International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

EXPLAINABLE ARTIFICIAL INTELLIGENCE MODELS FOR INTRUSION DETECTION SYSTEMS TO IMPROVE THE EFFICIENCY AND INTERPRETABILITY OF BLACK BOX MODELS

[1]Alycia Sebastian, [2]S.Silvia Priscila,[3]Praveen B.M,

[1]Alycia Sebastian is doing her PostDoc fellowship in Institute of Engineering and Technology, Srinivas University, working in the Information Technology Department as an Assistant Professor, Al Zahra College for Women, Sultanate of Oman.(email: alycia@zcw.edu.om) [0000-0003-0022-2431]

[2]S.Silvia Priscila working as an Associate Professor in the Department of Computer Science, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India(email: silviaprisila.cbcs.cs@bharathuniv.ac.in) [0000-0002-6040-3149]

[3]Praveen B.M is working as a Professor in the Department of CyberSecurity and Cyber Forensics, Institute of Engineering and Technology, Srinivas University (email: bm.praveen@yahoo.co.in) [0000-0003-2895-5952]

Abstract

In the field of cybersecurity, Intrusion Detection Systems (IDS) are commonly used to prevent and minimize threats. Systems for detecting intrusions aid in preventing threats and vulnerable points from entering computer networks. Many Machine Learning (ML) techniques are available for assisting development of IDS that perform effectively. In a broad range of tasks, ML based systems have demonstrated improved learning performance. But, the problem with the certain cutting-edge models is that they lack in explanation ability, transparency, and reliability. Explainable AI (XAI) approaches are employed in order to comprehend and clarify these AI models to security analysts. The present study proposes an architecture that uses SHAP for Intrusion detection Systems (IDS) to increase the interpretability of the model by extracting the explanations from the black box models like XGBoost, AdaBoost, Support Vector Machines (SVM) and Random Forest (RF) algorithms. The explanations provided by the SHAP analysis are validated with filter based Feature selection method. The experiment's findings show that the XGBoost model outperforms the other ML models and the SHAP analysis is performed for the ML model to study the efficacy of the explanations and the importance of features.

Index Terms— Intrusion Detection Systems, Machine Learning, Explainable Artificial Intelligence (XAI), Random Forest(RF), Support Vector Machine(SVM), AdaBoost, XGboost, SHapley Additive explanations(SHAP)

I. INTRODUCTION

The notion of "network security" refers to a broad range of strategies and technological advancements designed to safeguard the reliability, integrity, safety, and usability of a network and its information. It targets a wide range of threats and uses both hardware and software technologies in an attempt to prevent them from penetrating or propagating within

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

the network. Intrusion detection monitors a network or computer system for fraudulent activity, such as illegal

access, abuse, or alteration of system resources. ID tries to recognize such action in real-time or very near to real-time

and reacts appropriately to prevent further information loss or damage. (Munner et al, 2024). IDS are intended for examining network and system activity in order to identify any unusual patterns that could indicate the existence of an attack(Malik et al, 2022).

Many ML and Deep Learning(DL) techniques were used to classify data as normal or invasive (Abrar et al, 2020., Vinaykumar et al., 2019, Al-Omari et al., 2021). Specifically, these earlier AI-based investigations concentrated more on the classification precision of different AI algorithms without providing an understanding of their logic or functioning. This drawback emphasizes the critical need to take into consideration the relatively new field of XAI in order to improve the understandability of AI conclusions in IDS.

A. Intrusion Detection Systems

There are two major methods of detection that IDS are usually based upon. Signature-based IDS (SIDS) identifies malicious traffic based on whether the network traffic matches a database of known attack signature, whereas Anomaly-based IDS (AIDS) identifies abnormal or unusual system behavior. But these techniques are highly constrained in contemporary settings.

To begin with, they rely on the previous knowledge of attack signatures thus they are not applicable to any novel or zero-day attacks. Second, the storage and computation of extensive databases of attack patterns is a resource-consuming task, which is especially demanding in the case of IoT devices with small storage and processing capabilities.

The signature-based IDS tend to be effective when required to uncover known threats by comparing network traffic to an existing set of malicious indicators or Indicators of Compromise (IOCs). Such IOCs can be file hashes, suspicious URLs, malicious code snatches, specific email header, or behavioral patterns that frequently prelude an attack. The system scans network packets, matches them to its signature database and marks any traffic that has an appearance of a known attack or IOC.

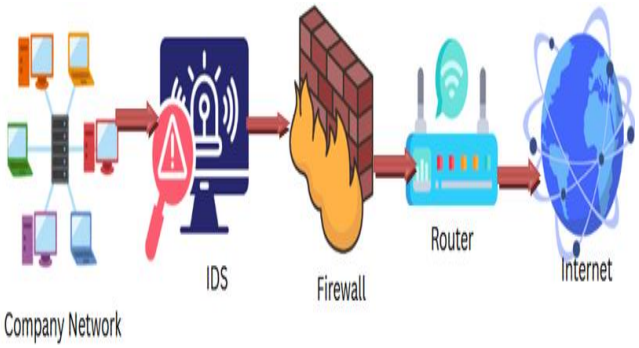

Figure 1. Intrusion Detection Systems

AIDS monitors aberrant activity, making it capable of identifying new or undiscovered threats that vary from regular ones (Sharma et al, 2024). Any variation from the existing

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

regulated baseline, such as a user attempting to log in outside of typical business hours, illegal device updates to a network, or an inflow of new IP addresses seeking to join to a network, may trigger an AIDS flag. The drawback in this situation is that a lot of benign actions will be reported only for being unusual. Because anomaly-based IDS has a higher chance of producing false positives, it may take more time and money to look into every signal pertaining to a potential danger(Satilmus et al., 2024).

B. Explainable Artificial Intelligence (XAI)

The phrase "explainable AI" refers to an AI model's expected effects and potential biases. It helps to define model accuracy, accountability, equity, and decision outcomes powered by AI. When using AI models in production, the company must be prepared to clarify AI in order to acquire the trust of its stakeholders. Explainability in AI helps an organization implement a strategy for sustainable growth for AI(Arrecho et al., 2024. Mohammed et al., 2024). The Black box model vs white box model is depicted in figure 2. XAI attempts to provide details about the reasoning and decision-making process of the model, in contrast to black-box AI models, which frequently serve as opaque decision-making systems.

