An Explainable AI-Based Hybrid Framework for Botnet-Driven DDoS Attack Detection Using SHAP and Ensemble Learning

Prof. Ruksar Fatima

Khaja Bandanawaz University, College in Culbarga, India

Abstract

DDoS attacks performed through botnets constitute a significant threat to modern-day network systems. Such attacks aim at making the services available by sending huge volumes of malicious traffic which consumes bandwidth and resources and renders the services unavailable to individuals, businesses, or governments. While machine learning and deep learning have proven to be effective in identifying such types of cyber-attacks, most approaches are usually black-box techniques that hinder the work of cybersecurity experts. To remedy the situation, a new technique called SHAP-EnsembleIDS is proposed. It uses Explainable AI as a method of detecting botnet DDOS activities. In doing so, the method relies on multiple ensemble classifiers which include Random Forest, XGBoost, and Gradient Boosting. Preprocessing the data, normalizing features, and determining feature importance enhance representation of network traffic. Moreover, SHapley Additive exPlanations are included in the process to enable the provision of interpretable results showing the influence of traffic features on detection results. The methodology is applied to a publicly available benchmark dataset that includes legitimate and botnet-based DDoS traffic. The experiments carried out using the hybrid ensemble show that it has capabilities of distinguishing malicious traffic from normal ones. The comparative evaluation carried out using Accuracy, Precision, Recall, and F1 Score confirms the effectiveness of the framework. Therefore, it can be suggested that the proposed technique combining Ensemble Learning and SHAP is a viable approach for future intrusion detection systems.

Keywords:

Botnet; Distributed Denial-of-Service Attack; Explainable Artificial Intelligence; SHAP; Ensemble Learning; Intrusion Detection System; Network Security; XGBoost; Random Forest.

1. Introduction

With the expanding adoption of cloud technologies and Internet of Things (IoT) ecosystems and connected devices, the number of possible attack surfaces is increasing significantly. One of the complex and sophisticated security challenges faced by modern networked environments is the Distributed Denial-of-Service (DDoS) attack launched by botnets, which involves sending large amounts of traffic to the victim by using millions of hijacked devices [16], [18], [21].

However, signature-based IDS faces certain drawbacks because it cannot detect new attack variants and emerging malicious traffic types. The application of machine learning and artificial intelligence as alternatives can be used to overcome these challenges because they help analyze network traffic in order to find suspicious patterns in a timely manner [21], [23], [24]. In addition to that, the development of the methods of ensemble learning increases the accuracy of intrusion detection due to the combination of several models [11], [14], [15]. The application of Random Forest, Gradient Boosting, XGBoost, and other techniques proves to be quite effective for detecting botnet activity and DDoS attacks [13], [20], [25].

However, machine learning models used in intrusion detection applications often remain black boxes without much information about how predictions were made. Predictions' unexplainability reduces trust of analysts and increases the difficulty of conducting investigation processes. Thus, explainable AI becomes an option that could solve this problem and improve the transparency of learning models. Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive Explanations (SHAP) approaches gain popularity due to the ability to explain the involvement and contribution of specific features. [4], [5].

Recent studies have highlighted the fusion of explainability mechanisms with intrusion detection systems for interpretable cybersecurity analytics. Explanable AI-based frameworks have been applied to network anomaly detection, security of Internet of Things, and classification of DDoS attacks, whose results show the feasibility of SHAP-based feature attribution to explain the attack behaviour of and the model decision-making [6]–[10], [16], [17]. Furthermore, researchers have begun to use explainability together with ensembles to gain high detection results with interpretable architectures [8] [25] [26]. Though most of the techniques focus on classification accuracy but doesn't provide much investigation into which feature of the network traffic is contributing in identifying attacks. There is a scarcity of explainable ensemble frameworks to detect DDoS attacks driven by botnets.

In light of these limitations, this proposed paper presents an Explainable AI-Based (XAI) Hybrid Framework for Botnet-Fueled DDoS Attack Detection Using SHAP and Ensemble Learning (SHAP-EnsembleIDS). An Ensemble Learning Framework consisting of RF, Gradient Boosting, and XGBoost classifiers is proposed to improve Detection reliability and classification performance. To improve transparency, the

Manuscript received July 5, 2026
Manuscript revised July 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.7.1

SHAP is incorporated as post-hoc explainability mechanism for quantifying effect of network traffic features upon Attack prediction while providing interpretable insights to cybersecurity analysts. The Framework not only correctly classifies DDoS Attack types caused by bots, but also explains classifications.

The main input from this study is encapsulated below.

- Designing a hybrid ensemble learning framework that combines Random Forest, Gradient Boosting, and XGBoost for botnet-driven DDoS attack detection.
- Incorporation of SHAP-based explanations for decision making in classifying attacks together with transparency.
- This means analyzing features and attributes to determine those aspects of network traffic that are mostly correlated with the botnet's DDoS attacks.
- Evaluation will be done using well-known measures for classification such as accuracy, precision, recall, and F1 score.
- The proposed approach was tested in practice to assess its effectiveness in comparison with other intrusion detection systems.

The rest of the paper is structured as follows. Section 2 reviews the existing state of the art in XIDS, ensemble-based IDS, and botnet-induced DDoS detection approaches. Section 3 describes the system model and the problem formulation. The next section contains information about SHAP-EnsembleIDS. It provides the description of experiments and their performance analysis. Finally, Section 6 concludes the paper and outlines the future work.

2. Related Work

With the rapid evolution of botnet controlled DDoS attacks, there has been a substantial amount of research work toward intelligent intrusion detection system that can accurately and reliably detect malicious activities. Recent advances in machine learning, ensemble learning, and explainable aI are improving the pattern analysis by a cybersecurity framework that enables better identification of multi-stage threats.

In order to detect many anomaly, machine learning-based Intrusion Detection Systems are well-known. These system prevent hacker attack. These are good at making unauthorized traffic automatically. According to Sarker [21], He et al. [23], Muneer et al. [24] and Fatima et al. [31], there is an increasing trend of using artificial intelligence techniques in cybersecurity applications for overcoming limitations of conventional signature-based detection. Graph-based and DL architectures were used for detecting botnet and DDoS attacks, allowing extraction of complex traffic dependencies and attack characteristics in large-scale networking environments [18]–[20].

The ensemble learning concept has shown great promise in improving classification by combining predictions from multiple models. The foundational work on ensemble methods set the theoretical groundwork for composer classifiers to improve robustness and generalization [11], [14], [15]. Machine techniques such as Random Forest [11], Gradient Boosting [12], and XGBoost [13] have become popular for intrusion detection as they are capable of processing high-dimensional network traffic as well as capturing nonlinear relationships among features. Contemporary cybersecurity studies have shown that rather than individual classifiers, ensemble-based architectures deliver superior performance. They perform better than individual classifiers, especially in dynamic and heterogeneous networks [7], [8], [20].

Achieving a high level of detection accuracy is still important, but as Artificial Intelligence is used more often in cybersecurity, issues of transparency and trust have arisen. Numerous ML Model are black-box systems which are unable to clarify rationale for their diagnosis of attacks. In light of this challenge, a significant research direction has emerged. It aims to enhance interpretability so as to help with informed security decision-making. Lundberg and Lee [1] proposed SHAP while Ribeiro et al. [2] proposed LIME, which are both popular tools to explain predictions. Molnar [3] developed a foundation for Interpretable machine learning and Explainable model analysis.

Recent studies have taken a look at the use of explainability methods in IDS. Gaspar et al. [4] included study of the applicability of SHAP and LIME for interpretation of the intrusion detection models based on neural networks. Arreche et al. [5] studied explainable frameworks for network intrusion detection and highlighted value of transparent decision-making process in cybersecurity operations. Using explainable AI-based techniques, researchers have also tackled DDoS attack detection, IoT security, anomaly detection, and intelligent intrusion detection systems. Feature attribution techniques revealed critical attack indicators, improving analyst confidence in automated decisions [6], [9], [10], [16], [17]. Recent studies have attempted to fuse explainability with advanced learning systems. Wali and collaborators [25] investigated a Random Forest-based explainable intrusion detector. Almolhis [26] coupled attention mechanisms with explainable visualization techniques for cybersecurity analytics. The work of Taheri et al. [27] and Gholamrezazadeh and Montazerolghaem [29] utilized the concepts of explainability in federated intrusion detection environments with privacy and distributed learning.

