the



most useful information for IDS. A number of earlier works using optimization-based feature selection techniques, such as adaptive bacterial foraging method and evolutionary optimization method [4, 5] showed that good feature reduction can significantly enhance classification accuracy on the basis of a reduction of the computation. These results are encouraging for the use of sophisticated feature selection algorithms in new IDSs.

Apart from the detection performance, explainability has become a most important demands for the cybersecurity application. Advanced ML models often have high rates of accuracy, but the “black box” leaves security analysts without knowledge of the logic behind intrusion predictions.

The need for more transparency, trustworthiness, and decision support in IDSs is being highlighted in recent research by the importance of XAI [6, 7]. Explainable models that are based on SHAP and other explainability techniques have become more prevalent in recent years to give interpretable explanations without loss of predictive accuracy.

Despite the progress of ML-based IDS, several technical problems remain unresolved. Network traffic datasets frequently contain class imbalance, redundant features, noisy observations, and heterogeneous attack patterns. These characteristics can cause a classifier to favor the majority class. Consequently, high overall accuracy alone may conceal missed attacks or excessive false alarms. Therefore, an IDS framework must consider feature relevance, false-positive behavior, minority-class detection, and model interpretability together.

Existing studies have examined feature selection, class balancing, ensemble classification, and XAI; however, these components are often evaluated independently. Filter-based feature selection can efficiently rank variables but may not identify the subset that works best with a particular classifier. Whereas wrapper-based selection accounts for classifier performance but is computationally more demanding.

Similarly, most ensemble works use fixed or equal voting without reporting how classifier weights and decision thresholds are determined. Additionally, preprocessing and oversampling are sometimes applied before cross-validation, which can transfer information from validation samples into model development and produce optimistic estimates.

Therefore, this study addresses the research problem of the development of a reproducible and leakage-controlled binary IDS that integrates class balancing, hybrid feature selection, optimized ensemble classification, and explainable prediction. Mutual Information and Recursive Feature Elimination are combined to obtain a compact feature subset. Also, XGBoost and LightGBM probabilities are aggregated using training-derived soft-voting weights and a decision

threshold. The operations, including SMOTE, normalization, feature selection, model fitting, and ensemble tuning, are also performed within the training portion of each fold. The framework is trained and evaluated on CICIDS2017 and UNSW-NB15 independently using stratified five-fold cross-validation. Therefore, the experiments represent multi-dataset evaluation rather than multi-dataset generalization. Based on this research problem, the contributions of the study are summarized as follows:

- A leakage-controlled intrusion detection pipeline is developed in which imputation, normalization, MI ranking, RFE, and SMOTE are fitted to the training portion of each cross-validation fold.
- A hybrid MI–RFE feature-selection strategy is implemented to obtain subsets of six features for CICIDS2017 and ten features for UNSW-NB15 while retaining binary classification performance.
- An optimized soft-voting ensemble is developed by combining the attack-class probabilities generated by XGBoost and LightGBM. The XGBoost weight and classification threshold are selected from training data rather than random fix.
- An ablation study is done, thereby identifying the predictive contribution and trade-offs associated with each component.
- Fold-wise results, mean and standard deviation, confidence intervals, out-of-fold confusion matrices, Wilcoxon signed-rank comparisons, and McNemar tests are reported.
- SHAP analysis is done to explain the influence of the selected traffic characteristics on binary benign-versus-attack decisions.

The rest article organization is as follows. Section 2 gives the related literature on IDS, ensemble learning, feature selection techniques, and explainable artificial intelligence. Section 3 offers the proposed intrusion detection framework, comprising preprocessing, feature selection, ensemble learning, and explainability components.

Section 4 defines the experimental setup, evaluation metrics, datasets, implementation details, experimental results, comparative analysis, ablation study, and explainability findings. Lastly, Section 5 concludes the paper and plans probable instructions for future research.

2. Literature Review

2.1. Machine Learning and Deep Learning-based Intrusion Detection

As cyber threats become more complex, the use of ML and DL techniques has gained great momentum for NIDS. The traditional signature-based IDSs are unable to detect new forms of attacks, which has prompted researchers to research intelligent anomaly detection frameworks. In recent years, ML

techniques like Random Forest (RF), XGBoost, Support Vector Machine (SVM), and LightGBM have successfully been able to classify malicious network traffic with high detection rates.

For instance, on the CICIDS2017 and UNSW-NB15 benchmark datasets, Waghmode et al. (2025) [9] tested several ML models and found significant improvements in intrusion detection accuracy over traditional methods. Likewise, Nassef et al. (2025) [10] used XGBoost-based learning for attack detection and achieved excellent classification results with existing network traffic datasets.

In addition, DL techniques have received a lot of interest for their automatic learning of complicated traffic patterns. There is extensive research using Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and Transformer-based architectures for IDS. Umer et al. (2025) [11] introduced a Transformer-based IDS that used a wrapper-based feature selection method, which successfully improved the detection accuracy of the model on the UNSW-NB15 dataset.

Similarly, Ahmed et al. (2024) [12] combined CNN with the XGBoost model to enhance ID system performance in IoT networks. While these progressions have been made, DL models frequently consume enormous computational assets and can lack interpretability, making it hard to apply them practically in cybersecurity operations.

2.2. Ensemble Learning and Feature Optimization

In recent years, multiple classifiers have developed to overcome the drawbacks of single classifiers and ensemble learning techniques are introduced. Ensemble models are constructed from several learning algorithms to provide the model with better predictive performance, robustness and generalization ability. Kumar et al. (2025) [13] proposed a hybrid IDS based on RFE and multiple classifiers on CICIDS2017, UNSW-NB15 and other benchmark datasets. They found that hybrid architectures can achieve higher accuracy as well as attack detection rates than individual classifiers.

Network traffic datasets also have redundant and irrelevant attributes, so feature selection is another vital part of an IDS. More et al. (2024) [14] showed that feature ranking techniques can be used in combination with ensemble learning to decrease the number of computations involved in the classification task without compromising its performance.

Similarly, Asry et al. (2025) [15] used RFE algorithm with SHAP features to select important network traffic features and increase the efficiency of the model. The optimization of features is emphasized in these studies, but the combination of several of these feature selection approaches in an integrated ensemble system has been underexplored.

2.3. Explainable Artificial Intelligence in Intrusion Detection

While often times the advanced ML and DL models provide high predictive performance, the black-box nature of these models is a challenge to cybersecurity practitioners. XAI has emerged as a promising solution as it offers explanations that are transparent to the model decisions.

Mohale, V. Z., and Obagbuwa, I. C (2025) [2] highlighted that explainability improves the trust and usability of cybersecurity systems by allowing analysts to grasp the reasoning behind traffic classification. Likewise, Muhammad et al. (2025) [16] explored LIME-based explanation methods and showed that LIME-based explanations improve the interpretability of ID predictions.

However, in the list of available XAI methods, SHAP has become very popular, owing to its capacity to offer both local and global descriptions, being consistent with the model's behavior.

Arreche et al. (2024) [17] suggested an explainable IDS that integrates SHAP analysis and emphasized its ability to enhance transparency and decision-making. However, most of the previous research studies are primarily interested in explainability or predictive accuracy separately, but not amalgamating them together in a holistic DIDS system.

2.4. Research Gap

This section shows advances in ML, DL, ensemble learning, feature optimization and XAI techniques for IDS, still several research gaps remain. Many studies validate their models with one set of benchmark data, limiting robustness under different network traffic characteristics. Second, previous studies often focus on prediction and offer little focus on security analyst explainability.

Third, feature selection, class-imbalance handling, ensemble classification, and explainability are investigated as separate enhancements rather than as complementary components of a unified IDS pipeline. Consequently, there remains scope for an integrated framework that addresses data quality, class imbalance, feature relevance, classification robustness, and model interpretability.

To address these gaps, the present study integrates preprocessing and SMOTE-based class balancing with MI-RFE feature optimization, XGBoost-LightGBM soft-voting ensemble learning, and SHAP-based explainability.

This framework is evaluated independently on CICIDS2017 and UNSW-NB15 to examine its performance under two distinct benchmark environments. Table 1 summarizes representative recent studies and positions the proposed framework relative to existing approaches.

Table 1. Recent studies related to intrusion detection systems

Author/Year	Dataset	Method	Key Focus	Limitations
Mohale, V. Z., and Obagbuwa, I. C (2025) [2]	IoT IDS data	ML + XAI	Improved interpretability in IDS using XAI	Focus mainly on explainability, not full hybrid feature-selection ensemble framework.
Waghmode et al. (2025) [9]	NSL-KDD, CICIDS2017, UNSW-NB15	ML classification	Multi-dataset intrusion detection	Binary classification focus; limited ensemble explainability
Asry et al. (2025) [15]	Cybersecurity dataset	SHAP + RFECV	Feature selection using SHAP and RFECV	Does not fully combine MI-RFE with XGBoost–LightGBM ensemble
Umer et al. (2025) [11]	UNSW-NB15	Wrapper FS + Transformer	Wrapper-based feature selection and attention model	Single dataset validation only
Ahmed et al. (2024) [12]	IoT IDS datasets	CNN + XGBoost + SHAP	Hybrid model with SHAP-based interpretability	IoT-focused; not cross-validated across CICIDS2017 and UNSW-NB15
Arreche et al. (2024) [17]	IDS datasets	XAI-IDS framework	End-to-end explainable AI for IDS	More explainability-oriented than hybrid feature optimization
More et al. (2024) [14]	UNSW-NB15	Stacking ML + feature selection	MI and Extra Tree feature selection	Single dataset; limited cross-dataset generalization
Kumar et al. (2025) [13]	CICIDS2017, NSL-KDD, UNSW-NB15, CSE-CIC-IDS2018	Hybrid IDS + RFE-RF	Multi-dataset hybrid IDS for edge computing	Explainability not central to framework
Nassef et al. (2025) [10]	UNSW-NB15, CICIDS2017, others	XGBoost + feature selection	DDoS detection using XGBoost	Attack-specific focus; less general NIDS framework
Muhammad et al. (2025) [16]	UNSW-NB15	LIME-based XAI IDS	Explainable intrusion detection	Uses LIME; limited ensemble-model comparison
Patil et al. (2025) [18]	CICIDS2017	Feature selection + ensemble learning	Optimized NIDS with feature selection	Uses only CICIDS2017; no UNSW-NB15 validation
Fiona Lawrence and Rajesh Nigam. (2025) [19]	CICIDS2017, NF-UNSW-NB15	LightGBM, XGBoost, CatBoost	Ensemble model comparison	Comparative work; limited integrated explainability
CICIDS2017 official dataset [20]	CICIDS2017	Benchmark dataset	Includes benign and common attacks with labeled flows	Dataset itself, not a model study
UNSW-NB15 official dataset [21]	UNSW-NB15	Benchmark dataset	Comprises nine attack groups.	Dataset itself, not a model study

Table 1 indicates some recent studies with results for feature selection, ensemble classifiers, attention-based models, and explainability techniques. However, these approaches provide limited combined evidence concerning training-fold-only preprocessing, optimized probability fusion, constituent-model comparisons, component-level ablation, and paired statistical testing.