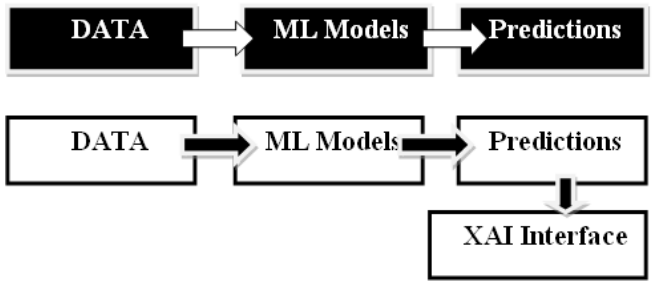

Figure 2 Black box Model Vs White Box Model

XAI is a collection of methods and strategies that allow end users to understand and rely on the output generated by ML algorithms. In light of the current environment of growing ethical issues around AI, transparency is especially crucial. AI systems especially are becoming more and more common in daily life, and the decisions they make can have a big impact. In theory, these technologies could assist in removing human bias from historically biased decision-making processes, such as setting bail or evaluating a borrower's eligibility for a home loan. Because of the biased nature of the data used to train them, implemented systems inadvertently maintained prejudiced attitudes despite efforts to eradicate racial prejudice from these procedures through AI. It is critical that AI systems are carefully examined and created in accordance with responsible AI (RAI) principles as the amount of time people rely on these systems to make critical decisions in the real world increases (violet turri , 2022). In XAI, these models' descriptions are essential with its features and labels. Then create global priority values for every aspect. The graphs are used to examine how each feature affects the AI model's choice; one can create expectations for the model's behavior. The major contributions of the study includes

- Examining the ML models to classify normal or attacks in the network traffic data
- Analysis of the ML model to contrast with other models
- Model Explanation using SHAP analysis for the predictions by the ML models.

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

In further sections, the study is arranged with the related works in the study field in the recent years in section 2 followed by the methodologies used in the study including ML for classification of network traffic data and XAI models for providing the interpretations. In section 4 , the performance evaluation of ML and XAI model were done and it is presented with the study findings and interpretations followed by conclusion and future work in section 5.

Ii. Related Works

A number of recent studies have started to investigate the use of XAI for IDS to generate interpretations. AI-based security controls, which incorporate ML algorithms into security controls like intrusion (Belavagi et al., 2023) and malware detection, are an increasingly common means to combat the sophistication of cyberattacks and the complexity of cybersecurity. These AI security controls are thought to be more successful than conventional heuristic- and signature-based controls. However, these AI-based security controls are becoming black-box systems as a result of the increasing use of sophisticated ML algorithms (Yayla et al.,2022)

Arrecho et al., 2024, provide an end-to-end framework for evaluating black-box XAI approaches for network detection of breaches on both a national and local level. Then, two well-known black-box XAI techniques, SHAP and LIME, were tested using six different evaluation metrics: precision, sparseness, reliability, efficacy, robustness, and completeness.

Bellegdi et al, 2014 examines the efficiency of XAI methods for improving the transparency of ML-based IDS to categorize a given network flow as either malicious or benign and lightBGM and XGboost outperforms the other methods. Then both global and local explanations for the LightGBM's predictions were given by SHAP and LIME approaches.

Hooshmand et al., 2024 suggests an ensemble model for anomaly ID. In this method, the K-means clustering and the Synthetic Minority Oversampling Technique (SMOTE) are used to address the problem of imbalanced data(SKM). SMOTE oversamples the minority class, while K-means undersamples it based on clusters. Using a Denoising Autoencoder (DAE), the top 15 features are chosen based on their higher weights in order to minimize the dimensionality of the data. To provide an explanation for anomalous IDS, the SHAP technique and the XGBoost algorithm are used.

In the last ten years, a number of IDS have been suggested to shield systems from cyberattacks. IDS systems based on ML proved exceptional efficacy against traditional cyber threats. However, as traditional ML-based techniques are susceptible to adversarial attacks, the rise of adversarial attacks in the cyber world emphasizes the necessity of updating these IDS. Wali et al.'s suggested IDS system (2023) combines XAI with the performance of traditional ML-based IDS to combat adversarial attacks. The table 1 lists a few recently studied ML models with XAI to generate explanations

Table -1 Recent studies using ML and XAI models for IDS

Author(s)	Year	ML/DL approaches used	XAI method employed	Dataset used	Evaluation metric /Score

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Arrecho et al.	2024	RF, DNN, LGBM, SVM, MLP, ADA, KNN	SHAP and LIME	RoEduNet-SIMARGL2021	Accuracy, sparsity, stability, efficiency, robustness, and completeness.
Bellegdi et al.	2024	LightGBM and RF, Adaboost, XGBoost	SHAP and LIME	CIC-IDS2018	F1-Score- LightGBM -0.979 XGBoost- 0.978
Hooshmand et al.	2024	XGBoost	SHAP	NSL-KDD and UNSW-NB15	Detection rate UNSW-NB-15 - 99.01% NSL-KDD-99.22%
Wali et al.	2023	RF Classifier	SHAP	CICIDS	Accuracy 98.5%
Hariharan et al.	2023	RF, XGBoost, Light Gradient Boosting	SHAP, LIME	Real time network traffic data	Accuracy

III. METHODOLOGY

The present study uses ML models like XGBoost, Adaboost, SVM and RF for the classification of IDS. The interpretations are generated with the SHAP and LIME following the application of the ML models like XGBoost, AdaBoost , SVM and RF and the architecture of the model is depicted in Figure 3.

A. XGBoost

XGBoost is an efficient and scalable version of gradient-boosted DTs (GBDT). It has one of the most popular ML models used when it comes to regression, classification, and ranking, and the model provides us with efficient parallel tree boosting. GBDT as a prediction algorithm is an ensemble model which trains a set of DTs to enhance forecasting accuracy, much like RF. In ensemble learning, two or more algorithms are used to generate a stronger and better model. Multiple DTs make up the model that is constructed by both GBDT and random forest. It is the construction and fusion of the trees that differs(Hooshmand et al,2024)

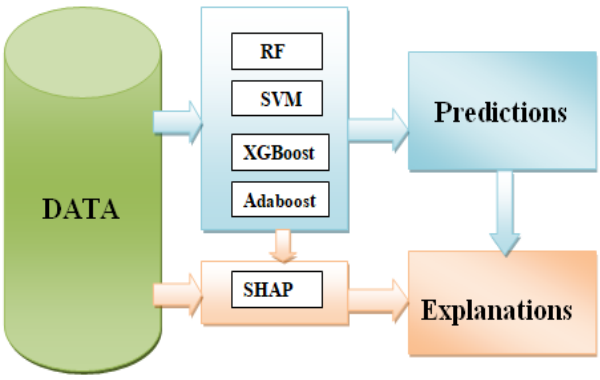

Figure 3.Overview of XAI models to interpret the ML models Predictions

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The boosting concept is used to improve the weak model by adding multiple weak models to form one strong model of prediction. Gradient boosting is a certain kind of boosting that enlists an objective and a gradient descent optimization method to construct weak models in a systematic, additive way, enhancing the performance of the overall models one step at a time. It is used to determine the desired results for the subsequent approach in order to eliminate mistakes. Figure 4 depicts how the gradient of the inaccuracy in relation to the prediction is used to identify the intended outcomes for each case.