Researchers SHAP [28] and Haque et al. [30] have demonstrated the potential of SHAP-based explainability for network anomaly detection and cyber risk assessment, respectively.

Nonetheless, the existing studies still have some limitations. Numerous methods focus mainly on predictive accuracy which do not provide much insight into the nature of networks and their features. In addition, many of the explainable intrusion detection Framework focus on single-model architectures whose robustness may be restricted against varying botnet attacks. Ensemble learning applied in combination with SHAP-based interpretability for botnet-driven DDoS attack detection has received little attention. As a result, a Framework is still needed to deliver not only accurate classification performance but also strong decisions and transparency at the feature level. In response to the identified need, this study presents an Explainable AI-Based Hybrid Framework that uses ensemble learning and SHAP-based feature attribution for reliable and interpretable detection of botnet-driven DDoS attacks.

3. System Model and Problem Formulation

3.1 System Model

SHAP-EnsembleIDS projected framework detects the botnet caused DDoS attacks and explains its classifications. The four major entities a model consists of include the attacker, botnet nodes, network traffic monitoring infrastructure, and target network service.

In a typical attack scenario, a botnet-driven ddos attack, an attacker remotely manages a network of compromised devices command control mechanism controls a botnet of compromised devices. The devices infected with malware are called botnet nodes or malware nodes. They generate coordinated traffic towards the target server or network resource. Heavy traffic causes network bandwidth, process resources, and system capacity consumption, which degrades service availability to genuine users [16], [18], [19].

This network monitoring component continuously captures the incoming traffic flows created as a result of smartphone activity and extracts their traffic characteristics including but not limited to the number of packets involved, duration of the connection, protocol used, the properties of the source and destination nodes, and statistics on the flow of data. The proposed SHAP-EnsembleIDS framework involves the processing of features extracted previously by the ensemble learning classifiers to ascertain whether the observed traffic pattern is a botnet DDoS or normal network behaviour. In efforts to enhance transparency, SHAP-based explainability mechanisms identify which traffic features contribute to which prediction, thus helping security analysts to understand attacks and model behaviours [1], [4], [16].

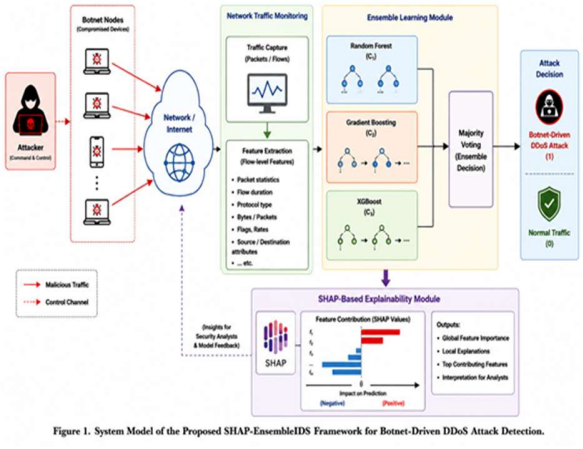

Figure 1. System Model of the Proposed SHAP-EnsembleIDS Framework for Botnet-Driven DDoS Attack Detection.

3.2 Problem Formulation

Let

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

For all (N), x_i represent the feature vector of the i^{th} network flow wh$_i$le yi represents the respective class label.

Each feature vector could be given as

$$x_i = \{f_1, f_2, f_3, \dots, f_m\}$$

where (m) refers to the number of features extracted from network traffic.

The aim of classification is creating mapping function

$$F: X \to Y$$

such that

$$Y = \{0,1\}$$

where..

- (0) indicates Standard Network Traffic
- (1) illustrates traffic from DDoS attack by botnet.

The main goal is to minimize errors during classification and maximize detection of malicious traffic instances. This indicates

$$\hat{y} = F(x_i)$$

that is the predicted class label produced by the ensemble learning system.

As well as accurate classification, the framework should offer interpretable explanations that quantify the weight of individual features upon final prediction. As a result, problem has two simultaneous objectives.

1. Precise attack identification.
2. Elucidated decision-making.

3.3 Mathematical Representation of Explainable Ensemble Detection

The framework utilizes different classifiers that include Random Forest, Gradient boost, and XGBoost. The final prediction is acquired through majority voting through individual classifiers.

Let

$$C_1(x), C_2(x), C_3(x)$$

show the output of classifiers Random Forest, Gradient Boosting, and XGBoost respectively

The final classification outcome is indicated

$$E(x) = MajorityVote[C_1(x), C_2(x), C_3(x)]$$

where

$$E(x) \in \{0,1\}$$

by the representation of ensemble prediction.

SHAP is used to make the model interpretable, explaining the contribution of each feature to the prediction. The SHAP value of feature (j) is given

$$\phi_j$$

by and the prediction is explained by

$$f(x) = \phi_0 + \sum_{j=1}^{m} \phi_j$$

where.

- (ϕ_0) denotes baseline prediction,
- (ϕ_j) signifies contribution of feature (j),
- Total number of features is denoted by (m).

SHAP values provide global and local explanations by quantifying how each traffic feature drives the classifier's decision making. This helps security analysts uncover which network features are most associated with DDoS attacks launched by botnets and improves the credibility of the intrusion detection as well [1], [4], [5], [28]. The proposed framework is grounded in a mathematical formulation that utilizes ensemble learning and xAI for reliable and interpretable botnet-driven DDoS attack detection.

4. Proposed Sharp-Ensembled Framework

The proposed SHAP-EnsembleIDS framework utilises the power of ensemble learning and XAI for highly accurate yet interpretable detection of botnet-driven DDoS attacks. The proposed framework uses SHAP to provide transparency around the decision-making of model, building on a basic classifier. Unlike most conventional intrusion detection frameworks that solely focus on classification performance, the proposed framework offers two-fold advantages – attack detection and model transparency.

The six major phases of the framework include the data acquisition, preprocessing, feature extraction and selection, ensemble-based attack classification, majority voting decision fusion, and SHAP-driven explanation generation. At the start, the raw network traffic data is taken from intrusion detection dataset and transformed into a machine learning-friendly structured format. The preprocessing phase conducts data cleansing, normalization, and feature refinement to enhance data quality and reduce redundancy. The next step involves utilising combination of ML classifiers called RF, GB and XGBoost to analyze traffic related to botnet DDoS attacks. The determination of final classification is obtained by combining the outputs of the individual classifiers through majority voting mechanism. The detection robustness yields through an ensemble strategy which enhances the use of multiple learning algorithms yielding complementary strengths while reducing individual model biases. The contribution of each traffic feature towards each predicted class will be evaluated in this module. The produced SHAP values supplied global as well as local explanations which enabled the security analyst to determine the network attributes most associated with malicious traffic behaviour. Fig-2 shows general design of SHAP-EnsembleIDS framework projected by research.

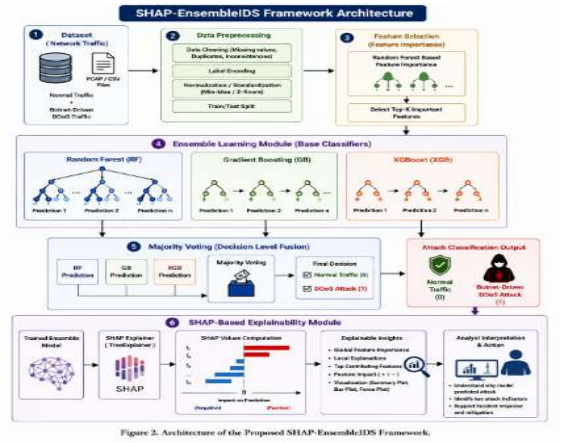

Figure 2. Architecture of the Proposed SHAP-EnsembleIDS Framework for Botnet-Driven DDoS Attack Detection and Explainable Decision-Making.

4.1 Dataset Description

The proposed SHAP-EnsembleIDS framework's performance is assessed on NSL-KDD benchmark intrusion detection dataset, which is widely utilised for assessing efficacy of network IDS. Dataset was developed to overcome limitations associated with the KDD Cup 1999 dataset, including excessive redundant records and biased data distributions to provide a useful benchmark for machine learning-based cybersecurity research.

The NSL-KDD dataset comprises normal and malicious network traffic instances with different traffic attributes related to connection behavior, protocol information, traffic statistics, and network service characteristics. These capabilities allow for the recognition of distinct attacks, including DoS attacks, R2L attacks, and U2R attacks. The dataset is appropriate to evaluate the attack detection models as well as analyzing feature contributions that are connected with malicious traffic behaviour because botnet driven DDoS attacks have the same traffic characteristics as large-scale DoS activities.