Thus, the proposed framework is positioned as an integrated and reproducible evaluation of these complementary elements rather than as a claim that its individual algorithms are independently novel.

3. Research Methodology

3.1. Proposed Framework

In this study, an optimized ensemble intrusion detection framework is designed to improve the detection accuracy, feature efficiency, model robustness and interpretability in heterogeneous network environments.

The framework combines the data preprocessing, hybrid feature selection, ensemble learning, and XAI under the same architecture and structure.

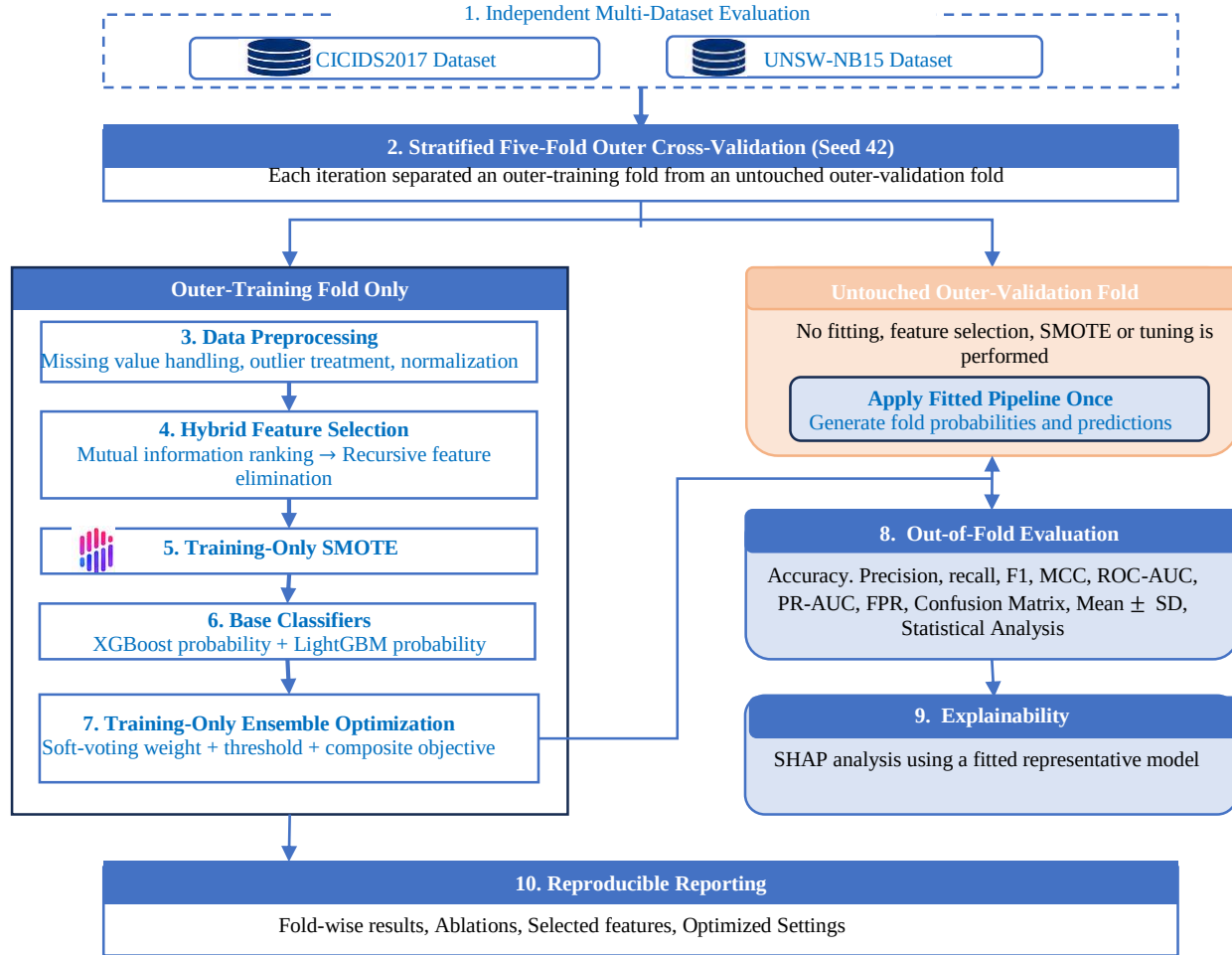


Fig. 1 Proposed methodology architecture

The workflow is divided into six steps: (i) data acquisition and preprocessing, (ii) feature optimization, (iii) building the ensemble model, (iv) cross-validation, (v) explainability analysis, (vi) performance evaluation. The raw network traffic data used in this work is taken from CICIDS2017 and UNSW-NB15 benchmark datasets.

The datasets are preprocessed before the processing, which includes missing value imputation, outlier removal, noise reduction, and SMOTE to balance the classes. Then, a hybrid feature selection mechanism based on the MI and RFE is utilized for identifying the informative network traffic attributes and for dimensionality reduction.

A feature subset that is optimized is then fed into the ensemble of XGBoost and LightGBM classifiers. Both classifiers are connected in an ensemble of classifiers based on voting to enhance the robustness of the classifiers and attack detection ability. The Stratified 5-Fold Cross-Validation is used to make sure that the model is reliable and is applied to other datasets. Finally, SHAP-based

explainability is used to generate interpretable explanations of model decisions. Figure 1 depicts the overall flow of the proposed framework.

3.2. Dataset Description

For effectiveness analysis of the proposed intrusion detection framework, experiments are performed on two popular benchmark datasets; namely CICIDS2017 and UNSW-NB15. The datasets are used in a variety of cybersecurity research because they possess realistic network traffic properties and have a broad coverage of present-day cyberattacks. The CICIDS2017 dataset consists of real-life network traffic generated in an enterprise environment, both benign and malicious.

The various attack categories in the dataset are Distributed Denial of Service (DDoS), Brute force attacks, Botnet, Web attacks, Infiltration attacks and Port scanning. Due to its diversity and realistic traffic patterns, CICIDS2017 is extensively used for evaluating intrusion detection models. The UNSW-NB15 dataset was developed to overcome some

of the drawbacks of previous benchmark datasets and contains current attack scenarios obtained in a controlled network environment. This dataset includes normal traffic and nine categories of attacks, such as exploits, reconnaissance, DoS, shellcode, worms, backdoors and fuzzers.

3.3. Data Preprocessing

Raw network datasets are likely to have missing values, duplicate records, outliers, noisy observations, and an imbalanced distribution of the attacks, which can adversely affect the performance of models. Hence, a complete preprocessing pipeline is developed to enhance the quality of the data before feature optimization and classification. First, duplicate records are detected and eliminated to avoid bias when training the model. Subsequently, missing values are dealt through imputation strategies.

After data cleaning, feature normalization is carried out to minimize the discrepancy on the level of features and to normalize them across all features of the network traffic. Normalization helps in the convergence of models and avoids features with larger numbers dominating the learning process, as seen in intrusion detection datasets, which have features with different ranges of numbers. The normalized value of a feature is defined as Equation (1):

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Where X represents the original feature value, X_{min} denotes the features minimum value, also X_{max} denotes the maximum feature value. Outlier detection is subsequently applied to identify abnormal observations that originate from data collection errors or extreme noise. The Interquartile Range (IQR) method is employed due to its robustness in handling skewed network traffic distributions. The lower and upper bounds are computed as Equations (2)-(4):

$$LB = Q_1 - (1.5 \times IQR) \quad (2)$$

$$UB = Q_3 + (1.5 \times IQR) \quad (3)$$

$$IQR = Q_3 - Q_1 \quad (4)$$

and Q_1 and Q_3 signify the first and third quartiles, respectively. The noisy and anomalous observations are removed, and class imbalance is addressed using SMOTE. Network intrusion datasets usually have uneven distributions of attacks, with minority classes of attacks having much lower numbers than that of normal traffic. Such imbalance often biases ML models toward majority classes and reduces attack detection effectiveness. SMOTE creates synthetic minority class samples by interpolation among neighboring observations. The synthetic instance is created by Equation (5):

$$X_{new} = X_i + \lambda(X_{nn} - X_i) \quad (5)$$

Where X_i is the minority class sample, X_{nn} is a nearest neighbors, λ is a random value within the interval $([0,1])$. The final pre-processed dataset obtained after cleaning, normalization, outlier handling, and class balancing is represented in Equation (6):

$$D_p = \{x_1, x_2, x_3, \dots, x_n\} \quad (6)$$

Where D_p denotes the optimized preprocessed dataset. This is the feature set that is fed into the hybrid feature selection stage. The proposed pre-processing method is intended to tackle some of the significant issues with intrusion detection datasets. The treatment of missing values increases the completeness of the data, the normalization process stabilizes learning, the removal of outliers eliminates noise, and the SMOTE process enriches the minority in the attack. The framework includes these preprocessing steps in a single pipeline, so the resulting dataset is cleaner and has a more balanced ratio of values, thus enabling better optimization of features and classification.