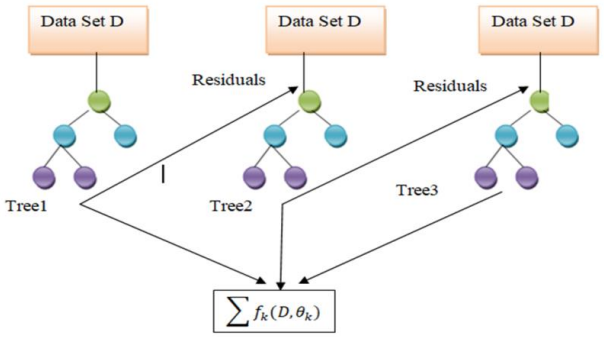

Figure 4 Illustration of XGBoost

B. Adaboost

The most beneficial applications for AdaBoost are binary classification tasks, where it improves DT performance. Any ML algorithm can be made to perform efficiently with the support of AdaBoost. These are models that when applied to classification attains accuracy slightly higher than random guessing. When used in classification problems (as opposed to regression) this method is also referred to as discrete AdaBoost. The concept of AdaBoost is based on the idea of attaching weights to error and that of combining several weak classifiers to create a powerful overall classifier. Its fundamental concept is to repeatedly train the data, varying the weights of the classifiers with each step in order to better predict challenging or misclassified examples. For instance, the proportion of weight of every single sample for the initial iteration is

$$\text{Weight}(Xi) = 1/N \qquad (1)$$

Where Xi is the ith sample and N is the number of samples.

A weak classifier is produced with the weighted training samples, and it is often a decision stump. The decision stumps assess a single input feature and produce +1.0 or -1.0 which is one of the two class labels. This implementation is particularly aimed at binary (two-class) classification problems. The overview of Adaboost algorithm is shown in figure 5 followed by the steps involved in this algorithm (Bellegdi et al. ,2024).

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

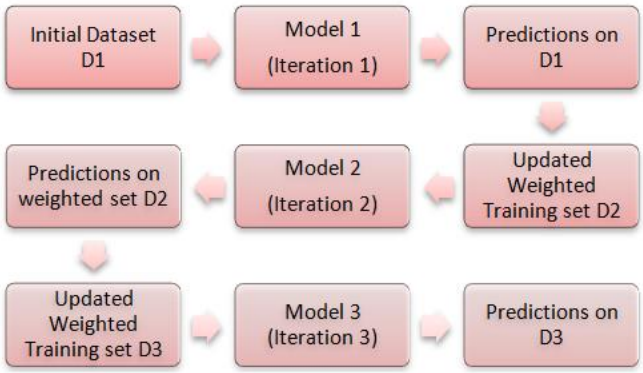

Figure 5 Overview Adaboost Algorithm

Algorithm Adaboost:

Step 1: Choose a training subset at random.

Step 2: The model is iteratively trained by choosing the training set according to the previous training's accurate prediction.

Step 3 : Outputs a wrongly classified samples a higher probability of categorization in the following iteration, it gives them a larger weight.

Step 4, the trained classifier's accuracy is used to determine its weight for each repetition. The classifier with the highest accuracy will receive more weight.

Step 5: Repeat the technique until all training data fits correctly or the maximum number of estimators is met.

C. Support Vector Machine (SVM)

The SVMs refer to a family of supervised learning techniques that are applied in the activities of regression, classification, and detection of anomalies. In contrast to methods that only minimize training errors, SVMs maximize the decision function by minimizing structural risk, to improve generalization and decrease the risk of overfitting. A portion of the input data is used to identify maximum margin hyperplanes between classes. A support vector is a set of vectors that define hyperplanes. When the input data cannot be split linearly, SVM classifies it using the highest margin hyperplanes after mapping it into a feature with a high dimension space, which might be infinite. Furthermore, SVM needs less training data to perform classifying in a highly dimensional feature space. Initially, SVM was designed to handle the classification of binary data problems. It has lately shown a great deal of potential in multiclass categorization.

Figure 6 depicts a straight line formed between two groups. This implies that every point of data on one side of the line will be connected with a category, while each data point on the other side will be allocated to a different category. This means that there might be an infinite number of lines accessible (Arrecho et al,2024).

The following steps are how SVM searches for the maximum marginal hyperplane:

- c

reate hyperplanes that effectively divide the classes.

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

choose the hyperplane that has the greatest segregation from the two closest data points.

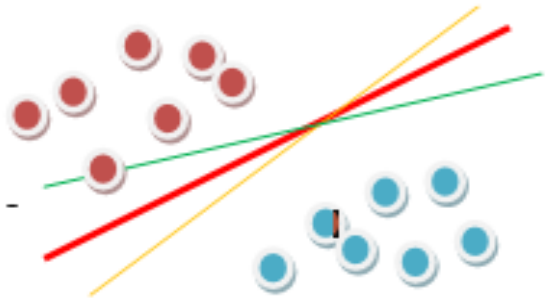

Figure 6 . SVM and possible decision boundaries

It selects the line that divides the data and is as far away as it can be from the closest data points. The decision boundary is found by utilizing the support vectors that result from the two nearest data points. It is not necessary for the decision boundary to be a line. Because one can locate the decision boundary with more than two characteristics, it is also known as a hyperplane.

SVM specializes at handling large amounts of data, which makes it an incredible instrument. It provides clear responses and is capable of understanding a variety of data formats. It also makes accurate predictions and remains stable in the context of anomalous data points. All things considered, it helps with a wide range of computer learning issues.