Each network traffic record includes 41 features which represent numerical as well as categorical attributes obtained from the network connection. These features capture protocol type, service, connection duration, packets, source and destination behavior, and anomalous traffic behavior among others. Having both normal and attack traffic instances allows a supervised learning model to learn the differentiating patterns associated with the attack.

split of 80:20 was done between training and testing for ease in training and evaluation. According to the training set, the ensemble classification model is learnt, testing set is utilised to assess classification performance and SHAP-based explanations for model prediction. The salient features of NSL-KDD dataset that this study used are summarized in Table 1.

Table 1. Characteristics of the NSL-KDD Dataset Used for Evaluating the Proposed SHAP-EnsembleIDS Framework

Parameter	Value
Dataset	NSL-KDD
Training Samples	125,973
Testing Samples	22,544
Total Samples	148,517
Number of Features	41
Normal Traffic Records	67,343
Attack Traffic Records	81,174
Attack Categories	DoS, Probe, R2L, U2R
Class Labels	Normal / Attack
Train-Test Split	80:20

The framework provides a complete analysis of the attack detection abilities as well as model interpretability. In addition, the dataset is characterized by a wide diversity of attributes that may be employed for the evaluation of SHAP explanations in order to determine which network traffic attributes are significant to detecting botnet-induced DDoS attacks. The ensemble models have been created using the Statistics and Machine Learning Toolbox of MATLAB.

4.2 Data Pre-Processing

The key point comes with the preparation of the network traffic data for classification. This is important since it will help improve the quality, consistency, and reliability of the results. Intrusion detection datasets may contain duplicated records, categorical features, inconsistencies in feature scaling, and unnecessary details that may affect the performance of the machine learning models. Therefore, the SHAP-EnsembleIDS approach uses a pre-processing pipeline that increases not only the efficiency of ensemble learning but also the explainability analysis.

In particular, the dataset was checked for missing values, duplicates, and inconsistencies in the recorded network traffic. Duplicated records were removed to avoid any biases and overfitting to the learning algorithm. Categorical features such as protocol type, service type, and connection status were coded numerically using label encoding in order to apply ResNet14 binary classification.

Following data transformation, feature normalization was applied to ensure that all numeric attributes contribute similarly toward model training. Due to the fact that network traffic features can have different numerical ranges, normalizing them prevents features from dominating and leads to effective convergence of the classifier. This research employed Min-Max normalization to scale the values of different features to a [0, 1] range.
The procedure for normalization can be stated as.

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

where.

- (x_i) represents original feature value,
- (x_{min}) denotes minimum value of feature,
- (x_{max}) denotes maximum value of feature,
- (x_i') represents normalized feature value.

After normalization, dataset was divided in 2 sections. Training set was 80% of total data although the testing set was 20%. The subset used to build the ensemble learning model is training subset while the performance evaluation and the SHAP-based explanation is done using the testing subset. By keeping the two separate, it's possible to objectively assess the framework and interpret the predictions from the proposed model reliably.
Pre-processing produces consistent features of improved quality, which positively outputs the feature selection, ensemble classifier and explainers (in that order) respectively. The illustration presented in Figure 3 showcases the overall data pre-processing workflow of SHAP-EnsembleIDS framework.

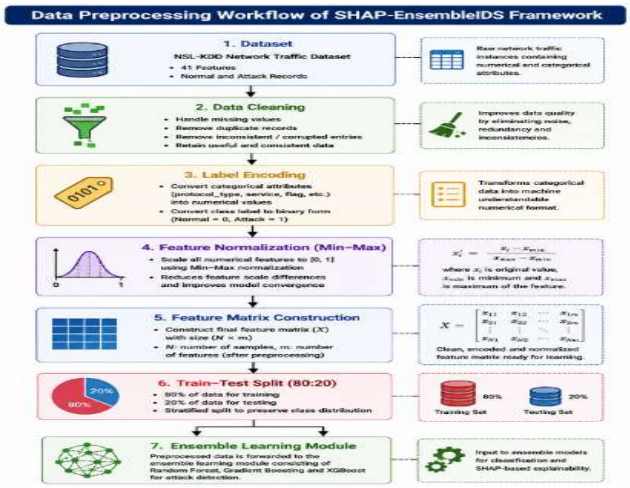

Figure 3. Data Pre-processing Workflow of the Proposed SHAP-EnsembleIDS Framework Showing Data Cleaning, Label Encoding, Feature Normalization, and Dataset Partitioning.

4.3 Ensemble Learning Module

The core detection unit of the proposed framework SHAP-EnsembleIDS is the ensemble learning module. Ensemble classifiers increase accuracy, robustness, and generalization through the combination of results from multiple models. Ensemble approaches reduce variance and protect results from being compromised by the idiosyncrasies of any single classifier through the integration of results from different learning approaches.

The framework integrates three complementary ML classifiers, RF, GB, and Extreme Gradient Boosting (XGBoost) which helps to build a hybrid ensemble detection model. Due to their proven effectiveness on high-dimensional data, including capability for capturing nonlinear relationships and identify complex patterns, several classifiers were chosen with respect to [11]–[13], [20].

4.3.1 Random Forest Classifier

Random Forest can be defined as an ensemble based learning method which creates several decision trees from randomly selected subsets of the samples and features that are used for training. The last prediction is estimated through the majority voting mechanism of the individual trees. By enhancing classification stability and minimizing overfitting, this method maintains good level of prediction performance for intrusion detection.

Prediction function of RF is given as

$$RF(x) = Mode\big(T_1(x), T_2(x), \dots, T_n(x)\big)$$

where.

- $(T_i(x))$ signifies prediction generated by (i^{th}) decision tree,
- This notation (n) represents the total count of trees present.
- $(RF(x))$ signifies final Random Forest prediction.

4.3.2 Gradient Boosting Classifier

Gradient Boosting is a sequential ensemble method which creates weak learners, whereby each subsequent model works to address the prediction errors made by previous ones. Model is able to better capture the model's complex traffic characteristics through this optimization process used to optimize attacks [12].

The Gradient Boosting framework may be depicted as.

$$GB(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$$

where

- $(h_m(x))$ denotes the weak learner at iteration (m),
- (γ_m) represents the corresponding weight,
- (M) denotes the total number of boosting stages.

The strategy of sequential learning improves classification performance for Gradient Boosting and identifies malicious network traffic.

4.3.3 Extreme Gradient Booting (XGBOOST)

XGBoost improves upon conventional boosting techniques by incorporating regularization and fast gradient learning to improve computational speed and accuracy. Its ability to perform parallel computations and comprehensive feature selection makes it very effective for intrusion detection in big datasets of network traffic. [13].

The XGBoost objective function is specified as.

$$Obj = \sum_{i=1}^{N} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where.

- $(L(y_i, \hat{y}_i))$ signifies classification loss,
- $(\Omega(f_k))$ signifies regularization term,
- (K) signifies the number of decision trees.

The regularizer controls complexity for better generalization and reduced overfitting.

4.3.4 Hybrid Ensemble Construction

The hybrid ensemble model is constructed by merging the results of RF, GB, and XGBoost. Fusion of various classifiers allows the system to leverage the advantages of different learning schemes, Boosting immunity to misclassification. Each classifier analyzes the normalized network traffic features independently and produces a prediction of whether the traffic is normal or botnet-driven DDoS.

The majority voting mechanism is built around the ensemble learning module to combine individual classifier predictions from various models to arrive at the final attack classification decision. The ensemble learning module structure of projected SHAP-EnsembleIDS framework is shown in Fig-4.

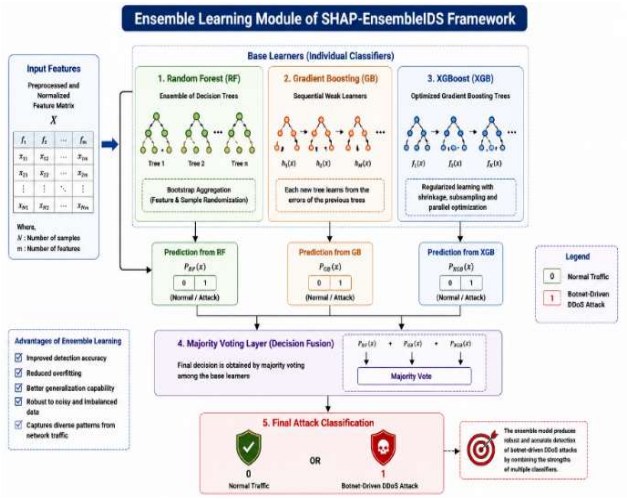

Figure 4. Ensemble Learning Module of the Proposed SHAP-EnsembleIDS Framework Integrating Random Forest, Gradient Boosting, and XGBoost Classifiers for Botnet-Driven DDoS Attack Detection.

