Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Explainable Machine Learning-Based Intrusion Detection System Using Random Forest and SHAP for Network Security

Mr. S. Muzibuddin
M. Tech.,(Ph.D)
Assistant Professor
Dept. of CSE (CS)
RGMCET , Nandyal

S. Shaheer
Dept. of CSE (CS)
RGMCET , Nandyal

G. Sai Jahnavi
Dept. of CSE (CS)
RGMCET , Nandyal

S. Vara Prasad
Dept. of CSE (CS)
RGMCET , Nandyal

Abstract: Network infrastructures are extremely dynamic and sophisticated with the growing rapidity of the Internet, cloud computing services, Internet of Things gadgets, and high-speed communication technologies.This has greatly advanced the rate and complexity of cyber-attacks. Intrusion Detection Systems (IDS) are important in tracking the network traffic and detection of any malicious activities that can threaten confidentiality, integrity and availability of systems.Conventional IDS systems like signature-based systems are based on the existing attack pattern and are effective in the identification of known threats.Nevertheless, they cannot identify the zero-day and developing attacks because they do not have previous signatures.Machine Learning (ML) based IDS systems have become increasingly potent alternatives, as they acquire the pattern based on the network traffic data and can extrapolate to identify previously unknown threats. Although they have a better detection mechanism, the ML models usually take the form of black-box.They give output of classification with no explanation of how they arrived at those decisions.Explainability is a crucial concern in the field of cybersecurity operations since, in this context, security analysts need to ensure they identify alerts, examine the trends of attacks, and defend themselves in their mitigation measures.Lack of transparency decreases the trust and makes it hard to deploy in the real-world setting. Hence, this paper suggests Explainable Machine Learning-based Intrusion Detection System, which combines classification features with interpretability technologies.

Indexed Terms: Intrusion Detection System (IDS), Machine Learning, Explainable Artificial Intelligence (XAI), Random Forest, SHAP (SHapley Additive exPlanations), Network Security, Cybersecurity, CICIDS2017 Dataset, Network Traffic Analysis.

I. INTRODUCTION

Network infrastructures have become very dynamic and complicated with the blistering development of the Internet, cloud computing, IoT devices, and high-speed communication technologies. This growth has greatly elevated the number of and the level of cyber-attacks. Intrusion Detection System (IDS) is crucial in the monitoring of the network traffic and recognizing malicious activities that jeopardize confidentiality, integrity and availability of systems. The conventional signature-based IDS works well in identifying already known attacks but does not identify zero-days and emerging threats because they are based on known attack patterns. In order to counter this weakness, machine learning (ML)-based IDS models have been proposed. Such systems are learned based on past network information and are capable of identifying unfamiliar attacks. The vast majority of ML-based IDS models, though, are black-box models, that is, they offer high detection accuracy but do not describe the decision-making process. There must be explainability in cybersecurity operations. Security analysts have to be familiar with alerts, attack pattern analysis and make effective mitigation decisions. Transparency deficiency minimises credibility and restricts the practical implementation. Consequently, the proposed research is an Explainable Machine Learning-based Intrusion Detection System, which uses explanatory methods alongside a high classification.

- **The Problem**

The fast growth of networked systems and cloud environments has augmented the opportunities of attacks like DDoS, brute-force, botnets, infiltration, and ransomware. Signature-based IDS have the disadvantage where they are unable to detect the unknown attacks and on the other hand, the ML-based anomaly detection systems though efficient, is not transparent. Majority of ML-based IDS frameworks offer unjustifiable prediction, which makes it challenging to confirm alerts and optimize the detection policy. False positive rates that are high also add extra

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

burden to security teams as well as efficiency in operations. Therefore, the primary issue that is discussed in this paper is the absence of system transparency and interpretability in ML-based IDS systems. It is necessary to have a powerful IDS model that can not only ensure a high detection rate but also give significant contexts to its predictions. It is necessary to include Explainable Artificial Intelligence (XAI) methods to enhance the trust, decision-making, and real-time cybersecurity performance.

- **Our Contribution**

The following are the major contributions of the proposed research: the proposing study formulates an explainable and high-performance IDS based on the use of the Machine Learning:

- Combined Explicable IDS Framework: The systematic IDS architecture is built, which incorporates the preprocessing of data, feature engineering, supervised classification, and explainability in a single pipeline. Interpretability is considered to be a constituent element and not an added value.
- Feature Selection and Correlation Analysis: The

feature selection and correlation analysis is carried out to eliminate redundancy to increase computational efficiency. Determining attributes with a high level of correlation improves the performance of the model and interpretability.