D. Random Forest (RF)

RF are training algorithms that are supervised to construct multiple DTs using training data. Their method is known as bagging, in which, the dataset is randomly bootstrapped to create a sample and complete DTs are parallel trained. The obtained prediction is based on the averaging of all the trees, which is shown in Figure 7. The main concept of RF is to combine the findings of multiple DTs to form more accurate and strong model.

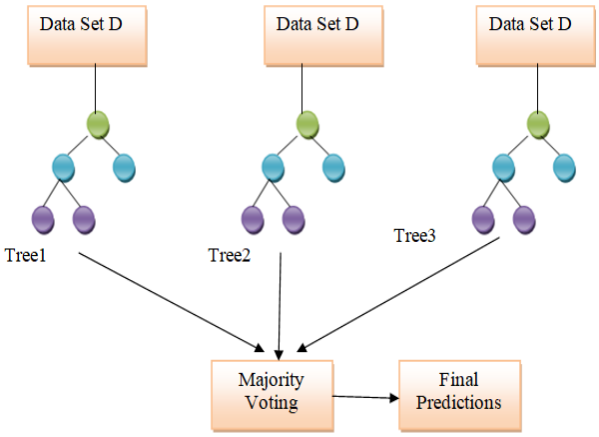

Figure 7. Random Forest Predictions

When multiple DTs are used, the predictions produced by each tree may be more accurate than the average. RF often incorporates far more information from numerous predictions than

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

a single DT, which makes it more accurate overall. The average of the DTs is used by RF to arrive at a final prediction for regression situations. A majority vote of the anticipated class can also be used to solve classification problems using RF.

E. SHapley Additive explanations(SHAP)

SHAP measures the effect of each layer in the ML model on the layer that comes after by computing SHapley values using a backpropagation technique. Propagating back through the layers and into the input layer, the process begins at the result layer. The backpropagation technique aids in identifying the input data that has the most influence on the decision made by the ML model under explanation. The contribution, positive or negative, of every characteristic in the model is shown by its SHAP value. It has two primary benefits: each record has a unique set of SHAP values and it can be computed for any model, not only basic, linear models. Choose a relevant example to receive an explanation for the black-box forecast(Gasper et al., 2024).

The features that influence a decision in a good or negative way are visually represented by SHAP's outcomes. SHAP generates a vector of importance scores, or SHAP values, as a mathematical explanation. At the time, SHAP was referred to as LIME's extension; although LIME concentrated primarily on local explanation, SHAP handled both local and global explanation(Dang et al.,2021). In comparison with explicitly calculating SHapley values, SHAP computes quickly for ML models. The explanation provided by SHAP for instance x is as follows:

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{2}$$

The explanatory model is represented by -g. The coalition vector, or simplified features, is denoted by z', with z' ∈ 0, 1 M. Whereas the 0 in z' indicates that the features in the data that was added are different from those in the original (the instance x), the 1 in z' indicates that the features in the new data are the same as those in the original data. The maximum size of a coalition is M. The feature attribution for feature j for instance x is denoted by ϕj ∈ R. The SHapley value is that. When ϕj is a high positive value, it indicates that feature j has a significant positive influence on the model's prediction (.M.Wang et al., 2020, Arrecho et al,2024)

IV. Experimental Results

The experiments were conducted to evaluate the ML models like XGBoost, Adaboost, SVM and RF . Then the outcomes of these model predictions will assist in interpreting the models using XAI models. The experiments were carried in python using suitable libraries to implement the classification and its explanations.

A. Dataset Description

The CICIDS-2017 dataset was used in this investigation to train the IDS. Selecting appropriate characteristics is a crucial aspect of intrusion detection. Some characteristics are necessary for all attacks, whereas others are just necessary for certain attacks, partially

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

required, or not required at all. The CICIDS-2017 dataset contains 76 features that are used to train and evaluate IDS. The figure 8 presents the sample features in the CICIDS-17 dataset.

SNo	Feature Name	SNo	Feature Name	SNo	Feature Name
1	Flow ID	27	Flow Packets/s	53	Fwd Packets/s
2	Source IP	28	Flow IAT Mean	54	Bwd Packets/s
3	Source Port	29	Flow IAT Std	55	Flow IAT Max
4	Destination IP	30	Flow IAT Max	56	Flow IAT Min
5	Destination Port	31	Flow IAT Min	57	Fwd IAT Total
6	Protocol	32	Fwd IAT Total	58	Fwd IAT Mean
7	Timestamp	33	Fwd IAT Mean	59	Fwd IAT Std
8	Flow Duration	34	Fwd IAT Std	60	Fwd IAT Max
9	Total Fwd Packets	35	Fwd IAT Max	61	Fwd IAT Min
10	Total Backward Packets	36	Fwd IAT Min	62	Bwd IAT Total
11	Total Length of Fwd Packets	37	Bud IAT Total	63	Bwd IAT Mean
12	Total Length of Bwd Packets	38	Bid IAT Mean	64	Bwd IAT Std
13	Fwd Packet Length Max	39	Bwd IAT Std	65	Bwd IAT Max
14	Fwd Packet Length Min	40	Bwd IAT Max	66	Bwd IAT Min
15	Fwd Packet Length Mean	41	Bwd IAT Min	67	Fwd PSH Flags
16	Fwd Packet Length Std	42	Fwd PSH Flags	68	Bwd PSH Flags
17	Bwd Packet Length Max	43	Bwd PSH Flags	69	Fwd URG Flags
18	Bwd Packet Length Min	44	Fwd URG Flags	70	Bwd URG Flags
19	Bwd Packet Length Mean	45	Bwd URG Flags	71	Fwd Header Length
20	Bwd Packet Length Std	46	Fwd Header Length	72	Bwd Header Length
21	Flow Bytes/s	47	Bwd Header Length	73	Fwd Avg Bytes/Bulk
22	Flow Packets/s	48	Fwd Packets/s	74	Fwd Avg Packets/Bulk
23	Flow IAT Mean	49	Bwd Packets/s	75	Fwd Avg Bulk Rate
24	Flow IAT Std	50	Min Packet Length	76	Bwd Avg Bytes/Bulk
25	Flow IAT Max	51	Max Packet Length	77	Bwd Avg Packets/Bulk
26	Flow IAT Min	52	Packet Length Mean	78	Bwd Avg Bulk Rate

Figure 8: Sample Features in the CICIDS-2017 dataset

B. Performance Evaluation of ML models

Firstly , the metrics such as Accuracy, Precision, and Recall were used to assess the effectiveness of the ML models.