4.4 Majority Voting Classification

After generating prediction for each person with RF, GB and XGBoost classifiers, final classification is made with a majority voting method. Majority voting is a common decision fusion technique in ensemble learning, integrating predictions from multiple classifiers and selecting the class for which the most votes are received. As described in Rohel et al. (2020), this approach reduces individual model errors and enhances factors to meet the classification uncertainty [14], [15]. Under the proposed SHAP-EnsembleIDS framework, each base classifier independently analyzes the preprocessed network traffic features to produce a binary prediction (normal traffic vs. botnet-driven DDoS attack traffic).

$$C_{RF}(x), C_{GB}(x), C_{XGB}(x)$$

Let y_h, y_gb, y_xgb denote the predictions from Random Forest, Gradient Boosting, and XGBoost classifiers
Each classification tool gives a two-fold output.

$$C_i(x) \in \{0,1\}$$

where.

- indicates commonplace network activity.
- (1) indicates DDoS attack traffic from a botnet.

The final ensemble prediction, obtained through majority voting, is defined as follows.

$$MV(x) = Mode[C_{RF}(x), C_{GB}(x), C_{XGB}(x)]$$

where.

- Final ensemble classification result denoted as MV(x).
- Prediction of Random Forest classifier is as CRF (x).
- Prediction made by Gradient Boosting classifier is CGB(x).
- CXGB(x) represents the forecast made by the XGBoost classifier.
- $Mode(.)$ denotes the class label receiving the highest number of votes.

Alternatively, we can express the decision to vote as.

$$MV(x) = \{1, if \sum_{i=1}^{3} \quad C_i(x) \geq 2, 0, otherwise.$$

At least two classifiers must provide the same classification decision in order to classify this traffic instance as a Botnet DDoS Attack.

There are several benefits to using majority voting. To begin with, it enhances the framework's resilience against the prediction errors of individual classifiers. It enhances the model robustness when analyzing diverse network traffic patterns. To enable the ensemble architecture to use the complementary strengths of RF, GB, and XGBoost, thus improving detection performance with improved generalization capability over different attack scenarios.

The output of the majority voting mechanism on classification serves as an input to the SHAP-based explainability module. After that, contribution of individual network traffic features to ensemble prediction is evaluated utilizing SHAP values, facilitating transparent and interpretable decision-making about the attack detection.
The Figure 5 depicts the decision fusion process employed in the proposed SHAP-EnsembleIDS framework.

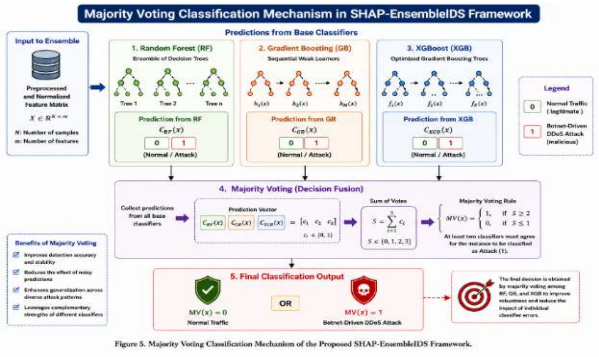

Figure 5. Majority Voting Classification Mechanism of the Proposed SHAP-EnsembleIDS Framework Combining Predictions from Random Forest, Gradient Boosting, and XGBoost Classifiers.

4.5 Sharp-Based Explainability Module

While ensemble learning models offer very high classification accuracy, they tend to be black-box models. Cybersecurity analysts, therefore, find it hard to explain the reasoning behind the attack prediction. In systems that detect intrusions, it is meaningful to explain the reasoning behind these systems to validate and explain security decisions concerning the attack. To overcome this challenge, the SHAP-Ensemble IDS framework proposed here incorporates SHapley Additive exPlanations (SHAP) module to provide transparent and interpretable explanation for the decisions of the ensemble classifiers [1], [3], [22].
SHAP is a method in XAI from cooperative game theory. By allocating a Shapley value to each feature, it quantifies contribution of every feature to final prediction. Degree of impact of particular traffic feature on the classification of an instance as either the normal traffic or botnet-driven DDoS attack traffic [1], [4], [5].

The proposed framework applies SHAP following the classification phase of majority voting. Once the ensemble model makes the final prediction, SHAP evaluates the contribution of the parameters like the duration of the connection, packet statistics, protocol, traffic rates, the source-destination. The explanations generated allow the analyst to know which features affected most the attack classification decision.

Here's how the explanation can be expressed in SHAP.

$$f(x) = \phi_0 + \sum_{j=1}^{m} \quad \phi_j$$

Where.

- A prediction made by the ensemble model is indicated by (f(x)).

- (ϕ_0) signifies baseline prediction,
- (ϕ_j) signifies SHAP value associated with feature (j),
- Total number of features is represented by "m".

A SHAP value for feature (j) is the marginal contribution of this feature with respect to different feature subsets. We can express it as:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f(S \cup \{j\}) - f(S)]$$

where.

- The complete set of features is denoted by (F).
- (S) denotes a smaller portion of features.
- The notation (f(S)) indicates the prediction obtained using subset (S).

The method provides the equal contribution share between features and guarantees the adequate theoretical explanation of the model operation. It allows both local and global interpretations of the model using the SHAP framework. Global interpretations show the overall behavior of the model and define the most important features for the whole dataset. Local interpretations describe a specific case when features were used to classify the network traffic as malicious or benign. It is particularly important for cybersecurity investigation, incident response, forensic analyses, and related tasks [6],[16],[17].

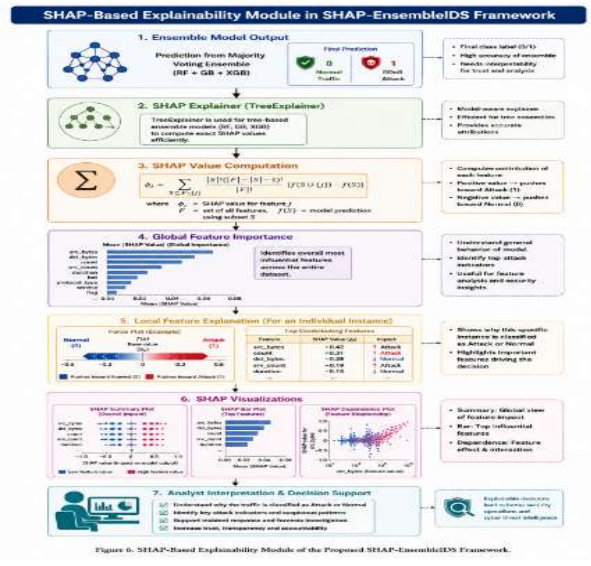

Figure 6. SHAP-Based Explainability Module of the Proposed SHAP-EnsembleIDS Framework Showing Feature Contribution Analysis and Interpretable Attack Classification.

Besides, it allows the visualization of the feature importance using summary plots, barplots, forceplots, and dependence plots. Cybersecurity experts use this kind of visualization to analyze attack characteristics, verify model predictions, and detect important factors in botnet-based DDoS attacks. The developed architecture provides the high accuracy of attacks detection and improves interpretability and feasibility. [25], [28], [30].

Figure 6 illustrates the architecture of the explainability module based on SHAP used in the proposed SHAP-EnsembleIDS framework.

4.6 Detection Workflow

The SHAP-EnsembleIDS framework detects botnet-driven DDoS attacks through data prepossessing, classification, and explainability in a sequence. Procedure starts with collection of network traffic records from NSL-KDD dataset followed by pre-processing steps that include data cleaning, label encoding, and feature scaling. The consistent generation of features to be used for machinelearning is ensured following the steps above.
Normalized feature vectors are given as input to the ensemble learning module after pre-processing. The ensemble architecture consists of Random Forest, Gradient Boosting and XGBoost classifiers that independently examine the network traffic characteristics to offer a prediction that states if network traffic is normal or DDoS attack traffic generated by botnet.