3.4. Hybrid Feature Selection

Modern intrusion datasets have much information about the traffic, but not all of these features are useful for attack detection. The redundant data can make the system more complicated to compute and might affect generalization. To overcome this drawback, a two-stage hybrid feature optimization framework is presented, which involves MI followed by RFE. The pre-processed dataset D_p obtained from the previous stage serves as the input to the feature selection module.

3.4.1. Mutual Information-based Feature Ranking

The dependency between an input feature and the target class is measured by the MI. The higher the MI score, the more intrusion detection is contributed, and that feature is considered more informative. The MI score between feature X and target class Y is given by Equation (7):

$$MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (7)$$

Where $p(x, y)$ denotes the joint probability distribution, $p(x)$ and $p(y)$ represent marginal probability distributions. All features are ranked according to their MI scores, and only the top-ranked features are retained for further optimization. The resulting feature subset is represented by Equation (8):

$$F_{MI} = \{f_1, f_2, f_3, \dots, f_m\} \quad (8)$$

Where m is less than n .

3.4.2. Recursive Feature Elimination

The subset of features selected from F_{MI} is then filtered with a RFE process. RFE is a method that is used to iteratively train a learning model, determine the importance of features and eliminate the least important features until the best reduced subset is found. The feature importance score at each iteration is estimated as $Imp(f_i)$ and features that have the least importance are eliminated from the model. The optimized subset of features is represented as in equation (9):

$$F^* = \{f_1, f_2, f_3, \dots, f_k\} \quad (9)$$

Where $k < m < n$. The optimized feature subset F^* becomes the input to the ensemble classification stage. The proposed two-stage feature optimization framework leverages the merits of the two methods like filter-based feature selection as well as wrapper-based feature selection. MI is efficient in eliminating irrelevant features and is still computationally efficient. The RFE technique then narrows down the feature set by evaluating the performance of the classifiers. The fusion of these two approaches yields the benefits of dimensionality reduction, computational efficiency and better classification performance than if used alone.

3.5. Ensemble Classification Model

The classification algorithm is designed from XGBoost and LightGBM to take advantage of the strengths of both algorithms and make the detection more robust in a heterogeneous network environment. The inspiration for using an ensemble architecture is based on the observation that individual classifiers often show varying performance across attack categories. XGBoost has high prediction ability and resistance to overfitting, while LightGBM has a high speed of training and process high-dimensional data efficiently. The integration of these models benefits from complementary learning features and enhances the overall intrusion detection effectiveness. XGBoost is a gradient boosting algorithm that builds Decision Trees (DTs) one by one to reduce prediction errors. The model prediction at iteration (t) is represented by Equation (10):

$$\hat{y}_i^{(t)} = \sum_{k=1}^t f_k(x_i) \quad (10)$$

Where f_k denotes the DT generated at iteration k , x_i signifies the input feature vector, also $\hat{y}_i^{(t)}$ denotes the predicted output. The objective function of XGBoost combines classification loss and regularization by Equation (11):

$$Obj = \sum_{i=1}^n L(y_i, \hat{y}_i) + \sum_{k=1}^t \Omega(f_k) \quad (11)$$

Where L represents the loss function, and Ω controls model complexity. With the optimized feature subset F^* , LightGBM builds the DTs by selecting splits that maximize information gain. The information gain is defined as in Equation (12):

$$Gain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \quad (12)$$

Where G_L and G_R represent gradient sums, H_L and H_R represent Hessian sums, and λ is the regularization parameter. The outputs generated by XGBoost and LightGBM are combined using a soft voting mechanism. For a network-flow observation x_i , XGBoost and LightGBM produce attack-class probabilities denoted by p_i^{XGB} and $\hat{p}_i^{LGB}$, respectively. The optimized soft-voting probability is calculated as in Equation (13):

$$p_i^{ENS} = wp_i^{XGB} + (1 - w)p_i^{LGB} \quad (13)$$

Where w in $[0,1]$ is the contribution of XGBoost and $(1 - w)$ is the corresponding contribution of LightGBM. A value of $(w > 0.5)$ gives greater influence to XGBoost, whereas $(w < 0.5)$ gives high influence to LightGBM. The final binary decision is obtained by Equation (14):

$$\hat{y}_i = \begin{cases} 1, & p_i^{ENS} \geq \tau \\ 0, & p_i^{ENS} < \tau \end{cases} \quad (14)$$

Where class 1 represents attack traffic, class 0 represents benign traffic, and τ denotes the decision threshold. The weight and threshold are selected exclusively from predictions generated within the training data of each outer fold. The outer validation fold is not used during configuration selection. Moreover, the ensemble approach gives a solid basis for the later explainability analysis of the ensemble through the framework of SHAP and allows for accurate and interpretable intrusion detection.

3.6. Cross-Validation Strategy

Stratified 5-Fold Cross-Validation is used on training and testing the proposed intrusion detection framework to ensure the reliability and generalizability of the proposed framework. Cross-validation uses many different train-test splits to give thorough estimation of a model's performance than a single train-test split. The approach proposed in this work is to divide the selected feature subset, F^* from the hybrid feature selection stage into 5 mutually exclusive folds while maintaining the original class distribution of the dataset. In every iteration, four folds train and one fold validates the result. This is repeated five times to ensure that each fold is used once for each test. The mean of all the folds is considered as by Equation (15):

$$CV_{avg} = \frac{1}{K} \sum_{i=1}^K M_i \quad (15)$$

Where K represents the number of folds, which is 5 in this research, M_i denotes the performance metric obtained from fold i . This stratified cross-validation in intrusion detection datasets is important because attack classes are often imbalanced. This ensures that class sizes are kept equal throughout all folds, giving a more accurate indication of the

performance of the model and minimizing the chance of biased results.

3.7. SHAP-based Explainability Analysis

Ensemble learning models are usually complex, and the decision rules in the models are hard to explain. In cybersecurity environments, security analysts need to be provided with clear explanations to better comprehend why certain network traffic instances are considered an attack. So, SHAP are included in the proposed framework to improve interpretability of the models. SHAP is a cooperative game theory approach and measures each feature impact on the ultimate forecast. The marginal contribution of a feature is calculated by observing its value with and without the other features to obtain the SHAP value of that feature. The SHAP explanation model can be formulated by Equation (16):

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i \quad (16)$$

Where $g(z')$ represents the explanation model, ϕ_0 denotes the baseline prediction, ϕ_i signifies the feature i contribution, z'_i specifies the presence or absence of feature i , and M is the total number of selected features. The resulting SHAP values have local and global interpretability. Local explanations explain what determines particular predictions, while global explanations explain what features in the data are most important.

3.8. Performance Evaluation Metrics

The proposed intrusion detection framework is tested with respect to various classification metrics that are widely used. They give a complete picture of the detection capability, classification reliability and overall predictive performance. The confusion matrix has four basic results True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Accuracy measures the proportion of correctly classified instances.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (17)$$

Precision evaluates the proportion of predicted attacks that are malicious.

$$Precision = \frac{TP}{TP+FP} \quad (18)$$

Recall measures the model's ability to properly recognize attack instances.

$$Recall = \frac{TP}{TP+FN} \quad (19)$$

The F1-Score offers a balanced evaluation by combining precision and recall.

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (20)$$

ROC-AUC compares the ability of the classifier with regard to various threshold values. The greater the AUC, the better the classification and the better the separation of attack and normal traffic classes. The proposed framework results are compared with existing ML and ensemble-based intrusion detection methods to evaluate its effectiveness.

Algorithm 1. Pseudocode for Proposed Methodology

- **Input:** Dataset D, Number of outer folds $K = 5$, XGBoost weight grid W, Threshold grid T, MI candidate counts M, RFE feature counts R and SMOTE ratios S
- **Output:** Out-of-fold probabilities and predictions
 - 1: Construct K stratified outer folds using random seed 42
 - 2: for each outer fold $k = 1, \dots, K$ do
 - 3: Divide D into outer-training data D_{tr} and validation data D_{val}
 - 4: Fit imputation and normalization using D_{tr} only
 - 5: Apply the fitted transformations to D_{val}
 - 6: for each candidate (m, r, s, w, τ) in $M \times R \times S \times W \times T$ do
 - 7: Select m features by MI using training data
 - 8: Refine the subset to r features using RFE
 - 9: Apply SMOTE with ratio s to the training subset only
 - 10: Train XGBoost and LightGBM
 - 11: Obtain training-validation probabilities p^{XGB} and p^{LGB}
 - 12: Calculate Equation (13)
 - 13: Calculate binary predictions using threshold τ
 - 14: Evaluate the composite optimization score J
 - 15: end for
 - 16: Retain the candidate having the highest training-only score
 - 17: Refit the selected pipeline using the complete D_{tr}
 - 18: Generate probabilities and predictions for untouched D_{val}
 - 19: Store fold metrics and out-of-fold predictions
 - 20: end for
 - 21: Aggregate the five folds using mean, standard deviation and 95% CI
 - 22: Perform Wilcoxon and McNemar comparisons

4. Results and Discussion

On two benchmark datasets, CICIDS2017 and UNSW-NB15, the proposed intrusion detection framework was tested. Although both datasets contain multiple attack categories, this study addresses binary intrusion detection. For CICIDS2017, all records labelled as BENIGN were assigned to class 0, whereas all attack categories were combined and assigned to class 1. For UNSW-NB15, normal observations were assigned to class 0 and all nine attack categories were combined as class 1. Therefore, all metrics correspond to benign-versus-attack classification.

Experiments were run using ML libraries in Python such as Scikit-learn, XGBoost, LightGBM, Imbalanced-Learn and

SHAP. This was compared with three baseline classifiers: DT, RF, SVM. In this study, the experimental findings are reported at two complementary levels. First, the original hold-out analysis is retained to illustrate class balancing, feature relevance, model interpretability, and threshold-dependent behaviour against DT, RF, and SVM. Second, the proposed framework and its variants are evaluated using leakage-controlled stratified five-fold cross-validation. The five-fold results create the confirmatory evaluation and are reported separately.

All experiments were conducted in the Kaggle notebook environment equipped with an NVIDIA Tesla T4 GPU. Python-based implementations of XGBoost, LightGBM, scikit-learn, and imbalanced-learn were used. A fixed random seed of 42 was applied to data partitioning, SMOTE, feature selection, and model training to support reproducibility. To prevent information leakage, every data-dependent operation was limited to the training portion of each outer fold.