(i)Accuracy: The proportion of samples that are correctly classified is compared with all samples, and the general accuracy of the model's predictions is evaluated. It offers a comprehensive evaluation of the model's efficacy.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

(ii)Precision: It is the percentage of positively recognized positive samples relative to the total samples projected to be positive.

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

(iii) Recall : also called sensitivity, recall is the proportion of correctly identified positive samples among all positive samples.

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

Table 2 : Methods and Metrics

S.No	Methods/Metrics	Accuracy	Precision	Recall
1	XGBoost	0.84	0.79	0.85
2	Adaboost	0.85	0.78	0.86
3	SVM	0.88	0. 79	0.88

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4	RF	0.90	0.80	0.89

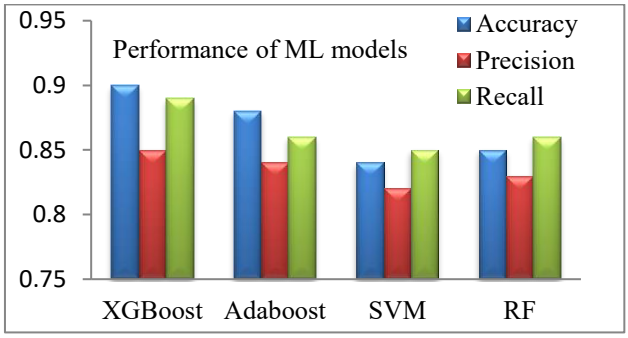

Figure 9. ML models Vs Metrics

From table 2 and figure 9, it is clear that the XGBoost outperforms the other methods. The weak models are combined to create a strong classifier. Next , the metrics are found to be more for Adaboost , which is another combined model to form a strong classifier followed by RF and SVM. These improvements are due to the usage of different models with the residuals of the previous predictions. These models classify the traffic data and it does not show on what basis it classifies. The ML models further used XAI models to generate explanations or interpretations.

C. Evaluation of XAI models

Information gain: This study uses 10 important items by the CICIDS-2017 data to maximize classification accuracy and give explainable insights. Such features are selected according to their information gain (IG) scores. Table 3 is a list of the IG values of the best 10 features and Figure 9 graphically shows these 10 features and the IG values of these features.

Information gain is a filter based feature selection measure which measures the contribution of a feature to predicting the target class. In decision tree algorithms, it is further applied to define the most useful features to divide the data. The characteristics that have high IG values play a greater role in learning new examples in the correct classification. IG is calculated by the difference between the change in class entropy pre- and post- split with a range of 0 (no gain) to 1 (maximum gain).The IG is mathematically expressed as follows

$$IG(E, A)= H(E) – H(E|A \quad (7)$$

In equation 1, IG(E, A) = IG of feature A in dataset E , H(E) = Entropy of the class distribution

H(E|A)=conditional entropy of the in dataset E for a given feature A, Entropy (E) is the measure of impurity . The formula for H(S) is

$$H(S) = -sum(p\_i * log2(p\_i)) \quad (8)$$

In equation(8) H(S) =entropy of set S. By dividing the dataset E into subgroups according to the values of feature A and calculating the entropy of each subset, the conditional entropy H(E|A) could be determined. The weighted average of the subsets' entropies, adjusted for the percentage of examples in each subset, is the conditional entropy. The features with the

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

highest IG's are considered for classification. This is one way of reducing the dimensions of the dataset(Mounica et al, 2024).

Table 3. Information gain -Selected Features

S.No	Feature Name	Information Gain
1	Average Packet Size	1.176
2	Packet Mean Length	1.163
3	Packet Length Std	1.138
4	Packet Length Variance	1.138
5	Flow IAT Max	1.117
6	Flow Duration	1.096
7	Forward Packets/s	1.064
8	Flow Bytes /s	1.047
9	Flow Packet /s	1.045
10	Total length of forward packets	1.011

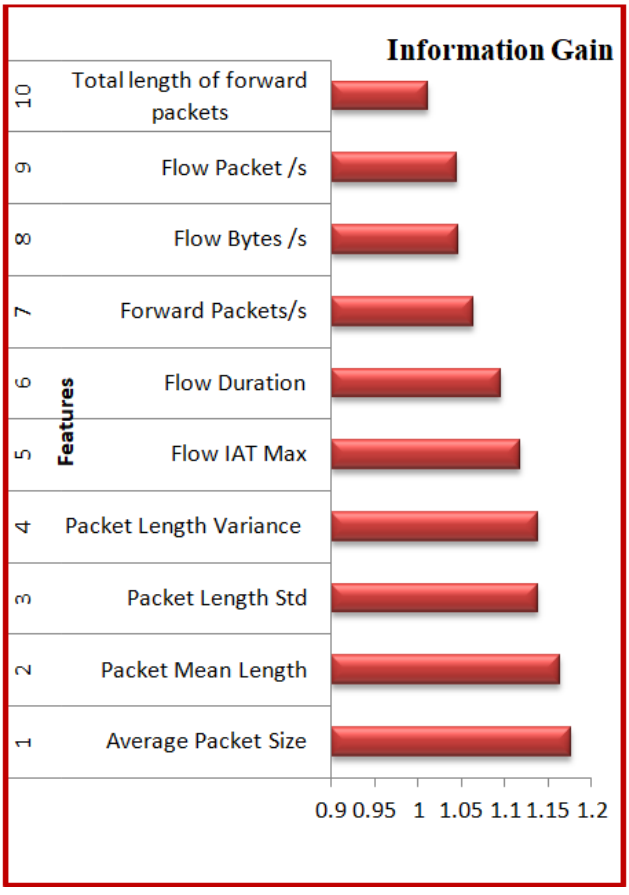

Figure .10 Features Vs Information gain

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The table 3 provides the information gain for each of the features considered for the SHAP explanations and the graphical representation for the 10 top features is shown in figure 10.