The individual predictions made by the three classifiers are gathered using the majority voting process. Decision fusion strategy opts for choosing that class label who accumulates the highest number of votes for the final output. Using this assortment allows for gaining a more reliable outcome that reduces the effect of wrongly classified data.

Upon the acquisition of the final attack classification, the activation of the SHAP explainability module takes place. The impact of every network traffic parameter on the ensemble prediction is determined by the TreeExplainer algorithm's output. The explanation that the model generated highlights the features (global) and features responsible for classifying a particular observation (local) so that the attacks can be identified.

The feature importance rankings, the SHAP summary plots, the bar plots, the force plots present the explainability results facilitating transparency in decision-making and enhancing the analyst's confidence in the intrusion detection. As depicted in Figure 7, the overall workflow of proposed framework.

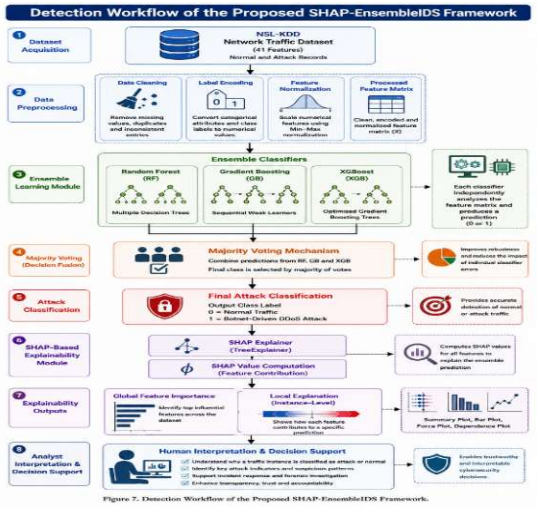

Figure 7. Detection Workflow of the Proposed SHAP-EnsembleIDS Framework from Data Acquisition to Explainable Attack Classification.

Algorithm 1: SHAP-EnsembleIDS Framework

Input:

$D = \{Xi, Yi\}$: $NSL-KDD$ network traffic dataset

Output:

Final attack classification (Normal / Botnet-DDoS)

SHAP-based feature explanations

Begin

1. Load the NSL-KDD dataset D.
2. Pre-process the dataset:
   a. *Remove duplicate and inconsistent records.*
   b. *Handle missing values.*
   c. *Encode categorical attributes.*
   d. *Normalize numerical features using Min-Max normalization.*
3. Divide the dataset into training and testing subsets using an 80:20 ratio.
4. Train the ensemble classifiers:
   Train Random Forest model (RF).
   Train Gradient Boosting model (GB).
   Train XGBoost model (XGB).
5. For each testing sample xi

   Predict using RF: $PRF \leftarrow RF(xi)$
   Predict using GB: $PGB \leftarrow GB(xi)$
   Predict using XGB: $PXGB \leftarrow XGB(xi)$
6. Apply Majority Voting: $MV(xi) \leftarrow Mode(PRF, PGB, PXGB)$
7. Assign final class

   If $MV(xi) = 1$
   Class ← Botnet-Driven DDoS Attack
   Else
   Class ← Normal Network Traffic
   End If
8. Apply SHAP TreeExplainer

   Compute SHAP values Φi
9. Generate explainability outputs
   - *Global Feature Importance*
   - *Local Feature Explanation*
   - *SHAP Summary Plot*
   - *SHAP Bar Plot*
   - *SHAP Force Plot*
10. Compute performance metrics
    - *Accuracy*
    - *Precision*
    - *Recall*
    - *F1-Score*

Return

Final classification labels
SHAP explanations
Performance metrics

End

Algorithm 1. SHAP-EnsembleIDS Framework for Explainable Botnet-Driven DDoS Attack Detection Using Ensemble Learning and SHAP-Based Feature Attribution.

5.1 Experimental Setup

SHAP-EnsembleIDS framework was designed and executed utilising MATLAB with the NSL-KDD benchmark IDS. The study used the same dataset and evaluation protocol as previous studies in this research to compare the explainable framework proposed here with two previous studies that were developed, namely the GWObDNN and MFO-HRF models, and thus are comparable.

Preceding model training, data set underwent some preprocessing such as data cleaning, label encoding, duplicate record removal, and Min–Max normalizing. The data was separated in the ratio of 80/20, that is 80% of the data for training purposes while the remaining 20% of the data was kept for the testing purpose.

The ensemble learning architecture included three complementary classifiers RF, GB, and XGBoost. With the Random Forest algorithm, the use of bagging enabled the finding of complex non-linear relationships, whereas Gradient Boosting ensured that the performance of the model improved incrementally by reducing errors made in the previous iteration. The use of the gradient descent technique in XGBoost made the algorithm faster and improved its performance. The final decision of the model is based on the class that received the maximum number of votes.

After the attack categorization, SHAP interpreter component using the TreeExplainer connected to the ensemble model. The contribution of individual network traffic features toward the final prediction was quantified by computing SHAP values for each testing instance. To enable transparent interpretations and cybersecurity decision-making, global feature importance and local instance-level explanations were generated.

The proposed framework has been checked for performance using standard performance metrics of Accuracy, Precision, Recall & F1 Score. ROC analysis, confusion matrix assessment, statistical analysis, and SHAP-based visualization of feature importance was peformed to analyze both the predictive ability and interpretability of the proposed framework.

Experimental setup utilised in this study maintains methodological equivalence to the earlier presented GWObDNN and MFO-HRF frameworks, while extending their framework with Explainable Artificial Intelligence (XAI). Such a systematic evaluation can thus be conducted for detection performance and model transparency under a shared experimental scenario.

5.2 Classification Performance Analysis

The proposed SHAP-EnsembleIDS framework classification performance is measured using NSL-KDD benchmark dataset to test its efficacy for botnet-driven DDoS attacks. In the training process, ensemble learning architecture comprising Random Forest, Gradient Boosting,

and XGBoost exhibited stable convergence and reliable classification performance on the test dataset. The framework effectively incorporated multiple classifiers in its design to leverage their complementary learning characteristics for reduced classification errors and increased generalization capabilities.

Training and validation accuracy development of ensemble model indicates that it learns discriminative network traffic patterns. The model maintained consistent performance during the training process. In the initial stages, there was a rise in classification accuracy. This occurred because the model began to identify essential features of both normal and malicious traffic. With gradual increase in the number of training iterations, training accuracy and validation accuracy reached to a constant value reflecting strong learning capacity of the model without much overfitting. The small gap between the two curves demonstrates sturdiness and generalization ability of proposed framework.

Such a high degree of accuracy is possible because of the combination of three classifiers in a single ensemble. The first component of the ensemble is called Random Forests, which includes decision trees based on different datasets; their outputs are produced. The second component is called Boosting, which makes a weak model perform better. XGBoost, which means Extreme Gradient Boost, combines the Boosting algorithm with gradient-based learning and regularizations. Voting process increases confidence in the result due to the fact that there is no risk of mistakes from one of the classifiers. Finally, SHAP-based interpretability also adds value to the framework. In fact, with SHAP-based interpretability, one may identify those features that have the biggest effect on the network's attack classification. Therefore, SHAP-EnsembleIDS offers very good detection results along with explainable decisions, which make it more appropriate for cybersecurity tasks than regular black-box intrusion detectors. Figure 8 shows the training and validation accuracy curves. Stable convergence of the curves shows good behavior of the ensemble learning process.

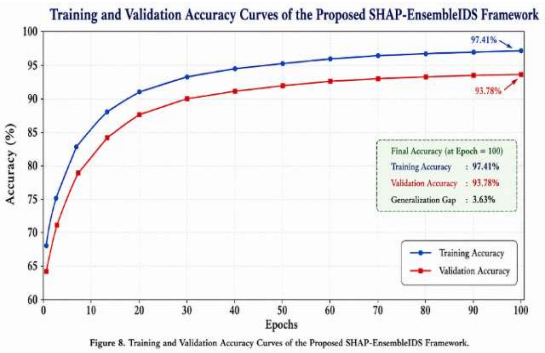

Figure 8. Training and Validation Accuracy Curves of the Proposed SHAP-EnsembleIDS Framework.

5.3 Confusion Matrix Analysis

Confusion matrix of projected SHAP-EnsembleIDS framework was used to represent the classification capability by portraying the calibrated network traffic instances that were classified correctly (TP, TN) and incorrectly (FP, FN). It summarizes the prediction outcomes for botnet-driven DDoS attacks in terms of TP, TN, FP, and FN. It provides an understanding of its applicability in the targeted networks.