Interquartile-range clipping bounds and RobustScaler parameters were estimated from the training fold and applied to the corresponding validation fold subsequently. MI ranking and RFE were also performed using only training-fold observations. SMOTE was applied after feature selection and only to the training fold. Ensemble weights and decision thresholds were selected from training-derived predictions without using the outer-validation observations. The validation fold remained unchanged until final evaluation.

4.1. Results on CICIDS2017 Dataset

The CICIDS2017 dataset was pre-processed and then is reduced to 1,667,238 samples, and six features were selected. The dataset was highly imbalanced prior to SMOTE, with 833,619 benign samples and 44,934 attack samples. Out of this, both classes had 833,619 samples after the SMOTE process, resulting in a balanced dataset for modeling. The pictorial representation is given in Figure 2.

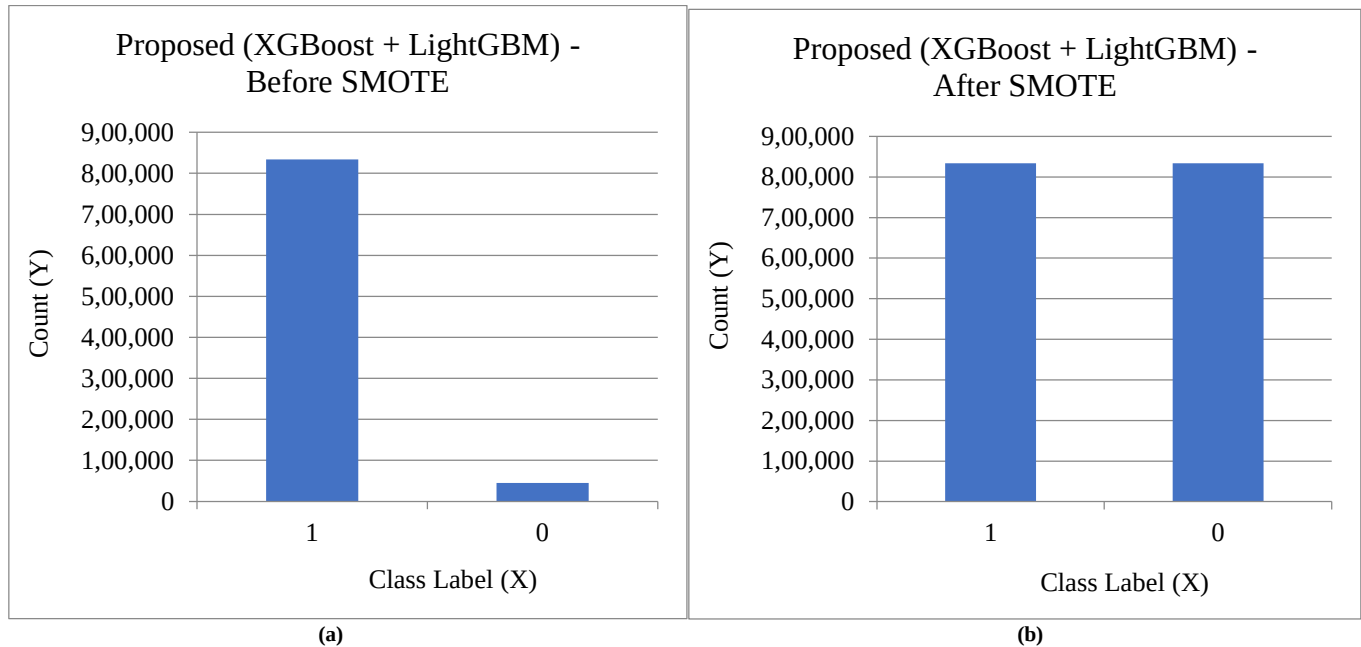


Fig. 2 Class distribution in the original hold-out analysis (a) Before, and (b) After SMOTE for CICIDS2017 dataset.

Table 2. Performance comparison on CICIDS2017 dataset in the original hold-out analysis

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
DT	0.9253	0.9021	0.9541	0.9274	0.9805
RF	0.9590	0.9948	0.9228	0.9575	0.9935
SVM	0.9061	0.9885	0.8218	0.8975	0.9526
Proposed XGBoost + LightGBM	0.9802	0.9840	0.9762	0.9801	0.9984

The proposed ensemble model, as in Table 2, showed the highest performance with 98.02% accuracy, 97.62% recall, 98.01% F1-score and 99.84% ROC-AUC. RF had slightly higher precision but lower recall, indicating it missed out more attacks. Hence, the proposed framework is more balanced and practically useful due to its ability of having high precision

and high recall. This is because SVM might be more difficult to model complex nonlinear traffic patterns. The recall results showed that DT had higher recall than SVM, while the precision results showed lower precision than SVM, meaning more false alarms. This is overcome by the proposed framework.

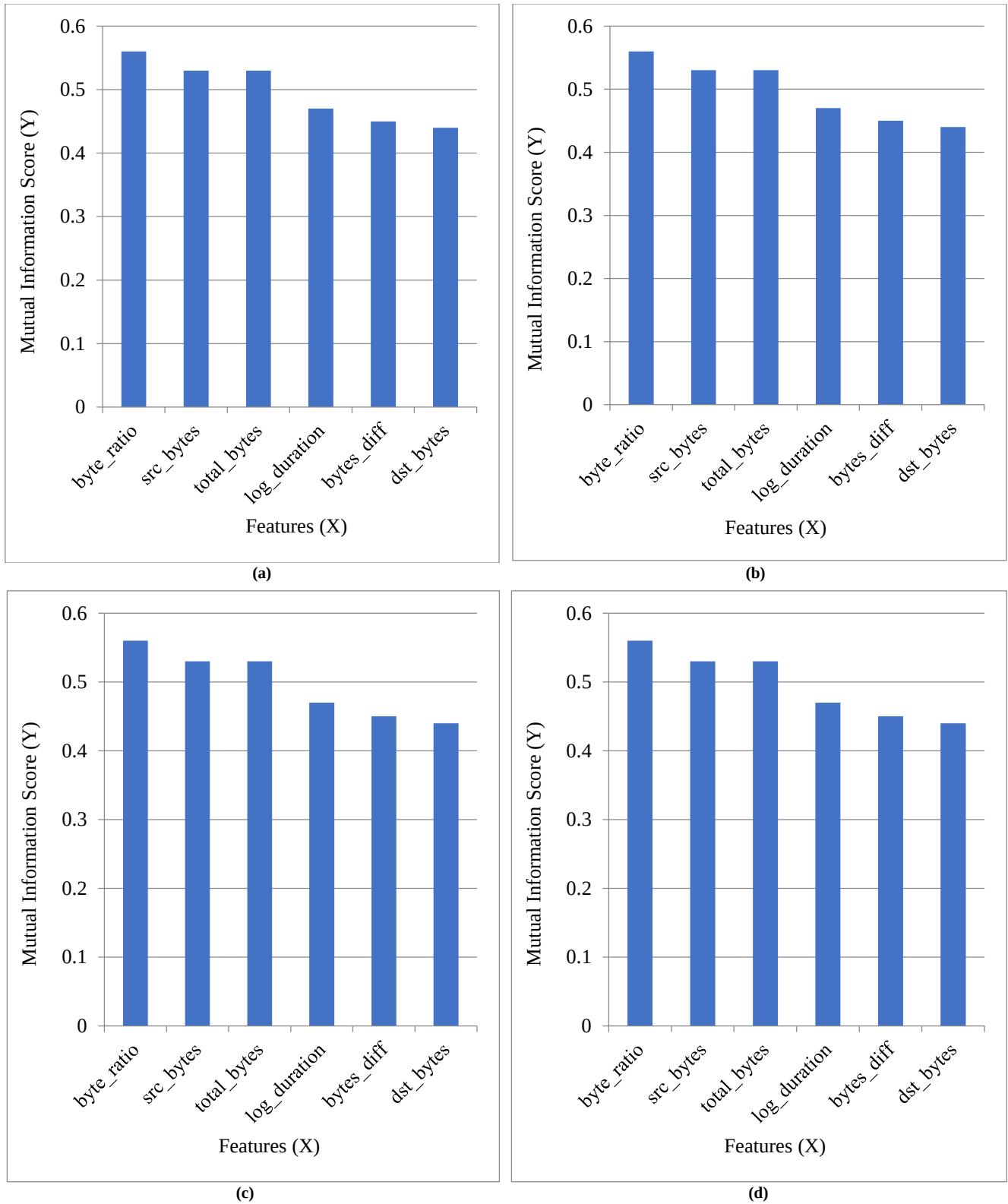


Fig 3. Mutual Information feature ranking for CICIDS2017 dataset in the original hold-out analysis (a) DT, (b) RF, (c) SVM, and (d) Proposed.

The final features that are selected by MI are byte_ratio, src_bytes, total_bytes, log_duration, bytes_diff, and dst_bytes and is given in Figure 3. The features are all concerning traffic

volume, byte distribution and communication duration – attributes that are very relevant for separating the benign and malicious flows.