- Random-Forest Ensemble Learning: Random Forest is used as a classifier to enhance the detection and decrease overfitting. Ensemble learning improves generalization and is stable in performance in the case of different attacks such as DDoS, brute-force attacks, and infiltrations attacks.
- SHAP-based Integration of Explainability: SHAP (SHapley Additive exPlanations) is incorporated in order to solve the black-box problem of ML models. The most significant features that determine the overall predictions have been identified in the global explanations. Local reasons explain why a given instance of network is a malicious or a normal network.
- False Positive Reduction and Operational Trust: The explainability mechanism is useful in assisting the analysts in comprehending misclassifications and in refining the security policies. This minimizes the false positives, increases the efficiency of operations, and increases the confidence in automated detection systems.

- **Content of the Paper**

The rest of this paper will be structured in the following manner: Part II provides the background research and other research in the area Intrusion Detection System, the use of machine learning to detect intrusions, and explainable Artificial Intelligence to be used in cybersecurity. It is in this section that the current methods have been examined and their weaknesses pointed out after which the research gap in this work is established. Section III explains the proposed system architecture, in which the architecture outlines the general framework, pipeline of data flow, the preprocessing mechanisms, the combination of the classification model and clarify the overall architecture along with the explainable module. Section IV describes the dataset description and preprocessing approach, such as data cleaning, feature selection, data normalization, and dealing with the issue of class imbalance. Section V describes the process of machine learning model development, such as the choice of the Random Forest classifier, the hyperparameters tuning, the training process, and the specifics of the implementation. Section VI gives the performance evaluation metrics and the results of the experiment. This part evaluates the outcomes of accuracy, precision, recall, F1-score, and the confusion matrix to confirm that the proposed system is effective or not. Section VII presents the explainability model of SHAP, which gives global and local explanation of model predictions.The techniques of feature contribution analysis and visualization are elaborated. Section VIII gives a detailed discussion of findings,

strengths, weaknesses, and implication of the proposed explainable IDS framework into practice. Lastly, IX brings the paper to a close and presents the research directions of the future such as real-time deployment, adaptive learning integration, and more sophisticated explainability techniques of next-generation cybersecurity systems.

II LITERATURE REVIEW

Intrusion Detection System (IDS) has been changing greatly taking into consideration the recent development in network technologies and the swift rise in cyber threats. Scientists have been working on the detection accuracy, false positive reduction, and flexibility to new attacks. This section provides a review of traditional IDS techniques, machine learning techniques, deep learning techniques, and Explainable Artificial Intelligence (XAI) in cybersecurity.

A. Traditional Intrusion Detection Systems: Early IDS The two common categories of IDS were signature and anomaly based systems. Signature-based IDS identify attacks by comparing network traffic with previously known attack pattern. They will work on familiar threats and have low false positives. Nonetheless, they are not able to identify the new attacks or zero-day attacks because they are based on the signatures that are already in place. Anomaly based IDS counter this weakness by modeling normal network behavior and detecting abnormal behavior as a possible threat. The common statistical detection techniques were threshold-based and rule-based. Whereas they are able to identify any unknown attacks, anomaly-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

based systems tend to raise high false positives since the way of normal traffic pattern might have changed. Hybrid IDS use both methods to enhance coverage of detection, but introduce complexity and computational needs in the systems.

B. Intrusion Detection based on machine learning: Because of the development of data-driven methods, the use of Machine Learning (ML) became common in IDS. Some of the algorithms used to classify network traffic include Decision Tree, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest. Decision Trees are easy to interpret and to understand yet it can overfit complicated data. SVM works better in higher dimensional space with a lot of sensitivity to parameters. KNN is easy and computationally costly with large data. An ensemble learning algorithm,Random Forest, has demonstrated good performance in intrusion detection. It uses a combination of decision trees which make it less prone to overfitting and enhances generalization. Experiments indicate a high precision on the benchmark datasets like NSL-KDD and CICIDS2017. Most ML-based IDS models are black-box systems in spite of their high detection rates reducing interpretability.