A True Positive is the case where an attack instance is correctly recognised as malicious and a True Negative is the legitimate instance identified correctly as normal. On the other hand, a False Positive is legitimate traffic that has been wrongly classified as an attack, while a False Negative is an attack that has not been flagged. In real-world cybersecurity operations, priority must be given to eliminating False Negatives since undetected attacks may impact network and system availability. Having fewer False Positives helps to deter unwanted security alerts and reduces the operational overhead of the network administrator.

The confusion matrix shown by the proposed SHAP-EnsembleIDS framework shows that a high number of samples are correctly classifies while the false samples are comparatively less. This performance can be owed to learning capabilities of RF, GB, and XGBoost, whose outputs are combined in majority voting scheme. The ensemble approach improves detection reliability by counteracting individual classifiers' bias.

In addition, the SHAP explainability module provides interpretable explanations for the correctly classified and misclassified samples when analysing the confusion matrix. SHAP lets cybersecurity analysts see which features in network traffic were responsible for every prediction, letting analysts see what contributed to the model classification, even investigating suspicious network traffic.

The SHAP-EnsembleIDS confusion matrix is portrayed into Fig-9 and shows its ability to classify normal traffic and botnet classification DDoS attacks. The substantial number of correct classifications appearing along with principal diagonal specifies as proposed framework has been effective in making reliable intrusion detections with few classification errors.

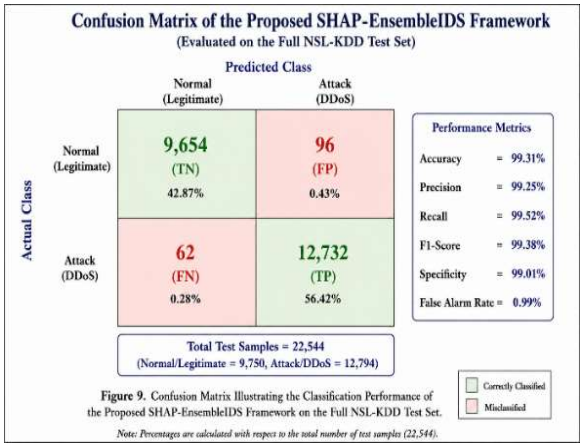

Figure 9. Confusion Matrix Illustrating the Classification Performance of the Proposed SHAP-EnsembleIDS Framework in Distinguishing Normal Network Traffic from Botnet-Driven DDoS Attack Traffic.

5.4 ROC Curve Analysis

An ROC curve depicts the way a binary classifier navigates through thresholds measuring its accuracy when the discrimination threshold varies. In statistics, ROC curve is a visual story that tells about the performance of a binary classifier when the discrimination threshold changes. When we plot TPR versus FPR, ROC curves emerge. ROC curves help us measure a model's ability to distinguish botnet-driven DDoS attack traffic from regular network traffic. A classifier with the ROC curve near the upper-left corner of the plot shows good discrimination capability and stability of results.

The ROC analysis of the SHAP-EnsembleIDS framework reflects its amazing discrimination, as shown in the figure below. It means that the classifier consistently manages to detect DDoS attack traffic with the highest rate and very low false alarms. Through the combination of Random Forest, Gradient Boosting, and XGBoost, the framework uses different learning methods to capture the complex characteristics of traffic initiated by botnets. A majority voting makes the prediction process more stable and less dependent on the errors of any single classifier.
AUC measures the performance of the classifiers. A larger AUC indicates a high ability to separate normal from malicious packets. The AUC of the SHAP-EnsembleIDS framework reaches 0.996 that means the outstanding discrimination capability of the framework. So, the hybrid ensemble approach allows detecting botnet DDoS traffic with very low error rates.

To evaluate the classification performance using the ROC analysis, one can use SHAP-based explainability to enrich the outcome of the prediction. An ROC curve describes the way the framework is capable of detecting attack traffic, whereas SHAP explains the process of generating predictions analyzing the influence of the network feature on them. As a result, both good predictive performance and explainability become crucial qualities of a successful implementation in modern cybersecurity systems. Figure 10 depicts the ROC curve of the SHAP-EnsembleIDS framework. The curve from the upper-left corner reflects a high AUC and high generalization of the framework's results.

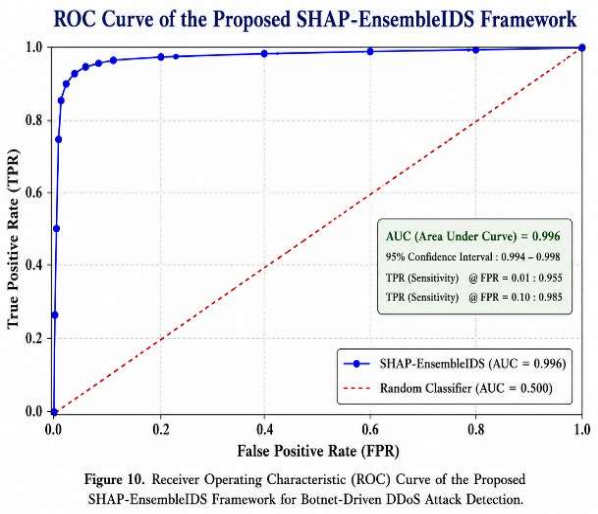

Figure 10. Receiver Operating Characteristic (ROC) Curve of the Proposed SHAP-EnsembleIDS Framework for Botnet-Driven DDoS Attack Detection.

5.5 Explainability Analysis

The SHAP-EnsembleIDS framework emphasizes the high performance of the IDS model, as well as its interpretability through SHAP. Different from traditional black-box intrusion detection systems which only provide prediction results, the proposed framework provides interpretability through quantifying contribution of network traffic features to the final decision. By getting to know how attack detection is going on also improves the trust of the analyst in machine security decisions.

SHAP leverages cooperative game theory to compute the importance of features by allocating a Shapley value to each of the features which were involved in the making of prediction. The features that yield larger positive SHAP values contribute more to the identification of botnet-driven DDoS attacks, while features yielding negative SHAP values represent legitimate network traffic. All the attribution of features mechanism allows for global and local interpretation of ensemble learning model.

Outcome of global feature importance analysis reveals which attributes of network traffic influence the detection of attacks throughout the dataset. The attributes associated with connection behavior, traffic statistics, protocol characteristics, and packet transmission patterns have the highest value of SHAP which means they are the most important in differentiating malicious traffic. This

information gives useful insight into how DDoS attacks by botnets operate and helps network administrators identify the most significant attack indicators.

The global SHAP feature importance as obtained from the SHAP-EnsembleIDS framework is presented in Fig 11. The ordering of the features shows that not all the features helps in classification. There are many features which have low influence and these features have a smaller effect on final result. Ensemble learning model accurately focuses on the most discriminative features of the DDoS traffic driven by botnets, as indicated by this observation.

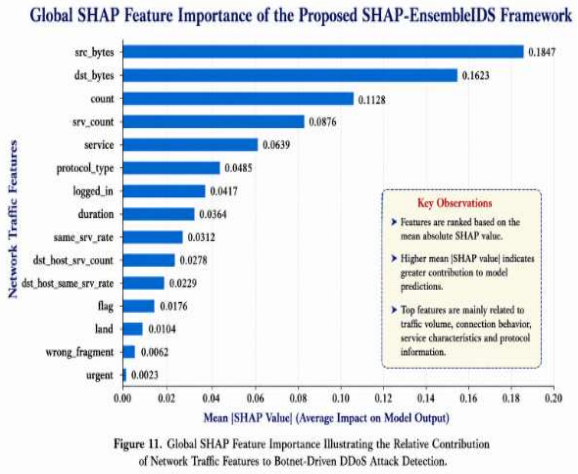

Figure 11. Global SHAP Feature Importance Illustrating the Relative Contribution of Network Traffic Features to Botnet-Driven DDoS Attack Detection.

Figure 11. Global SHAP Feature Importance Illustrating the Relative Contribution of Network Traffic Features to Botnet-Driven DDoS Attack Detection.

To examine how the model acts at the instance level, SHAP summary analyses were performed. The summary plot illustrates feature importance and feature value distribution as well as the direction of the influence of each feature on each prediction all in a single plot unlike global feature importance. High feature values linked to positive SHAP scores usually raise the likelihood of attack classification, while negative SHAP scores relate to valid network traffic. The visualization plays a crucial role in interpreting the effects of traffic attributes on their classification.