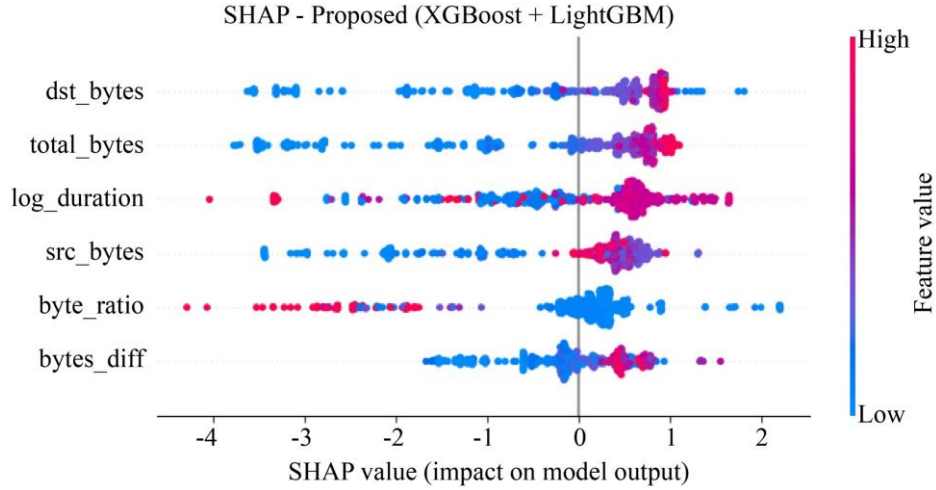


Fig. 4 SHAP summary plot for CICIDS2017 dataset in the original hold-out analysis

The SHAP analysis in Figure 4 also revealed that attributes related to bytes and duration had a significant impact on the model's predictions. Therefore, the proposed preprocessing, feature selection and ensemble-learning approach prove to be more accurate and easier to interpret according to the results obtained during the CICIDS2017.

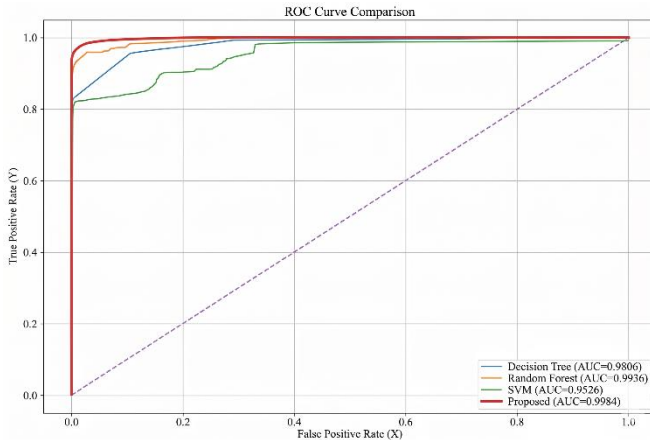


Fig. 5 ROC curve comparison on CICIDS2017 dataset in the original hold-out analysis

It is also revealed from the ROC curve analysis in Figure 5 that the proposed framework has better discriminative capability. The proposed ensemble using XGBoost and LightGBM achieved ROC-AUC value of 0.9984, which is better than other tested methods like DT (0.9806), RF (0.9936), and SVM (0.9526). This means that the proposed framework still has a better separation between the attack and benign traffic at different classification thresholds.

The proposed framework is also effective, based on the Precision-Recall (PR) curve as in Figure 6. The proposed framework achieved the highest PR score of 0.9985 followed by RF (0.9940), DT (0.9742) and SVM (0.9628). This result is significant for the imbalanced intrusion detection problem,

for which the performance of PR is particularly useful. The proposed framework still has a high precision even if recall increases, which indicates that the framework could detect more attacks without giving too many false alarms. In general, the ROC and PR curves confirm both the accuracy of the proposed ensemble approach and its stability when the threshold changes. This validates its viability for IDSs that demand a high detection rate and low false alarm rate.

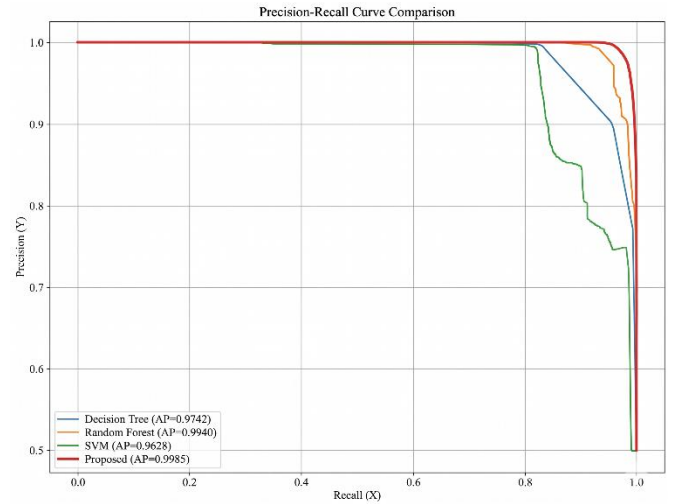


Fig. 6 Precision-Recall curve comparison on CICIDS2017 dataset in the original hold-out analysis

4.2. Results on UNSW-NB15 Dataset

UNSW-NB15 is a more difficult benchmark dataset as it includes a variety of attack categories and realistic network traffic characteristics. The above preprocessing, outlier removal, noise filtering and feature engineering made the data 53,872 samples with 10 selected features. The dataset prior to SMOTE was of moderate imbalance with 26,936 attacks and 14,413 benign samples. SMOTE brought both classes equal size to 26,936 per class to fairly train the model, and is illustrated in Figure 7.

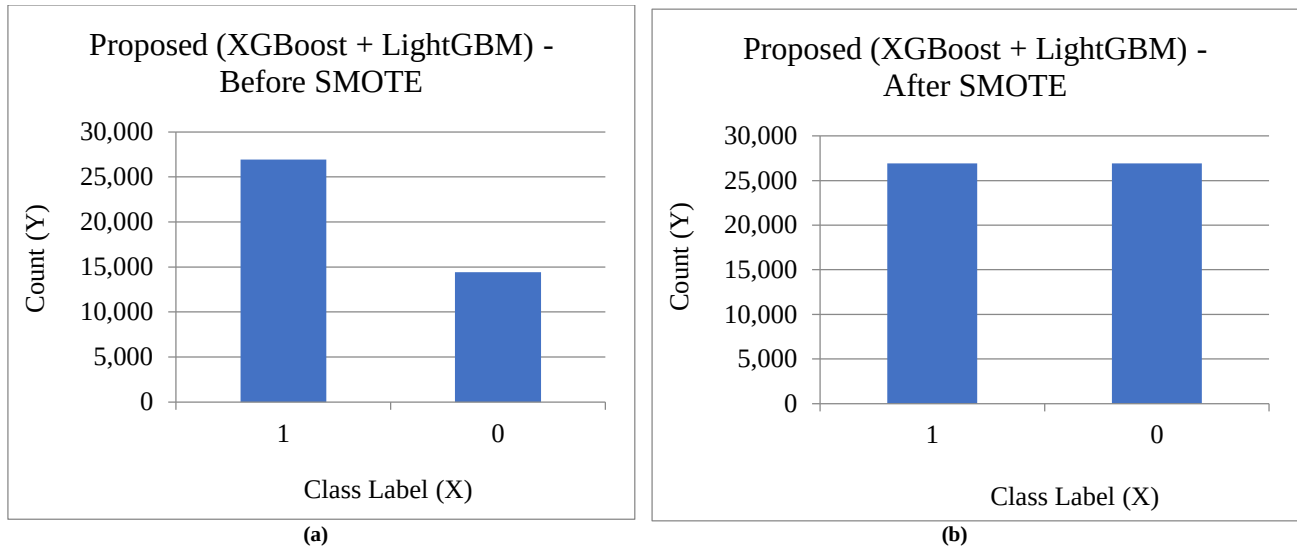


Fig. 7 Class distribution in the original hold-out analysis (a) Before, and (b) After SMOTE for UNSW-NB15.

The features identified by the MI analysis are: byte_ratio, total_bytes, sbytes, dbytes, sttl, dttl, packet_ratio, rate, packet_total, and dpkts. This is illustrated in Figure 8. The

features reflect packet behaviour, volume, and the properties of the protocols, which are very relevant for intrusion detection.