C. Deep Learning in Detection of Intrusion: Deep learning is an automatic method of learning complicated features in the form of raw network data.

Convolutional Neural Networks (CNN) identify the spatial features, whereas Recurrent Neural Networks (RNN) and long-short-term memory (LSTM) networks learn temporal features of sequence traffic data. These techniques are very sensitive to detection, particularly when the data to be analyzed is large. They, however, demand much computational resources and extensive training data. Besides, they have a multi-layered structure that lowers their transparency, and are hard to read in security-relevant contexts.

D. Cybersecurity Explainable Artificial Intelligence: The more complicated became ML and deep learning models, the more interpretability was required. Explainable Artificial Intelligence (XAI) is an attempt to make the predictions of a model interpretable. Explainability in cybersecurity aids analysts in validating their alerts, analyzing patterns of attack, and supporting the actions taken to forestall attacks. LIME and SHAP are model-agnostic methods that are used extensively. The SHAP, which works on the basis of cooperative game theory, gives a prediction with contribution scores on every feature. It provides: Global accounts - finding general significant aspects. Local explanations- explaining individual predictions. New researches incorporate SHAP to ML-based IDS to enhance transparency without impacting the accuracy in a considerable way. Explainability is, however, not considered as a part of the IDS framework in most works but an extra analysis tool.

E. Research Gap Analysis:Though new progress has been made in ML and deep learning-based IDS, it still has a number of challenges. Majority of the systems are more concerned with accuracy of detection rather than its interpretation. Black-box models do not allow real-world implementation where accountability and compliance must be established. Also, systems that rely on anomaly tend to generate high false positives, and this raises the burden on security personnel. It is evident that there is a need to have an accurate and interpretable IDS framework. The system is expected to have both global and local explanations, minimise false positives and aid real-time monitoring. In pursuit of this gap, the proposed study combines ensemble learning and SHAP-based explainability to come up with a transparent and high-performing system of intrusion detection.

III METHODOLOGY

The Explainable Machine Learning-based Intrusion Detection System suggested is systematic and designed in the manner that proper detection and decision making can be achieved. The methodology amalgamates data processing, feature optimization, and supervised learning as well as explainability methods into one framework. The steps of the process are aimed at enhancing the performance of detection and preserving interpretability and low false positives. All the workflow steps entail the following:

- Dataset Collection
- Netowork Traffic input
- Data PreProcessing
- Feature Selection
- Model Training(Random Forest)
- Performance Evaluation
- SHAP Explainability
- Final Prediction(Normal/Attack +Explainability)

➢ **Data Set Collection:**

The quality of the dataset that is used to train and test a Machine Learning-based Intrusion Detection System (IDS) largely determines its performance. The CICIDs2017 dataset is utilized in this paper due to its realistic nature of a network traffic and the presence of current cyber-attack cases. The dataset includes normal (benign) traffic and various forms of attacks which include DDoS, brute-force, botnet attacks, web attacks and infiltration. It was gathered through several days which is useful to capture the various network behaviors and patterns of attacks. The network records consist of over 80 features, including the number of packets and the number of bytes, the duration of a flow, the protocol type, the flag information, etc. The data is already labelled and thus can be used in supervised learning. Application of CICIDS2017 guarantees sound training, ability to compare it with other models of IDS and realistic estimation of the working of the system.

S. No	Attack Category
1	Benign(Normal)
2	DDOS
3	Brute Force

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

4	Botnet
5	Port Scan
6	Web Attack-SQL Injection
7	Web Attack-XSS
8	Infiltration
9	Heartbleed

Fig : Attack Categories Present in CICIDS2017

➢ **Network Traffic Input:**

The first step in the construction of the proposed Explainable IDS is the Network Traffic Input stage. In this research, the network traffic data was borrowed on the CICIDS2017 dataset in Kaggle. The dataset was initially designed by the Canadian Institute of Cybersecurity to model the real enterprise network behavior and various cyberattack models. This study packet data. Each flow can be characterized as the uses flow-based data that is an alternative to raw communication between a source and a destination at a certain time. This makes computation much simpler than packet-based analysis. The data set is saved in CSV format and it has a significant number of numerical values including flow length, packets, data transfer rate and protocol related values. Every record would be normal (benign) or one of the type of attacks: DoS, DDoS, brute-force, or port scanning. It is at this phase that the data is loaded into the machine and features and labels are separated appropriately. It is relevant to ensure that data are organized correctly, as it directly influences the model training, the level of detection, and SHAP-based explainability.

