International Journal of Innovative Research in Engineering
Volume 7, Issue 2 (March - April 2026), PP: 239-243.
https://www.doi.org/10.59256/ijire.20260702029
www.theijire.com
ISSN No: 2582-8746

Intrusion Detection System Using Ensemble Learning and SHAP Explainability

Bibash Basnet[1], Prajwal Rai[2], Subarna Sapkota[3], Bibek Gautam[4]
[1] *Padmashree College, Nilai University, Nepal.*
[2]*Kantipur City College, Purbanchal University, Nepal.*
[3]*Nepal College of Information Technology, Pokhara University, Nepal.*
[4]*Pulchowk Campus, Tribhuvan University, Nepal.*

How to cite this paper:
Bibash Basnet[1], Prajwal Rai[2], Subarna Sapkota[3], Bibek Gautam[4], "Intrusion Detection System Using Ensemble Learning and SHAP Explainability", IJIRE-V7I2-239-243.

Copyright © 2026 by author(s) and Fifth Dimension Research Publication. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Abstract: *Intrusion Detection Systems (IDS) play a critical role in monitoring network traffic and identifying malicious activities that threaten digital infrastructure security. Traditional IDS models struggle with accuracy, interpretability, and reliability as cyber-attacks continue to evolve. This study develops an ensemble learning-based IDS that combines Random Forest and XGBoost algorithms through a Voting Classifier to enhance detection performance and stability. The system was trained, validated, and tested on the NSL-KDD benchmark dataset, which underwent preprocessing including encoding, normalization, and feature selection. SHAP (SHapley Additive exPlanations) was integrated to provide feature-level interpretability for each classification, addressing the black-box nature of ensemble models and increasing transparency. The system achieved strong performance with 99.26% training accuracy, 98.97% validation accuracy, and 98.86% test accuracy. SHAP analysis identified key contributing features such as src_bytes, dst_bytes, and flag, enabling deeper understanding of the model's decision-making process. Comparative analysis with previous ensemble-based IDS studies demonstrated superior performance over CNN-RF hybrids and RF-SVM-LIME explainable ensembles. Overall, the developed IDS represents an accurate and interpretable solution for detecting malicious traffic and supporting informed cybersecurity decisions.*

Key Words: *Intrusion Detection System, Ensemble Learning, Random Forest, XGBoost, SHAP, Explainability, NSL-KDD, Network Security, Machine Learning.*

I.INTRODUCTION

The growing reliance on computer networks and interconnected systems in the modern digital age has elevated cybersecurity to a strategic priority for organizations worldwide. Network intrusions, whereby unauthorized individuals or malicious actors gain access to computer networks, represent significant threats to digital infrastructure. Such intrusions can be perpetrated by external attackers or insiders and may aim to steal sensitive information, disrupt services, or install backdoors for future attacks. The most prevalent intrusion techniques include malware infections, port scanning, brute-force attacks, phishing, and DoS/DDoS attacks, which can result in substantial financial losses, legal liability, information breaches, and reputational damage [1].

To identify and respond to such illegitimate activities, organizations have implemented Intrusion Detection Systems (IDS), which monitor network activities to identify suspicious behavior or policy violations [2]. Intrusion detection systems are generally categorized into signature-based systems, which detect established attack patterns, and anomaly-based systems, which identify deviations from normal traffic behavior. Nevertheless, conventional IDS, particularly those based on fixed rules and signatures, often fail to recognize novel and advanced attacks due to high false positive rates and lack of explainability.

Machine learning, especially ensemble learning, represents one of the most promising directions for addressing these limitations by combining multiple models to enhance accuracy and robustness. This study develops an Intrusion Detection System using ensemble learning methods that combines Random Forest and XGBoost through a Voting Classifier, enhanced with SHAP (SHapley Additive exPlanations) for interpretability. The IDS is trained on the NSL-KDD benchmark dataset and performs binary classification to distinguish between normal and malicious traffic. The system achieves 98.86% test accuracy while providing transparent feature-level explanations, enabling security analysts to understand and trust classification decisions.

II.RELATED WORK AND RESEARCH GAP

Intrusion Detection Systems play a vital role in monitoring the network traffic and detecting unauthorized activities. Traditional IDS based on the signature methods have limited capability in identifying advance attacks. Ensemble learning

methods have significantly improved IDS detection capabilities. This section reviews related ensemble based approaches for intrusion detection.