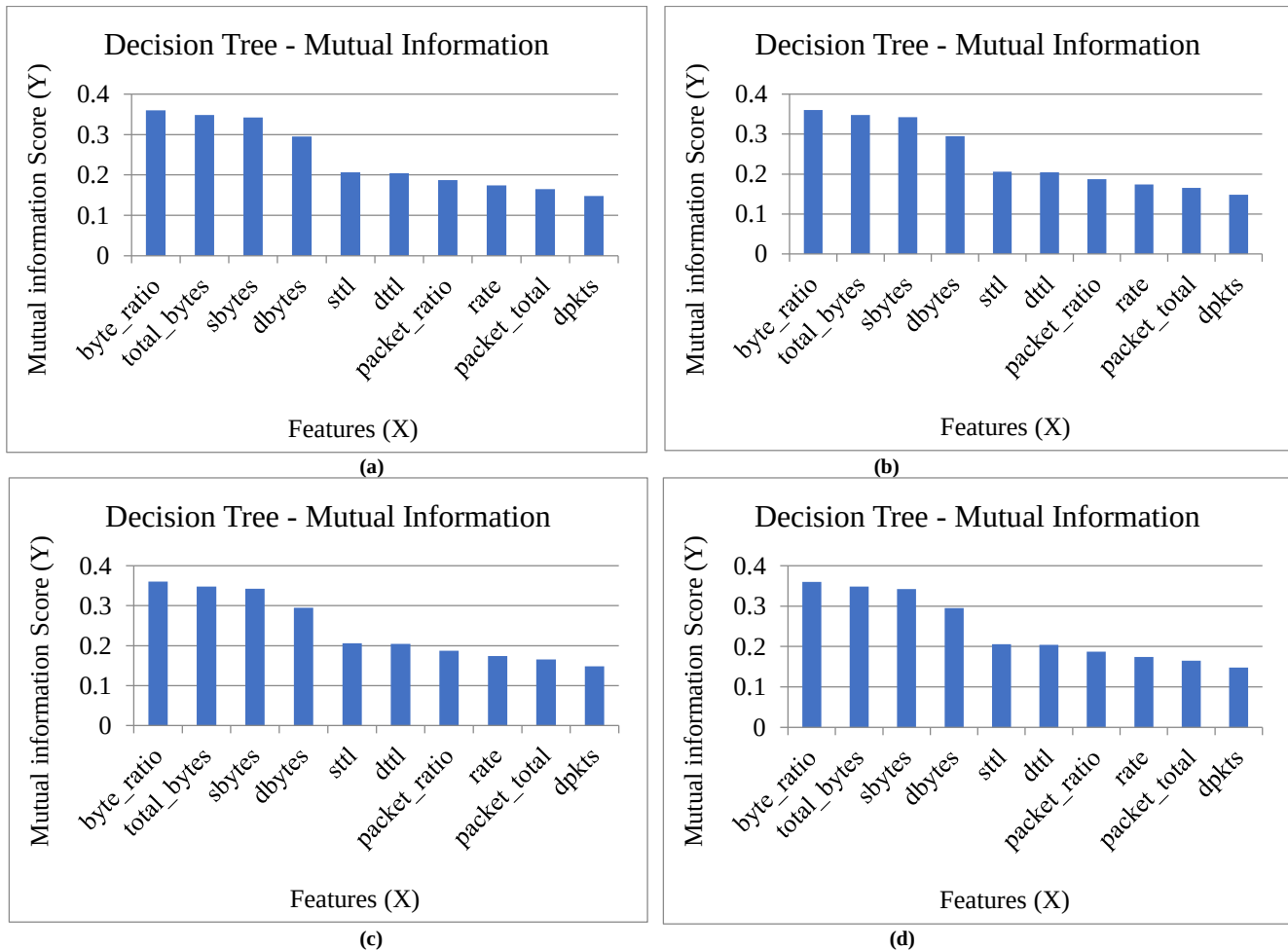


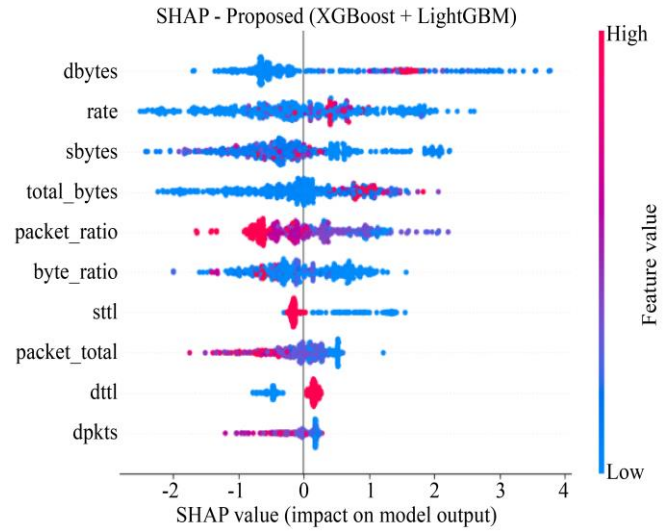
Fig 8. Mutual Information feature ranking for UNSW-NB15 in the original hold-out analysis (a) DT, (b) RF, (c) SVM, and (d) Proposed.

Table 3. Performance comparison on UNSW-NB15 dataset in the original hold-out analysis

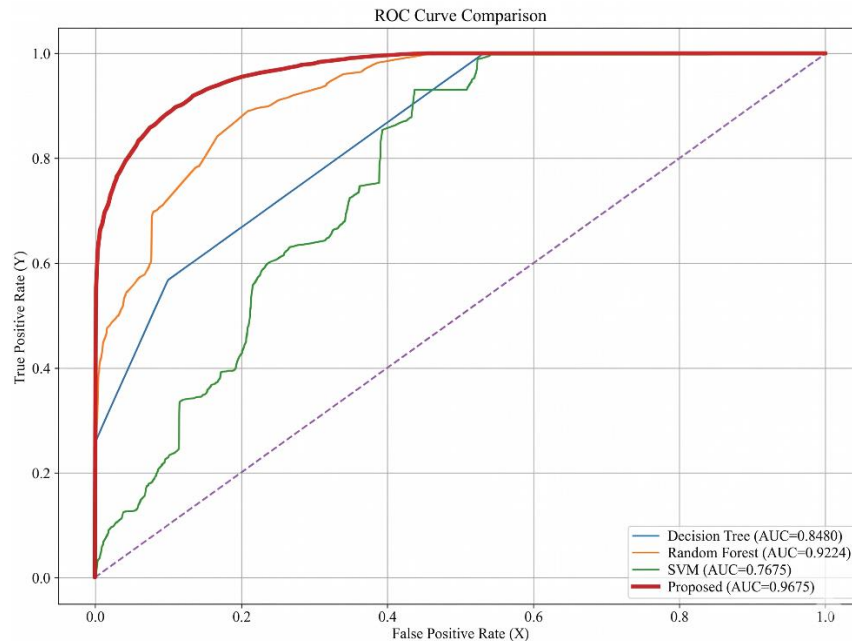
Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
DT	0.7378	0.8583	0.5696	0.6848	0.8519
RF	0.8064	0.7308	0.9710	0.8338	0.9236
SVM	0.7242	0.6514	0.9645	0.7776	0.7662
Proposed (XGBoost + LightGBM)	0.8939	0.8958	0.8914	0.8936	0.9682

Results in Table 3 showed that the proposed framework outperformed all the baseline models in all cases. The proposed ensemble reached an accuracy of 89.39% and a ROC-AUC of 96.82% that performs well at attack and benign classes level with higher discrimination capability. DT achieved the lowest value for recall (56.96%), indicating that it has weak performance in detecting attack traffic. RF and SVM showed high recall accuracy, but with a lower level of precision, suggesting that they had a greater number of false alarms. On the contrary, the proposed framework kept the trade-off between precision and recall in a balanced manner and achieved the highest F1 score among all the evaluated model.

The overall performance of all classifiers in UNSW-NB15 was decreased compared to CICIDS2017. This is expected behaviour due to the more diverse attack patterns and realistic traffic distributions in UNSW-NB15. However, the proposed ensemble still possessed good classification performance and thus performed well in various intrusion detection scenarios. The SHAP analysis in Figure 9 additionally indicated that dbytes, rate, total_bytes, packet_ratio and byte_ratio were the most important predictors in the prediction process. All of these properties help to differentiate malicious network activity from normal network activities.

**Fig. 9 SHAP summary plot for UNSW-NB15 in the original hold-out analysis**

The ROC and PR curves presented in Figures 10 and 11 further confirm the effectiveness of the proposed approach. Generally, the ensemble outperforms the baseline classifiers for all the thresholds, which is indicative of its performance to achieve a high detection rate and low false classifications.

**Fig. 10 ROC curve comparison on UNSW-NB15 in the original hold-out analysis**

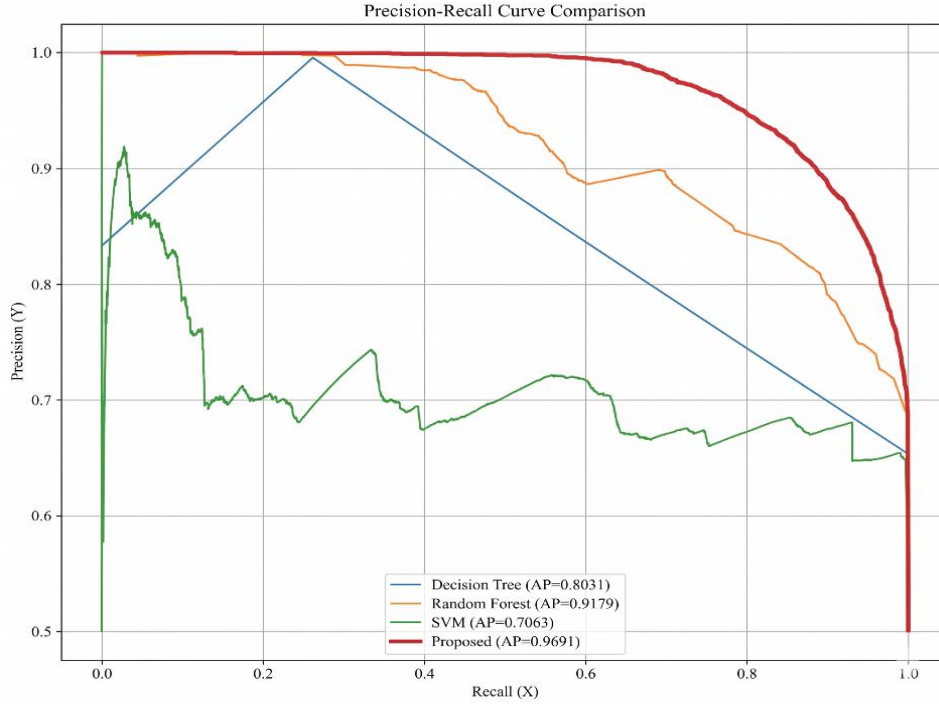


Fig. 11 Precision-Recall curve comparison on UNSW-NB15 in the original hold-out analysis

4.3. Statistical Analysis

Each model configuration was evaluated using stratified five-fold cross-validation with random seed 42. For every metric, the arithmetic mean and sample standard deviation were calculated from the five outer-fold values.

The 95% confidence interval was estimated by using two-sided Student's (t)-distribution by Equation (21).

$$CI_{95\%} = \bar{M} \pm t_{0.975, K-1} \frac{s}{\sqrt{K}} \quad (21)$$

Where $\bar{M}$ is the mean metric value, s is the sample standard deviation, and $K = 5$ is the number of folds. Fold-wise F1-scores were compared using the paired two-sided Wilcoxon signed-rank test. The minimum two-sided exact Wilcoxon probability was limited, and the results were interpreted together with effect magnitude rather than as stand-alone evidence, as only five-fold-level pairs were available. Exact McNemar tests were performed using paired out-of-fold predictions. McNemar's test assesses whether the observations correctly classified by only one of the two models differ significantly. Statistical significance was assessed at $\alpha=0.05$.

Table 4. Five-fold performance of the proposed ensemble

Dataset	Accuracy	Precision	Recall	F1-score	False Positive Rate (FPR)	ROC-AUC
CICIDS2017	0.992304 ± 0.000241	0.981806 ± 0.001824	0.975983 ± 0.001280	0.978884 ± 0.000646	0.004046 ± 0.000418	0.999562 ± 0.000023
UNSW-NB15	0.903948 ± 0.002546	0.888137 ± 0.010388	0.835288 ± 0.006409	0.860844 ± 0.002876	0.058156 ± 0.006412	0.968351 ± 0.000725

Table 4 presents the mean and sample standard deviation obtained across the five outer folds for the proposed optimized ensemble on both datasets.