➢ **Data Preprocessing**

Preprocessing of the data is a very important phase in the creation of a reliable and explainable Intrusion Detection System. The raw CICIDS2017 network flow dataset documentation has very high-dimensional traffic features, class imbalance, and statistical variations that are prone to a detrimental impact on the learning of the model. Thus, a preprocessing pipeline is organized to improve the quality of data, provide numerical stability, and better classification. The goal of preprocessing is to convert raw records of the traffic flows into a clean and normalized and machine-learnable format and leave meaningful behavioral features that can be used to explain behavior using SHAP.

1. Data Cleaning and Data Integrity Check:

The first stage of data analysis is to detect data inconsistencies including duplicate records, missing values and infinite values created during flow feature identifiers like Source IP, Destination IP and Timestamp, calculation. The duplication of entries is eliminated to avoid biased learning. The elements with values of infinity (Inf, -Inf) that are obtained in computation through division are replaced with a null value, and are then treated in a statistical imputation. Checks on the rows that contain too many missing values result in their removal to ensure the integrity of the data set.Such non-behavioral identifiers as Source IP, Destination IP and Timestamp are dropped since they are not easy to generalize and can lead to overfitting. Because the IDS is expected to identify specific behavioral patterns as opposed to the specific hosts, the storage of the IP addresses can create data leakage.

In the preprocessing of data, the inconsistencies that occur are the case of duplicates, missing values, and infinite values (Inf-Inf) which are created in the course of processing flow features and are dealt with with caution. Redundant records are eliminated to avoid biased learning and infinite records are legitimized with null records and dealt with with proper statistical imputation tools. The datasets are of high quality and reliability because records with too many missing values are removed. Also, non-behavioral identifiers, including Source IP, Destination IP, and Timestamp are eliminated to prevent excessive overfitting and data leakage to ensure that the model acquires generalized traffic behavior patterns and not particular host information.

S.No	Issue identified	Cleaning Action	Purpose
1	Duplicate Records	Remove duplicate rows	Encoded to numeric values
2	Missing values(NaN)	Remove duplicate rows	Ensure dataset completeness
3	Infinite Values (Inf,-Inf)	Replaced with 0 or median	Ensure dataset completeness
4	IP address columns	Dropped Source &Destination IP	Prevent overfitting
5	Timestamp Column	Removed	Not required for classification
6	Categorical Labels	Encoded to numeric values	ML model compatibility

Table: Data Cleaning Operations Performed on CICIDS2017 Dataset

2. Target Label Transformation:

The CICIDS2017 dataset contains multiple attack categories. For intrusion detection, a binary classification framework is adopted:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

- Benign → 0
- Attack (all categories) → 1

This simplifies the detection objective to identifying malicious traffic irrespective of attack subtype. Label encoding ensures compatibility with supervised machine learning algorithms.

3. Feature Scaling and Normalization:

Variations in magnitude are high with respect to network traffic features. As an example, flow duration can be in millions, but flag counts can be 010. In the absence of scaling, the process of learning might be dominated by features with higher values. In order to make feature contribution similar, Min-Max normalization is employed, which is as follows:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Such transformation normalizes all the values of features between 0 and 1. Normalization is known to improve the convergence of models, eliminate numeric instability, and also improve the interpretability of SHAP feature contributions.

4. Handling Class Imbalance:

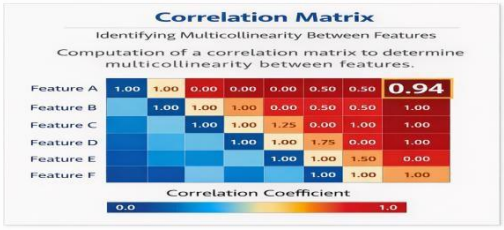
The sample of CICIDS2017 includes considerably more benign traffic than samples of attacks. Unbalanced data can lead to the biasing of the majority class of the classifier. To mitigate this issue: It uses stratified train-test splitting.In the training of Random Forest, class weighting is used.