B.Yogesh and Reddy (2022) conducted a comparative study emphasizing the Random Forest Classifier for intrusion detection [3]. Using the NSL-KDD dataset, they compared Support Vector Machine, K-Nearest Neighbors, Logistic Regression, Naive Bayes, and Random Forest. The Random Forest Classifier achieved approximately 99% accuracy, significantly exceeding other algorithms such as Naive Bayes (85%) [4]. Confusion matrix analysis revealed very low false detection rates, demonstrating that ensemble techniques achieve superior performance compared to individual classifiers.

Doost et al. (2025) presented a hybrid system combining Convolutional Neural Networks and Random Forest [5]. CNN extracted essential features from network datasets while eliminating redundant data, and Random Forest performed final classification. The hybrid CNN-RF model attained 97.4% accuracy and 98.4% precision on KDD99 and UNSW-NB15 datasets. Performance comparison showed that CNN-RF excelled in accuracy, precision, and recall while maintaining balanced execution time.

Patil et al. (2022) proposed an ensemble system combining Random Forest, Decision Trees, and Support Vector Machines with explainable AI [6]. The ensemble model achieved 96.25% detection rate on CICIDS-2017 dataset, higher than individual models. Local Interpretable Model-Agnostic Explanations (LIME) identified contributing features, illustrating that ensemble learning with explainable AI increases both detection capability and transparency.

Although machine learning-based intrusion detection systems have demonstrated encouraging performance, several limitations remain. Single classifier models struggle to capture complex attack patterns, while CNN-based hybrid approaches require high computational resources, limiting real-time deployment. Although Random Forest models have shown strong performance, their integration with gradient boosting algorithms such as XGBoost has not been extensively explored for intrusion detection. Furthermore, most existing explainable IDS solutions rely on LIME, which provides only local explanations, whereas SHAP offers more comprehensive feature-level interpretability. To address these gaps, this study proposes a voting-based ensemble combining Random Forest and XGBoost with SHAP-based explanations to improve detection accuracy, interpretability, and computational efficiency.

III. METHODOLOGY

The development of the Intrusion Detection System using ensemble learning follows a systematic workflow that includes data collection, preprocessing, feature selection, model training, explainability integration, and deployment. This section details every stage of the methodology.

A. Dataset Description

The proposed system is tested using the **NSL-KDD dataset**, which is a popular intrusion detection benchmark test data set. NSL-KDD is a refined variant of KDD99 dataset, which was meant to remove duplicate records and minimize bias in the evaluation. It is made up of network traffic logs that are classified as normal traffic or any of the following attacks: Denial of service (DoS), Probe, User to root (U2R), and Remote to local (R2L). The records have 41 features of network traffic such as basic, content-based and traffic attributes. The data is publicly accessible and widely implemented to facilitate reasonable comparison with the current research on IDS.

B. Data Preprocessing and Balancing

Data preprocessing is a critical step in transforming raw network traffic data for machine learning. Missing values and duplicate records were identified and removed to ensure data integrity. Categorical features including protocol_type, service, and flag were converted into numerical form using label encoding. A binary label was created with 0 representing normal traffic and 1 representing attack traffic. Numerical features were normalized using Standard Scaler to ensure balanced feature contribution. To address class imbalance, stratified splitting was applied during dataset partitioning to preserve the original class distribution across training, validation, and test sets, improving generalization performance.

C. Feature Selection:

Feature selection was performed to identify the most important features for intrusion detection. A Random Forest Classifier was trained on the complete feature set to determine feature importance scores, indicating each feature's influence in differentiating between normal and attack traffic.

Based on these scores, the top five features were selected:src_bytes, dst_bytes, flag, dst_host_srv_count, and diff_srv_rate. Focusing on these features improves efficiency by reducing computational complexity and prediction time. All subsequent models were trained using only the selected features.

D. Ensemble Model Development

The ensemble model combines Random Forest and XGBoost. Random Forest was selected for its robustness and resistance to overfitting through bagging and random feature selection, configured with 150 decision trees. XGBoost was chosen for its gradient boosting capability, where successive trees correct previous errors to capture complex patterns.

The models were combined using a soft Voting Classifier, where class probabilities are averaged to generate the final prediction. This approach leverages the stability of Random Forest and the optimization strength of XGBoost, achieving better accuracy and generalization than individual models.