The small fold-wise standard deviations observed on CICIDS2017 indicate stable performance across the five partitions. UNSW-NB15 produced greater variability, particularly for precision and false-positive rate, reflecting its more heterogeneous traffic distribution and more difficult separation between benign and attack observations.

Table 5. Proposed-model 95% confidence intervals

Dataset	Metric	95% confidence interval
CICIDS2017	Accuracy	0.992005–0.992603
	F1-score	0.978082–0.979686
	Precision	0.979541–0.984071
	Recall	0.974394–0.977572
	FPR	0.003527–0.004565
	ROC-AUC	0.999533–0.999591
UNSW-NB15	Accuracy	0.900787–0.907109
	F1-score	0.857273–0.864415

	Precision	0.875239–0.901035
	Recall	0.827330–0.843246
	FPR	0.050194–0.066118
	ROC-AUC	0.967451–0.969251

4.4. Ablation Study and Error Analysis

A component-level ablation study was conducted to examine the individual contributions of MI, RFE, SMOTE, XGBoost, LightGBM, and optimized ensemble aggregation. Every configuration was evaluated using the same stratified five-fold cross-validation partitions and random seed of 42. All preprocessing and data-dependent operations were fitted to the training portion of each outer fold. The evaluated configurations were defined as follows:

- A0: Full proposed framework with MI, RFE, training-fold SMOTE, XGBoost, LightGBM, optimized soft-voting weights, and an optimized decision threshold.
- A1: Proposed framework without MI.
- A2: Proposed framework without RFE.
- A3: Proposed framework without both MI and RFE.
- A4: Proposed framework without SMOTE.
- A5: Standalone XGBoost using the selected features.
- A6: Standalone LightGBM using the selected features.
- A7: XGBoost–LightGBM ensemble using equal weights and a fixed threshold of 0.50.

Table 6 presents the mean and standard deviation obtained across the five CICIDS2017 outer-validation folds.

Table 6. Component-Level ablation results on CICIDS2017 (Mean $\pm$ Standard Deviation)

Variant	Accuracy	Precision	Recall	F1-score	ROC-AUC	FPR
A0 Full optimized ensemble	0.9923 $\pm$ 0.0002	0.9818 $\pm$ 0.0018	0.9760 $\pm$ 0.0013	0.9789 $\pm$ 0.0006	0.9996 $\pm$ 0.0000	0.0040 $\pm$ 0.0004
A1 Without MI	0.9923 $\pm$ 0.0002	0.9820 $\pm$ 0.0019	0.9759 $\pm$ 0.0015	0.9789 $\pm$ 0.0006	0.9996 $\pm$ 0.0000	0.0040 $\pm$ 0.0004
A2 Without RFE	0.9923 $\pm$ 0.0002	0.9818 $\pm$ 0.0018	0.9760 $\pm$ 0.0012	0.9789 $\pm$ 0.0006	0.9996 $\pm$ 0.0000	0.0041 $\pm$ 0.0004
A3 Without MI and RFE	0.9924 $\pm$ 0.0001	0.9823 $\pm$ 0.0011	0.9760 $\pm$ 0.0012	0.9791 $\pm$ 0.0004	0.9996 $\pm$ 0.0000	0.0039 $\pm$ 0.0003
A4 Without SMOTE	0.9907 $\pm$ 0.0007	0.9962 $\pm$ 0.0025	0.9527 $\pm$ 0.0063	0.9739 $\pm$ 0.0021	0.9996 $\pm$ 0.0000	0.0008 $\pm$ 0.0006
A5 Standalone XGBoost	0.9902 $\pm$ 0.0008	0.9607 $\pm$ 0.0060	0.9866 $\pm$ 0.0020	0.9734 $\pm$ 0.0021	0.9995 $\pm$ 0.0000	0.0090 $\pm$ 0.0014
A6 Standalone LightGBM	0.9900 $\pm$ 0.0007	0.9601 $\pm$ 0.0047	0.9865 $\pm$ 0.0022	0.9731 $\pm$ 0.0018	0.9995 $\pm$ 0.0001	0.0092 $\pm$ 0.0011
A7 Equal-weight ensemble	0.9904 $\pm$ 0.0005	0.9614 $\pm$ 0.0043	0.9869 $\pm$ 0.0019	0.9740 $\pm$ 0.0013	0.9996 $\pm$ 0.0000	0.0089 $\pm$ 0.0010

The CICIDS2017 results show that removing MI, RFE, or both produced only minor changes in performance. A3 obtained a higher F1-score than A0, indicating that MI and RFE did not independently increase the mean predictive performance on this dataset. Their contribution is therefore interpreted as dimensionality reduction and the production of a compact feature representation rather than as a guaranteed improvement in classification accuracy.

The SMOTE removal in A4 reduced recall from 0.9760 to 0.9527 and reduced the F1-score from 0.9789 to 0.9739. Although A4 achieved higher precision and a lower FPR, it missed a larger number of attacks. This result demonstrates the contribution of training-fold SMOTE to improving attack sensitivity.

Standalone XGBoost and LightGBM obtained F1-scores of 0.9734 and 0.9731, respectively. Their FPR were 0.0090 and 0.0092, compared with 0.0040 for the proposed ensemble. The equal-weight ensemble achieved an F1-score of 0.9740 and a FPR of 0.0089. Thus, optimization of the voting weights

and decision threshold improved the CICIDS2017 F1-score and reduced false alarms relative to the individual classifiers and the equal-weight ensemble.

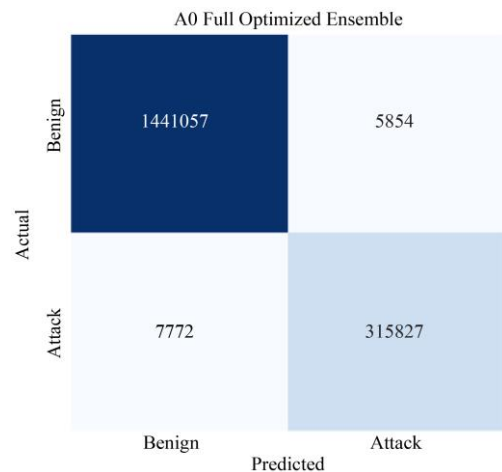


Fig. 12 Out-of-Fold confusion matrix of the proposed optimized ensemble on CICIDS2017

Figure 12 presents the out-of-fold confusion matrix of the proposed model. The counts correspond to a FPR of approximately 0.0040 and demonstrate that the proposed

configuration maintained a low false-alarm rate while detecting most attack observations. Table 7 presents the corresponding ablation results for UNSW-NB15.

Table 7. Component-Level ablation results on UNSW-NB15 (Mean $\pm$ Standard Deviation)

Variant	Accuracy	Precision	Recall	F1-score	ROC-AUC	FPR
A0 Full optimized ensemble	0.9039 $\pm$ 0.0025	0.8881 $\pm$ 0.0104	0.8353 $\pm$ 0.0064	0.8608 $\pm$ 0.0029	0.9684 $\pm$ 0.0007	0.0582 $\pm$ 0.0064
A1 Without MI	0.9042 $\pm$ 0.0020	0.8886 $\pm$ 0.0091	0.8355 $\pm$ 0.0091	0.8612 $\pm$ 0.0028	0.9684 $\pm$ 0.0008	0.0579 $\pm$ 0.0058
A2 Without RFE	0.9048 $\pm$ 0.0018	0.8891 $\pm$ 0.0098	0.8367 $\pm$ 0.0063	0.8621 $\pm$ 0.0015	0.9689 $\pm$ 0.0010	0.0577 $\pm$ 0.0062
A3 Without MI and RFE	0.9058 $\pm$ 0.0017	0.8912 $\pm$ 0.0095	0.8375 $\pm$ 0.0069	0.8634 $\pm$ 0.0017	0.9692 $\pm$ 0.0009	0.0565 $\pm$ 0.0060
A4 Without SMOTE	0.9052 $\pm$ 0.0031	0.9042 $\pm$ 0.0205	0.8213 $\pm$ 0.0163	0.8604 $\pm$ 0.0033	0.9687 $\pm$ 0.0010	0.0484 $\pm$ 0.0122
A5 Standalone XGBoost	0.9018 $\pm$ 0.0034	0.8842 $\pm$ 0.0243	0.8343 $\pm$ 0.0216	0.8580 $\pm$ 0.0021	0.9677 $\pm$ 0.0008	0.0610 $\pm$ 0.0168
A6 Standalone LightGBM	0.9041 $\pm$ 0.0044	0.8918 $\pm$ 0.0305	0.8330 $\pm$ 0.0269	0.8607 $\pm$ 0.0032	0.9686 $\pm$ 0.0009	0.0567 $\pm$ 0.0209
A7 Equal-weight ensemble	0.9043 $\pm$ 0.0041	0.8900 $\pm$ 0.0280	0.8357 $\pm$ 0.0241	0.8613 $\pm$ 0.0028	0.9686 $\pm$ 0.0008	0.0578 $\pm$ 0.0191

The UNSW-NB15 ablation results show that the exclusion of MI and RFE did not reduce performance. A3 produced the highest mean accuracy, F1-score, and ROC-AUC. Consequently, the contribution of MI and RFE on UNSW-NB15 is interpreted as reducing the feature space from 17 input attributes to 12–14 fold-specific features while retaining comparable predictive performance. The findings do not support a claim that feature selection improves every performance metric.

Removing SMOTE increased precision and reduced the FPR, but decreased recall from 0.8353 to 0.8213. This result indicates that SMOTE improved attack sensitivity at the cost of a moderate increase in false alarms. The optimized ensemble obtained a higher F1-score than standalone XGBoost and a comparable F1-score to standalone LightGBM. However, the equal-weight ensemble and feature-selection removal variants obtained higher mean values for certain metrics. Therefore, the UNSW-NB15 results indicate competitive rather than superior performance for the optimized ensemble. The benefits of the complete framework are its compact fold-specific representation, training-derived operating point, reproducible validation procedure, and balance between attack detection and false-alarm behaviour.