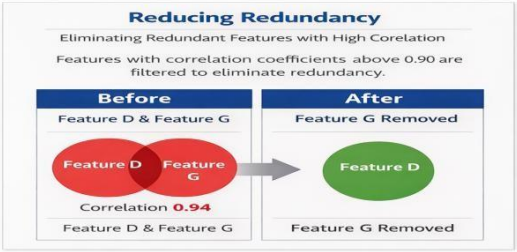
Class	Sample Proportion
Benign	~85%
Attack	~15%

Table: Example Class Distribution

➢ **Redundancy Analysis and Correlation:**

Features with high correlation can bring redundancy and add complexity to the model with no increase in performance. The computation of a correlation matrix is done to determine multicollinearity between features. Features that have correlation coefficients above 0.90 are considered and filtered on to eliminate redundancy without losing prediction ability.





➢ **Feature Selection:**

Dropped features that displayed a lot of correlation (correlation = +0.90).

Eliminated redundant and meaningless characteristics.

The help of the feature importance was provided to reduce top relevant features.

Reduced dimension, more effective model. Output:

Fast training and more understandable.

➢ **Model Training:**

Divide data into 20 and 80 training and testing respectively.

Random forest was used as the classifier. Acquired

normal traffic and attack traffic patterns. Tuning of the

application.

Output: Correct Classified IDS model.

➢ **Performance Evaluation:**

Evaluated using:

Accuracy

Precision

Recall F1-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Score

Confusion Matrix

Focused on high Recall and low False Positives.

Results: Reliable intrusion detection results.

- **SHAP Explainability:**

SHAP prediction of a model.

Identified key factors which influence attacks.

Provided international and personal description of predictions.

Result: Open and trustful artificial intelligent system.

- **Final Prediction**

New network data are given to trained model. Model

forecasts: Attack or normal.

SHAP is the description of the reason why the prediction was made.

Label of prediction and probability of outputs.

End Product: Intelligible and correct AI-based Intrusion Detection System.

IV. CONCLUSION & FUTURE ENHANCEMENT

The proposed project will create an AI-based Intrusion Detection System which will use machine learning to label network traffic as either normal or malicious. The feature selection is used to eliminate duplicate and high correlation features and enhance efficiency and lower the computational complexity. It is found that the trained model performs well in terms of accuracy, precision, recall, and F1-score and therefore is able to detect effectively with low false positives. The system brings the additional benefits of SHAP explainability, which helps get a clear understanding of the contribution of the features, making the model more of a black-box than a visible and reliable cybersecurity solution. The system can be further developed in future with addition of sophisticated deep learning modules like CNN and LSTM to identify a more sophisticated and dynamic patterns of attacks. It may also be implemented in real time network setup to monitor all time and respond to threats more quickly. More enhancements can also involve automated response systems, zero-day attack detection and connecting it with cloud and enterprise security systems to establish a scalable, intelligent, and responsive defense system.

V. REFERENCES

[1] L. Yang and A. Shami, "IDS-ML: An open source code for Intrusion Detection System development using Explainable Artificial Intelligence for Intrusion Detection Systems," *IEEE Access*, 2024.

[2] Y. Zhang and J. Liu, "SHAP-Based Explainable Intrusion Detection Using Ensemble Learning," *Computer Networks*, 2024.

[3] S. Sharma and R. K. Jha, "Advanced Hyperparameter Optimization Techniques for Network Intrusion Detection Systems," *Engineering Applications of Artificial Intelligence*, 2025.

[4] M. Alazab, W. Sun, and R. Abawajy, "Deep Learning Based Explainable Intrusion Detection in Software- Defined Networking," *IEEE Transactions on Network and Service Management*, 2025.

[5] T. Nguyen and H. Lee, "Efficient Feature Selection for High-Dimensional Intrusion Detection Using Meta- heuristic Algorithms," *Expert Systems with Applications*, 2024.

[6] A. Singh and P. Gupta, "Real-Time Intrusion Detection With XAI: A SHAP Perspective," *Journal of Information Security and Applications*, 2025.

[7] W. Lee and S. J. Stolfo, "A Framework for Constructing Features and Models for Intrusion Detection Systems," *ACM Transactions on*

[8] *Information and System Security*, 2023.

[9] X. Wang, P. S. Yu, and X. Zhang, "Feature Engineering and Explainability for Network Anomaly Detection," *Information Sciences*, 2025.

[10] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.