E. SHAP Integration for Explainability

To address the black-box characteristics of ensemble models and provide transparency in decision-making, SHAP (SHapley Additive exPlanations) was integrated into the system. SHAP relies on game theory concepts and calculates the contribution of each feature to a given prediction. SHAP values are computed to display feature importance for each classification made by the ensemble model, showing which features influenced the prediction toward either Normal or Attack.

F. Evaluation and Deployment

Model performance was evaluated using accuracy, confusion matrix analysis, precision, recall, and F1-score metrics. The model was deployed using a Tkinter-based graphical user interface (GUI) optimized for CPU-only execution with error handling for invalid inputs.

IV. RESULT AND DISCUSSION

The Intrusion Detection System with ensemble learning and SHAP explainability was tested on the NSL-KDD dataset. The system combines Random Forest and XGBoost classifiers using a Voting Classifier strategy to classify network traffic as either normal or attack. This section presents experimental findings and performance analysis.

A. Performance Metrics

Dataset	Accuracy	Precision	Recall	F1-Score
Training	99.26%	99.31%	99.21%	99.26%
Validation	98.97%	98.82%	99.13%	98.97%
Testing	98.86%	98.69%	99.03%	98.86%

Table I Performance Metrics Across Data Splits

The ensemble model demonstrated exceptional performance across all evaluation datasets. Training accuracy reached 99.26%, validation accuracy was 98.97%, and testing accuracy was 98.86%. These results indicate that the model generalizes well to unseen data with minimal performance gap between training and testing. The small performance difference confirms that the model does not suffer from significant overfitting and maintains substantial stability and reliability across different data subsets. The consistently high performance across all three datasets confirms that the ensemble method successfully combines the complementary strengths of Random Forest and XGBoost to generate robust and dependable predictions for network intrusion detection.

B. Confusion Matrix Analysis

The confusion matrix indicates strong classification performance, with low false positive (1.3%) and false negative (1.7%) rates. This balanced error distribution confirms that the ensemble model is not biased toward either class and effectively detects malicious traffic while minimizing false alarms.

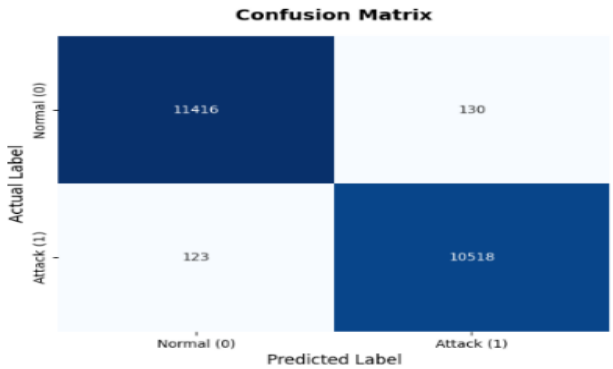

Figure 1: Confusion Matrix.

C. ROC Curve and AUC Score

Figure 2 presents the ROC curve plotting false positive rate versus true positive rate across classification thresholds. The curve is steeply inclined toward the upper-left corner, indicating excellent discrimination between normal and attack traffic. The Area Under the Curve (AUC) was 0.999, approaching perfect classification. This high AUC value indicates that the ensemble model maintains high sensitivity while preserving high specificity across threshold values.

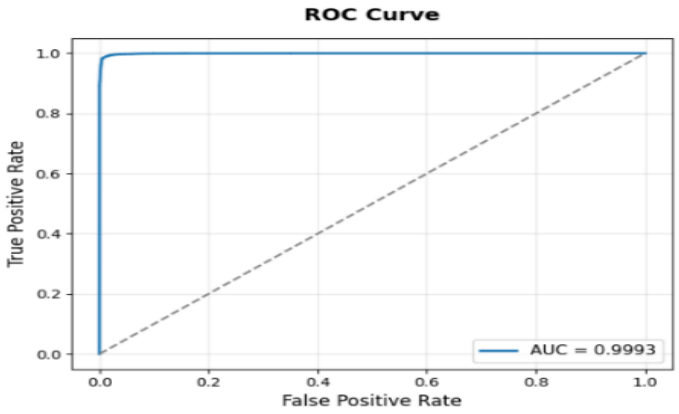

Figure 2:ROC curve of the proposed ensemble IDS.

D. Comparative Analysis

Comparative analysis shows that the proposed ensemble achieves 98.86% test accuracy, outperforming CNN-RF hybrids (97.4%), RF-SVM-LIME ensembles (96.25%), and Naive Bayes approaches (85%). In addition to higher accuracy, SHAP provides more comprehensive interpretability than LIME by offering both local and global feature importance. These results demonstrate that the proposed system achieves an effective balance between detection performance and explainability.