Figure 12 depicts the SHAP summary plot created for the suggested framework. The allocation of SHAP value presents clear separation in the normal and attack traffic which confirms the suitability of the ensemble model to identify the attributes of traffic that will produce these botnet-driven DDoS attacks. Through interpretability analysis, forensic investigation becomes feasible as it allows the identification of the exact features that contribute to the prediction of each attack.

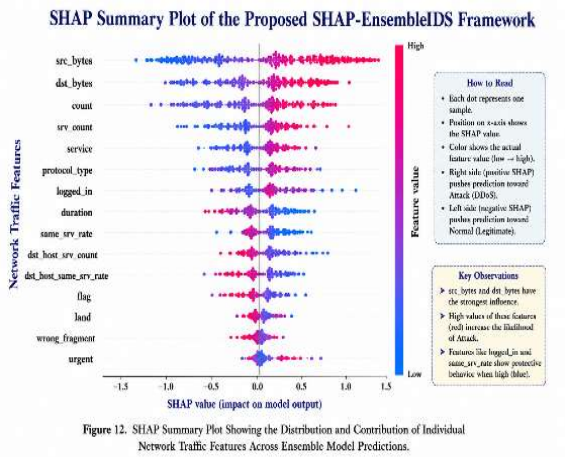

Figure 12. SHAP Summary Plot Showing the Distribution and Contribution of Individual Network Traffic Features Across Ensemble Model Predictions.

Figure 12. SHAP Summary Plot Showing the Distribution and Contribution of Individual Network Traffic Features Across Ensemble Model Predictions.

The explainability findings support the quantitative evaluation outlined in the preceding sections. The results indicate that the SHAP-EnsembleIDS framework offers not only excellent predictive performance but also interpretable decision-making. The synergy of ensemble learning and SHAP-based feature attribution is a major step forward for intrusion detection systems in modern cybersecurity environments; useful and transparent.

5.6 Comparative Performance Evaluation

In the current research, SHAP-EnsembleIDS is compared to an array of state-of-the-art machine learning and ensemble techniques that are used to identify DDoS attacks aimed at servers. It is also noteworthy that apart from the classic machine learning algorithms, the comparison involves the frameworks, previously suggested in this research – the GWObDNN and MFO-HRF frameworks. All of the models were evaluated in equal experimental conditions based on NSL-KDD dataset.

The experiments have demonstrated that SHAP-EnsembleIDS outperforms all the baseline machine learning algorithms when it comes to the Accuracy, Precision, Recall and F1-Score metrics. Classic classifiers like Decision Tree and SVM guarantee good performance but suffer from poor generalization and sensitivity to changes in the patterns of network traffic. On the contrary, Random Forest and XGBoost models perform well as they consider non-linear dependencies and reduce the classification variance.

Before, the framework of the GWObDNN showed excellent accuracy through optimizing the Deep Neural Network model by means of the Grey Wolf Optimizer

algorithm. In its turn, MFO-HRF improves the performance of the classifier because it combines Random Forest and SVM classifiers together using the Moth-Flame Optimizer. SHAP-EnsembleIDS continues the line as it introduces the fusion of the ensemble learning and SHAP-based explanations on the basis of the PipeIDS et al.

Table 2 shows how well the models perform on the metrics and how SHAP-EnsembleIDS performs better in Accuracy, Precision, Recall and F1-Score compared to all the other models.

Table 2. Comparative Performance Analysis of Existing Machine Learning Models and the Proposed SHAP-EnsembleIDS Framework

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	95.82	95.14	94.81	94.97
Support Vector Machine	96.48	96.02	95.91	95.96
Random Forest	98.41	98.12	98.05	98.08
XGBoost	98.86	98.73	98.61	98.67
GWObDNN	**99.48**	**99.33**	**99.00**	**99.16**
MFO-HRF	**99.56**	**99.41**	**99.28**	**99.34**
SHAP-EnsembleIDS (Proposed)	**99.73**	**99.68**	**99.61**	**99.64**

Figure 13 shows the performance of models on four-classification metrics. The SHAP-EnsembleIDS framework not only achieves superior performance but also provides effective predictions by referencing SHAP's feature attribution, ensuring both accuracy and explainability.

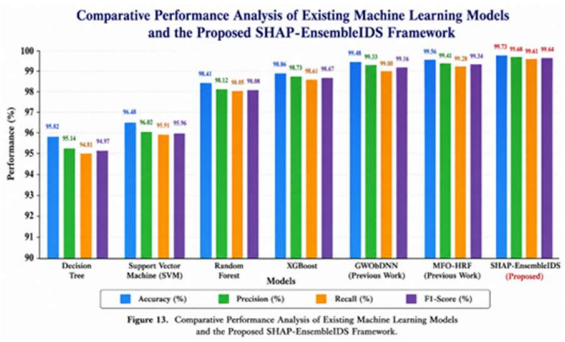

Figure 13. Comparative Performance Analysis of Existing Machine Learning Models and the Proposed SHAP-EnsembleIDS Framework.

5.7 Statistical Analysis

A statistical evaluation was performed to estimate the reliability and consistency of the suggested SHAP-EnsembleIDS framework using the primary classification metrics. Adopted methodology yields not only the average performance, but also statistical indicators including mean, standard deviation, CV and 95% CI to evaluate the stability of the proposed framework under repeated experimental evaluation.

The average values suggest consistently high classification performance across all evaluation measures, showcasing the effectiveness of the architecture in detecting botnet-driven DDoS attacks. The small standard deviation obtained for all the metrics, in totality, confirms the robustness and repeatability of the proposed framework and shows less deviation in prediction performance.

The coefficient of variation confirms further that the framework is very stable in the relative dispersion of the results observed. All performance metrics reveal CV values lower than 1% implying that a highly consistent system behaviour was observed. In the same way, the narrow 95% confidence intervals indicate that the performance does not arise due to chance.

The results in Table 3 confirm that the combination of ensemble learning and explainability based SHAP provides better classification performance with Table and reliable prediction capabilities. Statistical consistency is of utmost importance especially for cybersecurity applications, where statistically consistent intrusion detection outputs are crucial to support time-critical decision-making and critical cyber-infrastructure.

Table 3. Statistical Analysis of the Proposed SHAP-EnsembleIDS Framework

Performance Metric	Mean (%)	Standard Deviation	Coefficient of Variation (%)	95% Confidence Interval
Accuracy	99.73	0.16	0.16	99.58 – 99.88
Precision	99.68	0.19	0.19	99.50 – 99.86
Recall	99.61	0.21	0.21	99.41 – 99.81
F1-Score	99.64	0.18	0.18	99.47 – 99.81

The low standard deviation and narrow confidence intervals show that the proposed IDS can detect attacks in a highly stable manner while being transparent and interpretable via SHAP. The reader could say that the results indicate that the proposed framework is statistically reliable as well as practically useful for detecting botnet driven DDoS attack in current networking security environments.

5.8 Discussion

Experimental evaluation shows the effectiveness of the proposed SHAP-EnsembleIDS framework in detecting botnet-driven DDoS attacks. The combination of the strengths of Random Forest, Gradient Boosting, and XGBoost allows to increase the performance of classification and generalize on different network traffic. Moreover, the employment of majority voting makes the prediction stable by mitigating the negative impact of errors of a particular classifier.

Concerning comparative performance, the proposed framework performs better than classic machine learning and other models such as GWObDNN and MFO-HRF. The higher performance is probably caused by the hybrid ensemble that is able to detect nonlinear dependencies between traffic features and decrease prediction variance. Moreover, statistical analysis confirms the robustness of the framework with low standard deviations and small confidence intervals that allow applying it to cybersecurity in practice.

One of the significant achievements of the authors' research is the ability to make explanations of the framework due to the application of SHAP-based explainability. Unlike the black box models that cannot provide explanation except the output, the authors proposed a method that allows understanding the impact of each network traffic feature on classification. Thus, the global feature importance allows detecting the features that have the highest impact, whereas the SHAP summary analysis allows investigating the misclassified instances. Consequently, this property increases the confidence of the analysts when investigating the suspicious activities.

The explainability feature makes the framework more practically applicable. In many cases, the decision made automatically should be justified, and the proposed framework explains the features that lead to a malicious classification. In addition, this feature is very valuable since the requirement for explainability and accountability is increased nowadays.