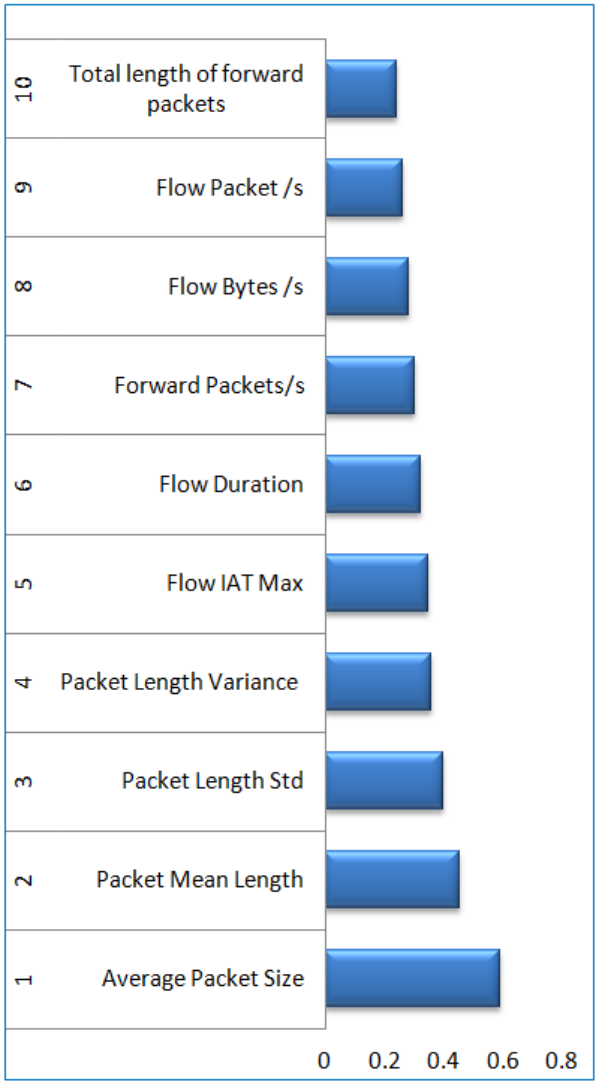

Figure 11 Analysis using SHAP for XGBoost Algorithm

The figure 11 presents the analysis of SHAP with the XGboost model that outperforms in this study. The results clearly indicate that the feature selected using a filter based information gain much coincides with the results of the SHAP analysis. It clearly explains that if the features are selected with the utmost importance, the model's efficiency also improves and the analysis reveals that the selected features decides the accuracy of the ML models.

SHAP values are a valuable tool for understanding the significance of characteristics in ML models, including applications that use the CICIDS dataset, which is frequently used in network intrusion detection research. SHAP values provide information concerning how every attribute contributes to the model's predictions by measuring its marginal contribution. The summary plot displays the importance of each feature and the SHAP values in the model. The figure 12, 13, 14 and 15 depicts the summary plot for explaining the features importance in detecting the intrusions efficiently for Random Forest, Support vector machine, AdaBoost and XGboost algorithms.

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

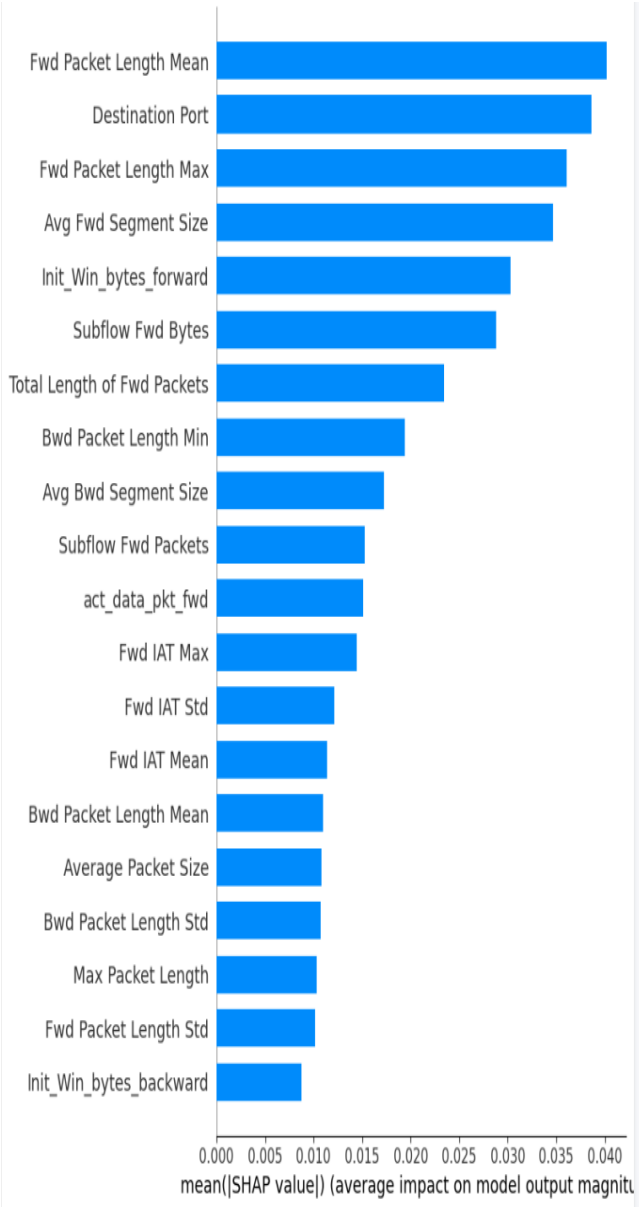

Figure 12 SHAP values CICIDS dataset using Summary plot for Random Forest

Figure 12 shows the Y axis indicating characteristic names in order of relevance from top to bottom, and the X axis representing SAHP scores. The average length of packet at the top of the summary plot represents an average length of the packets transmitted in the forward direction.. This feature contributes more in the case of random forest models in determining the intrusions as normal or malicious. For example, in a SHAP summary figure, users may notice a clear gradient in which greater Fwd Packet Length Mean values substantially favor predictions in the "malicious" class. RF handles feature interactions efficiently. The SHAP values emphasize packet size and flag characteristics. Feature importance is typically distributed across many top features.