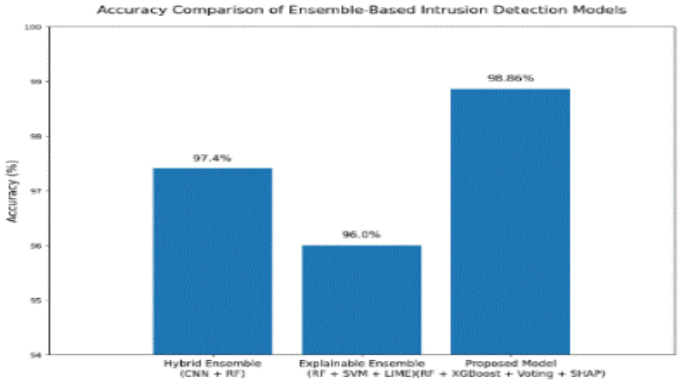

Figure 3: Comaprison with previous Ensemble Models.

E. Discussion

The results demonstrate that combining Random Forest and XGBoost through soft voting effectively leverages their complementary strengths for intrusion detection. The minimal performance gap across datasets confirms balanced learning without overfitting. SHAP integration enhances interpretability, enabling security analysts to understand and trust model decisions.

V. CONCLUSION

This paper presents an Intrusion Detection System using ensemble learning that incorporates Random Forest and XGBoost classifiers with soft voting to achieve high accuracy without compromising interpretability. The system was trained on the NSL-KDD dataset, achieving 98.86% test accuracy, 98.69% precision, and 99.03% recall. Feature selection reduced 41 network attributes to 5 most significant features, decreasing computational complexity by 87.8%. SHAP explainability provides feature-level interpretations enabling security analysts to understand classification reasoning. ROC analysis demonstrated excellent discrimination capability (AUC 0.999). Comparison with previous approaches showed superior performance over CNN-RF (97.4%) and RF-SVM-LIME (96.25%) models. The Tkinter-based GUI enables both manual and batch predictions with SHAP explanations, making advanced intrusion detection accessible without machine learning expertise. This work demonstrates that combining ensemble learning with explainable AI creates effective and interpretable network security solutions.

VI.FUTURE WORK

Though the proposed ensemble-based intrusion detection system achieved high accuracy on the NSL-KDD dataset, future research should test the model on more recent and realistic intrusion detection datasets such as CICIDS2017, CSE-CIC-IDS2018, or UNSW-NB15 to better represent modern network traffic and attack patterns. Retraining and validating the model on these datasets would enhance its applicability. The framework can also be adapted to IoT and edge-based network environments through environment-specific training and deployment at gateway nodes. Integration of online or incremental

learning mechanisms would improve adaptability to changing traffic patterns, while further enhancement of explainability and adversarial resistance would strengthen its suitability for real-world cybersecurity applications.

VIII.CONFLICT OF INTEREST

The authors declare that they have no conflict of interest related to the publication of this manuscript.

References

1. M. A. Hossain and M. S. Islam, "Ensuring network security with a robust intrusion detection system using ensemble-based machine learning," Array, vol. 19, p. 14, 2023.
2. R. Tahri, Y. Balouki and A. Lasbahani, "Intrusion Detection System Using machine learning Algorithms," ITM Web Conf., vol. 46, p. 4, 2022.
3. B.Yogesh and G. S. Reddy, "Intrusion detection System using Random Forest Approach," Turkish Journal of Computer and Mathematics Education, vol. 13, no. 02, pp. 725 - 733, 2022.
4. M. P. Shirurkar and M. More , "Implementation of the NSL-KDD Dataset to Study the Naive Bayes," Panamerican Mathematical Journal,vol.35, 2025.
5. P. A. doost, S. S. Moghadam, E. Khezri, A. Basem and M. Trik , "A new intrusion detection method using ensemble classification and feature selection," Scientific Reports volume, vol. 15, no. 1, p. 13642, 2025.
6. S. Patil, V. Varadarajan, S. Mazhar, A. Sahibzada, N. Ahmed, O. Sinha, S. Kumar, K. Shaw and K. Kotecha, "Explainable Artificial Intelligence for Intrusion Detection System," Electronics, vol. 11,2022.