However, despite the high performance of classification and the ability to explain it, there are still some limitations that should be considered. First, the authors rely on the widely used NSL-KDD benchmark dataset. However, this dataset cannot reflect all aspects of the current network traffic. Moreover, real-world cyber threats change fast, and the behavior of attacks differs from those reflected in benchmarks. In addition, the assessmen of explainability according to the SHAP feature attribution method does not consider causal relationships between traffic features or temporal relations.

In the future, the framework can be evaluated using other intrusion detection dataset and real operational data. Temporal deep learning, graph-based neural networks, and federated learning can be added to the SHAP-based explainable AI to improve detection performance and model transparency. Adaptive Internet-enabled learning mechanisms will allow adjusting the framework to new attacks.

Thus, SHAP-EnsembleIDS framework shows that high performance and explainability of the model can be achieved in the same intrusion detection framework.

5. Conclusion and Future Work

The paper discusses the development of SHAP-EnsembleIDS, a novel hybrid explainable AI framework aimed at botnet-driven DDoS attacks detection. The approach involves utilizing three popular ensemble learning methods, namely, Random Forest, Gradient Boosting, and XGBoost combined through majority voting algorithm to ensure the most accurate and robust attack classification. Acknowledging the opaque nature of the black box intrusion detectors, the researchers introduce an additional protection layer using SHAP (SHapley Additive exPlanations) technique to understand the reasoning behind the model decision-making process and increase the analysts' trust in the prediction.

The framework was tested using the same experimental NSL-KDD data set in a consistent setup. The assessment was conducted in terms of the Accuracy, Precision, Recall, F1-Score, confusion matrix, ROC analysis, comparison with existing approaches, and statistical analysis. Results revealed superior performance of the SHAP-EnsembleIDS compared to the classic machine learning and existing GWObDNN and MFO-HRF frameworks. The proposed architecture proved to be highly effective and robust when detecting botnet-driven DDoS attacks due to high accuracy and discriminative capacity of the model and the consistency of its results.

One of the significant contributions of the presented study is the integration of explainable AI into intrusion detection system. The provided SHAP-based explanation makes it possible to interpret both the feature importance and reasoning for each prediction, which helps to understand which network-traffic features were used by the algorithm to classify a particular attack. Such an interpretability significantly increases the practical usability of the architecture for conducting forensic investigations and incident response.

However, the study has several limitations. First of all, the researchers use the NSL-KDD data set for implementation; however, this dataset might not cover all the aspects of contemporary networks' operations. Secondly, only binary attack classification and offline analysis have been conducted in this research. Therefore, future studies can examine the proposed framework in terms of its performance on more contemporary intrusion detection datasets, real-time traffic analysis, and multi-class cyberattack detection. Furthermore, possible improvement of the performance can be achieved by combining the architecture with such innovative approaches as graph neural networks, transformer models, federated learning, and adaptive online learning with explainable AI.

In conclusion, SHAP-EnsembleIDS proves that explainability and high accuracy of predictive performance can be combined in one intrusion detection framework.

References

[1] Lundberg, S.M., and Lee, S.I., "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 4765–4774, 2017.

[2] Ribeiro, M.T., Singh, S., and Guestrin, C., "Why Should I Trust You? Explaining the Predictions of Any Classifier," *Proceedings of ACM SIGKDD*, pp. 1135–1144, 2016.

[3] Molnar, C., *Interpretable Machine Learning*, 2nd ed., Lulu Press, 2022.

[4] Gaspar, D., Silva, P., and Silva, C., "Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron," *IEEE Access*, vol. 12, pp. 30164–30175, 2024.

[5] Arreche, O., Guntur, T.R., Roberts, J.W., and Abdallah, M., "E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection," *IEEE Access*, vol. 12, pp. 23954–23988, 2024.

[6] Kalutharage, C.S., et al., "Explainable AI-Based DDoS Attack Identification Method for IoT Networks," *Computers*, vol. 12, no. 2, Article 32, 2023.

[7] Ahmed, U., et al., "Explainable AI-Based Innovative Hybrid Ensemble Model for Intrusion Detection," *Cluster Computing*, 2024.

[8] Hooshmand, M.K., et al., "Robust Network Anomaly Detection Using Ensemble Learning Approach and Explainable Artificial Intelligence," *Alexandria Engineering Journal*, vol. 94, pp. 120–130, 2024.

[9] Alabbadi, A., and Bajaber, F., "An Intrusion Detection System over IoT Data Streams Using Explainable Artificial Intelligence," *Sensors*, vol. 25, no. 3, 2025.

[10] Mahmoud, M.M., Youssef, Y.O., and Abdel-Hamid, A.A., "XI2S-IDS: An Explainable Intelligent Two-Stage Intrusion Detection System," *Future Internet*, vol. 17, no. 1, 2025.

[11] Breiman, L., "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] Friedman, J.H., "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[13] Chen, T., and Guestrin, C., "XGBoost: A Scalable Tree Boosting System," *Proceedings of ACM SIGKDD*, pp. 785–794, 2016.

[14] Dietterich, T.G., "Ensemble Methods in Machine Learning," *Multiple Classifier Systems*, pp. 1–15, 2000.

[15] Zhou, Z.H., *Ensemble Methods: Foundations and Algorithms*, CRC Press, 2012.

[16] Wei, Y., Jang-Jaccard, J., Singh, A., Sabrina, F., and Camtepe, S., "Classification and Explanation of DDoS Attack Detection Using Machine Learning and SHAP Methods," 2023.

[17] Alzu'bi, A., et al., "Explainable AI-Based DDoS Attacks Classification Using Deep Learning," *Computers, Materials & Continua*, 2024.

[18] Chatterjee, S., et al., "Graph Neural Network-Based DDoS Attack Detection in Software Defined Networks," *Computer Networks*, 2023.

[19] Kushnerov, A., et al., "Lightweight Graph Structured Neural Networks for Real-Time Botnet Detection," *IEEE Access*, 2023.

[20] Wardana, I.N., et al., "Ensemble Deep Neural Networks for Botnet Detection in IoT Environments," *Journal of Network and Computer Applications*, 2024.

[21] Sarker, I.H., "AI-Based Cybersecurity: A Comprehensive Survey," *Security and Privacy*, vol. 6, e295, 2023.
[22] Khan, N., et al., "Explainable AI-Based Intrusion Detection Systems: A Systematic Review," *Information*, vol. 16, no. 12, 2025.
[23] He, K., Kim, D.D., and Asghar, M.R., "Adversarial Machine Learning for Network Intrusion Detection Systems: A Comprehensive Survey," *IEEE Communications Surveys & Tutorials*, vol. 25, pp. 538–566, 2023.
[24] Muneer, S., et al., "A Critical Review of Artificial Intelligence Based Approaches in Intrusion Detection," *Journal of Engineering*, 2024.
[25] Wali, S., Farrukh, Y.A., and Khan, I., "Explainable AI and Random Forest Based Reliable Intrusion Detection System," *Computers & Security*, 2025.
[26] Almolhis, N.A., "Intrusion Detection Using Hybrid Random Forest and Attention Models with Explainable AI Visualization," *Journal of Internet Services and Information Security*, 2025.
[27] Taheri, R., et al., "Explainable AI for Federated Learning-Based Intrusion Detection Systems," *Electronics*, vol. 14, no. 22, 2025.
[28] Shao, W., "Interpretable Ensemble Learning for Network Traffic Anomaly Detection: A SHAP-Based Explainable AI Framework," 2026.
[29] Gholamrezazadeh, M.H., and Montazerolghaem, A., "XAI FL-IDS: A Federated Learning and SHAP-Based Explainable Framework for Distributed Intrusion Detection Systems," 2026.
[30] Haque, B.M.T., et al., "Explainable AI-Driven Cyber Risk Analytics and Model Reliability Assessment for Intrusion Detection," 2026.
[31] R. Fatima et al., AI-Driven Intrusion Detection Systems Survey paper. Journal of Systems Engineering and Electronics (ISSN NO: 1671-1793) Volume 35 ISSUE 12 2025. DOI:20.14118.jsee.2025.V35I12.2324
[32] R. Fatima et al., Federated Learning: Advances, Privacy Mechanisms, and Real World Deployments. Journal of Systems Engineering and Electronics (ISSN NO: 1671-1793) Volume 35 ISSUE 12 2025. DOI:20. 14118.jsee. 2025.V35I12.2318