As shown in Figure 13, the proposed model correctly classified 26,625 benign flows and 13,033 attack flows. The UNSW-NB15 confusion matrix reflects the greater difficulty of this dataset compared with CICIDS2017. Particularly, the higher number of missed attacks is consistent with the mean recall of 0.8353. Nevertheless, the model maintained a mean precision of 0.8881 and ROC-AUC of 0.9684 across the five folds.

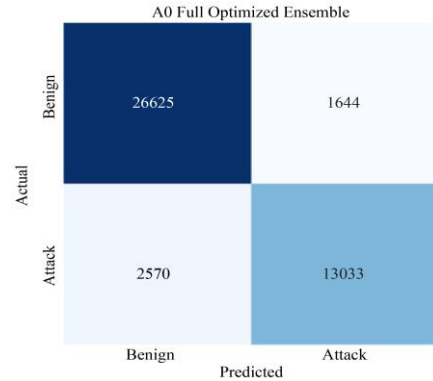


Fig. 13 Out-of-Fold confusion matrix of the proposed optimized ensemble on UNSW-NB15

Overall, the component-level results show that SMOTE influences the trade-off between attack detection and false alarms, while optimized probability aggregation provides a clear improvement over the individual classifiers on CICIDS2017 and competitive performance on UNSW-NB15. MI and RFE contribute to compactness and feature stability rather than consistent gains in predictive accuracy. These findings provide a more precise assessment of the contribution of each framework component.

4.5. Comparative Discussion

The experimental results show the proposed framework achieves good and stable performance on both datasets. The accuracy for the CICIDS2017 dataset was 98.02%, and for the UNSW-NB15 dataset, 89.39% for the framework on CICIDS2017. The proposed ensemble still outperforms all baseline classifiers, although performance has dropped on UNSW-NB15 because attack behaviors are more complicated.

Table 8. Comparison of the proposed framework with previously reported IDS approaches

Study	Dataset	Method/Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	ROC-AUC (%)
Ahmed et al. [12]	CICIDS2017	SHAP-based feature selection with hybrid bagging–boosting and SMOTE	98.47	-	-	98.47	-
Arreche et al. [17]	CICIDS2017	RF with XAI-IDS-selected features	99.00	96.00	96.00	96.00	98.00
Arreche et al. [17]	CICIDS2017	LightGBM with XAI-IDS-selected features	97.00	92.00	92.00	92.00	56.00
Proposed framework	CICIDS2017	MI–RFE + XGBoost–LightGBM + SHAP	0.9923 ± 0.0002	0.9818 ± 0.0018	0.9760 ± 0.0013	0.9789 ± 0.0006	0.9996 ± 0.0000
Proposed framework	UNSW-NB15	MI–RFE + XGBoost–LightGBM + SHAP	0.9039 ± 0.0025	0.8881 ± 0.0104	0.8353 ± 0.0064	0.8608 ± 0.0029	0.9684 ± 0.0007

The comparison with existing approaches in Table 8 demonstrates that the proposed framework achieves competitive performance across both benchmark datasets. On CICIDS2017, Arreche et al. [17] reported an accuracy of 99% using RF with XAI-IDS-selected features, which is slightly higher than the 98.02% obtained by the proposed framework.

However, the proposed approach attained a precision of 98.40%, recall of 97.62%, F1-score of 98.01%, and ROC-AUC of 99.84%, indicating a strong balance between classification performance and discriminative capability. In comparison, the LightGBM configuration reported by Arreche et al. [17] achieved 97% accuracy and an F1-score of 92%, while the proposed ensemble obtained higher values for both measures. Also, the existing Ahmed et al. [12] has got 98.47% accuracy with CICIDS dataset.

The findings show that the addition of SMOTE-based balancing, MI feature selection, SHAP guided interpretability and ensemble with XGBoost-LightGBM are significant in boosting the intrusion detection performance. In addition, the ablation study showed that both feature selection and ensemble learning are key elements of the proposed framework.

Overall, the framework has a high detection capability, good generalization across datasets, and increased interpretability, which is useful for real-world network intrusion detection applications.

5. Conclusion and Future Work

This study proposed an optimized ensemble IDS including the extensive feature selection method, hybrid feature selection, XAI, and comprehensive data preprocessing in network intrusion detection. Finally, the proposed framework integrated two features of network traffic: MI-RFE to select the most informative features, and then used an XGBoost–LightGBM soft-voting ensemble to classify the attacks. The framework was tested on the CICIDS2017 and UNSW-NB15 datasets. The results of the experiments showed that the proposed framework consistently outperforms the conventional ML classifiers such as DT, RF and SVM. The framework obtained an accuracy of 98.02% and ROC-AUC of 99.84% on the CICIDS2017 dataset and good performance on the more challenging dataset UNSW-NB15 with an accuracy of 89.39% and ROC-AUC of 96.82%. In addition, the SHAP-based explainability was used to give transparent insights into feature contributions, thus improving the interpretability and applicability of the model. The framework can be further tested with actual network traffic streams, with the inclusion of DL architectures, and the extension to IoT and cloud-based intrusion detection environments. Further, adaptive feature selection and online learning mechanisms may be investigated to enhance detection capabilities for new cyber-attacks.

Data Availability Statement

This work uses data from <https://www.unb.ca/cic/datasets/ids-2017.html> and <https://research.unsw.edu.au/projects/unswnb15-dataset>

References

- [1] Ramya Chinnasamy et al., “Deep Learning-Driven Methods for Network-based Intrusion Detection Systems: A Systematic Review,” *ICT Express*, vol. 11, no. 1, pp. 181-215, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Vincent Zibi Mohale, and Ibidun Christiana Obagbuwa, “A Systematic Review on the Integration of Explainable Artificial Intelligence in Intrusion Detection Systems to Enhancing Transparency and Interpretability in Cybersecurity,” *Frontiers in Artificial Intelligence*, vol. 8, pp. 1-10, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]

- [3] C. Rajathi, and P. Rukmani, "Hybrid Learning Model for Intrusion Detection System: A Combination of Parametric and Non-Parametric Classifiers," *Alexandria Engineering Journal*, vol. 112, pp. 384-396, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] R. Rajeshwari, and M.P. Anuradha, "Network Intrusion Detection Using Feature Selection Techniques: Bacterial Forage Optimization Algorithm," *International Journal of Intelligent Engineering and Systems*, vol. 17, no. 5, pp. 1-16, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [5] R. Rajeshwari and M. P. Anuradha, "Intrusion Detection using Dynamic Feature Selection: An Adaptive Bacterial Foraging Method," *SN Computer Science*, vol. 5, no. 6, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [6] Saahira Banu Ahamed Maricar et al., "An Improved Explainable Artificial Intelligence for Intrusion Detection System," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 14s, pp. 108-115, 2024. [[Google Scholar](#)] [[Publisher Link](#)]
- [7] A.L. Samed, and Seref Sagiroglu, "Explainable Artificial Intelligence Models in Intrusion Detection Systems," *Engineering Applications of Artificial Intelligence*, vol. 144, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Marija Gombar, Amir Topalovi, and Mirjana Pejic Bach, "Cost-Aware Lightweight Deep Learning for Intrusion Detection: A Comparative Study on UNSW-NB15 and CIC-IDS2017," *Electronics*, vol. 15, no. 8, pp. 1-28, 2026. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] PratikWaghmode et al., "Intrusion Detection System based On Machine Learning using Least Square Support Vector Machine," *Scientific Reports*, vol. 15, no. 1, pp. 1-23, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [10] Mohammad Nassef, "Boosting Intrusion Detection Against DDoS Attacks using a Feature Engineering-Based Fine-Tuned XGBoost Model," *International Journal on Semantic Web and Information Systems*, vol. 21, no. 1, pp. 1-39, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [11] Muhammad Umer et al., "Network Intrusion Detection Model Using Wrapper-Based Feature Selection and Multi-Head Attention Transformers," *Scientific Reports*, vol. 15, no. 1, pp. 1-15, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Usman Ahmed et al., "Hybrid Bagging and Boosting with SHAP-based Feature Selection for Enhanced Predictive Modeling in Intrusion Detection Systems," *Scientific Reports*, vol. 14, no. 1, pp. 1-32, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Amit Kumar, Vivek Kumar, and Abhay Pratap Singh Bhadauria, "Optimizing Intrusion Detection in Edge Computing Network: A Hybrid ML Approach with Recursive Feature Elimination," *International Journal of Intelligent Engineering and Systems*, vol. 18, no. 1, pp. 1-11, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Shweta More et al., "Enhanced Intrusion Detection Systems Performance with UNSW-NB15 Data Analysis," *Algorithms*, vol. 17, no. 2, pp. 1-16, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Chadia E. L. Asry et al., "Enhancing Cybersecurity: A High-Performance Intrusion Detection Approach Through Boosting Minority Class Recognition," *PLoS One*, vol. 20, no. 3, pp. 1-22, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [16] Aoun E. Muhammad et al., "L-XAIDS: A LIME-based Explainable AI Framework for Intrusion Detection Systems," *Cluster Computing*, vol. 28, no. 10, pp. 1-34, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Osvaldo Arreche, Tanish Guntur, and Mustafa Abdallah, "XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems," *Applied Sciences*, vol. 14, no. 10, pp. 1-41, 2024. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] Dharmaraj R. Pati; et al., "Optimizing Network Intrusion Detection Systems through Ensemble Learning and Feature Selection using the CIC-IDS2017 Dataset," *Informatica*, vol. 49, no. 4, pp. 1-30, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [19] Fiona Lawrence, and Rajesh Nigam, "Enhancing Intrusion Detection Systems with Ensemble Models and Hybrid Feature Selection Techniques," *Journal of Information Systems Engineering and Management*, vol. 10, no. 23s, pp. 937-954, 2025. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Intrusion Detection Evaluation Dataset (CIC-IDS2017), Canadian Institute for Cybersecurity, University of New Brunswick, 2017. [Online]. Available: <https://www.unb.ca/cic/datasets/ids-2017.html>
- [21] The UNSW-NB15 Dataset, University of New South Wales, 2015. [Online]. Available: <https://research.unsw.edu.au/projects/unsw-nb15-dataset>