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

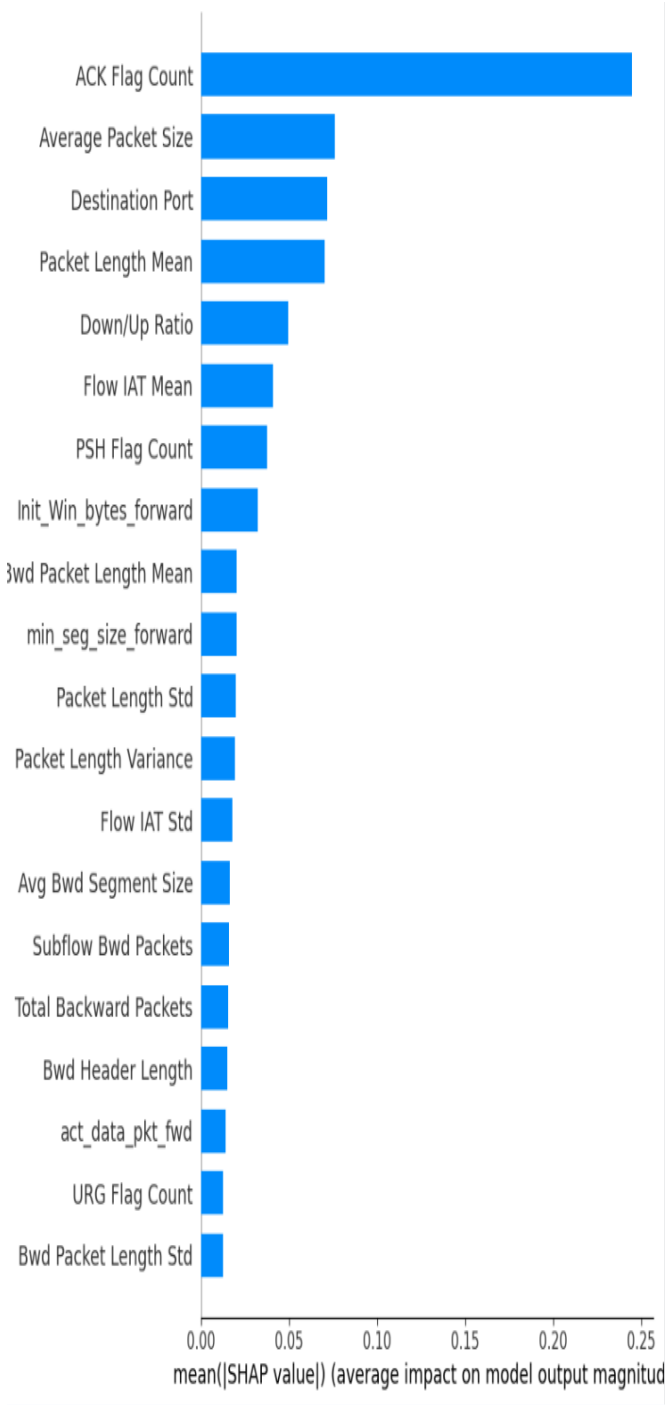

Figure 13 SHAP values CICIDS dataset using Summary plot for SVM

Figure 13 presents the summary plot for the SVM. The ACK Flag count , which shows high importance in the SVM model, represents the number of packets in a network flow that contain the ACK (Acknowledgement) flag. In TCP communication, this flag indicates that data received has been acknowledged. High SHAP values imply that anomalous ACK patterns have a considerable influence on the model's ability to predict malicious traffic. SVMs are sensitive to specific properties such as ACK Flag Count, particularly when used with kernel approaches.
SHAP frequently has fewer dominating elements and focuses on clear segregation.

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

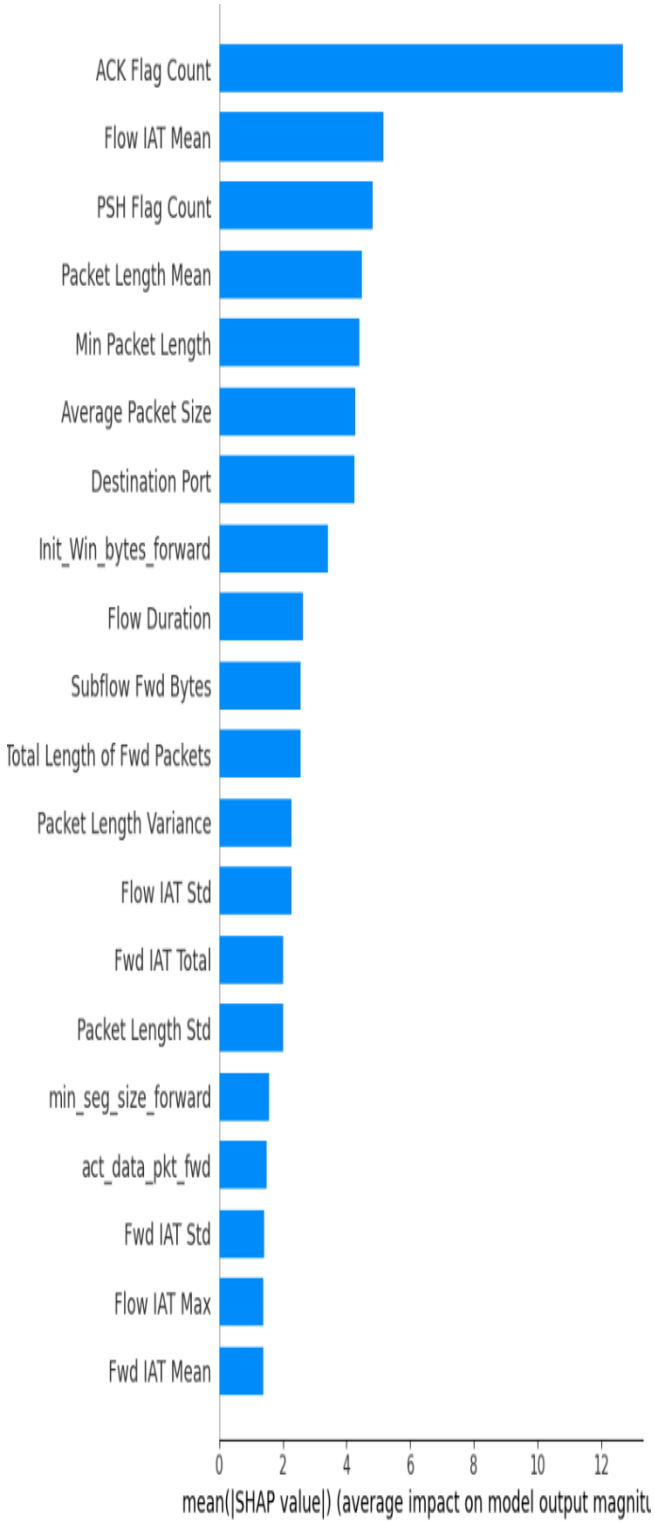

Figure 14 SHAP values CICIDS dataset using Summary plot for AdaBoost

Figure 14 displays the summary plot for the AdaBoost algorithm. This algorithm also shows that the ACK Flow Count as the Attacks such as SQL injection and brute force may not produce anomalous ACK counts, making them less significant in these circumstances. Adaboost concentrates on aspects where misclassification was widespread, such as flow duration. SHAP values tend to prioritize "hard-to-classify" traits.

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

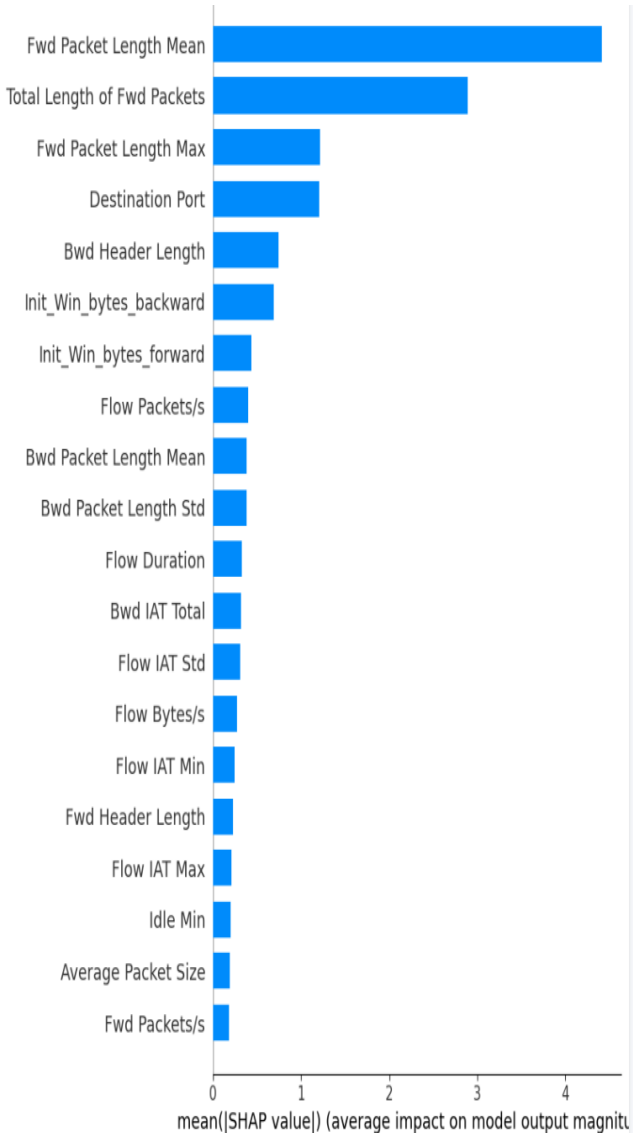

Figure 15 SHAP values CICIDS dataset using Summary plot for XGBoost

Figure 15 shows the summary plot for the XGBoost model. In tree-based models such as RF or XGBoost, Fwd Packet Length Mean may rank high in feature relevance due to its strong association with the target labels. XGBoost shows large differences in SHAP contributions, frequently rating Fwd Packet Length Mean and Flow Bytes/s high. It handles feature duplication effectively, resulting in strong interpretations. SHAP values in DL or other models can validate their contribution to the decision boundary.

The RF methods outperforms the SVM , Adaboost and XGBoost algorithms in terms of the classification performance and the feature ranking with the information gain and the SHAP values in the summary plot clearly shows the top ranking features in detecting the attacks. Furthermore, the SHAP explanations are further evaluated using the following metrics with the survey analysis.

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Descriptive Accuracy: The descriptive accuracy illustrates the importance of the most invasive elements once they are removed from the AI model. For example, deleting a key intrusion detection component should impair the efficacy of the AI model's intrusion prediction. A feature has a high XAI power if its removal results in a considerable loss of accuracy. For example, if the average size of the packet is not used for categorization, the model's accuracy suffers significantly.

Sparsity: The sparsity with an XAI method is the extent to which few features play a significant role in the model decision making. The number of feature importance scores below a very small value is counted to measure it. As an example, when 18/20 intrusion detection features have importance values near to zero, the model predicts the 20 features strongly; that is, the two features play a significant role in the model prediction, which is high sparsity. The advantage of high-sparsity XAI techniques is that security analysts can monitor networks with fewer key metrics. The top ten features have been chosen in this study because of the manner in which the features are important in the dataset.

Efficiency: The time required for a XAI technique to generate a description is an indicator of its efficacy. This number is crucial because it indicates how effectively the XAI technique applies to real-world systems, which is more practical when explanations are provided rapidly rather than slowly. Since the ultimate aim is to aid security analysts, it is predicted that the algorithm would be able to swiftly create exact XAI explanations to recognize breaches.. The time taken for the XAI models are more or less similar to the black box models and this has not been experimented due to the online execution environment.

Stability: Stability of an XAI method is how repeatable it is in generating explanations that remain consistent when run on the same conditions several times. It is also tested by determining the number of features that are reproducible in repeated experiments. These stability analyses can assist security analysts to select the most reliable XAI approach to detect network intrusion. Stability testing was not done in this study because of the limitations of the platform used.

Robustness: Robustness characterizes how an XAI technique provides the same type of explanation despite changes in input data, but with minor alterations. Such changes may be due to the ill intent manipulation or noise. The other valuable property, completeness, indicates the ability of the method to cover all possible network activities, including rare or edge cases. In case an XAI strategy is not robust enough or complete, the attackers can take advantage of such vulnerabilities, resulting in erroneous or wrong system outputs.

The model generates explanations without violating the normal feature selection technique like filter based approach, but still experiments has to be conducted to find the efficiency, stability and the robustness of the XAI models in providing explanations. Table 4 shows the comparison of model evaluation metrics with and without SHAP.

Received: August 03, 2025

International Journal of Applied Mathematics
Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table :4 Performance with and without SHAP

Metric	Without SHAP	With SHAP
Descriptive Accuracy	Key features identified manually or via feature importance (e.g., Random Forest, XGBoost)	SHAP provides clear feature contributions and highlights how accuracy drops with feature removal
Sparsity	Features with low correlation identified using traditional methods (e.g., correlation matrix)	SHAP directly identifies which features contribute little to no decision-making (Shapley values near 0)
Efficiency	Faster prediction with less detailed explanations	Potentially slower due to Shapley value calculations but offers better insights
Stability	Feature importance may vary across experiments	Stability of SHAP explanations can be tested based on feature importance consistency
Robustness	Performance tested with perturbations to input data	SHAP explanations remain consistent with slight data modifications if the model is robust

SHAP-based in contrast to non-SHAP-based evaluation metric comparison table is the presentation of the trade-off and benefit of using SHAP to provide model interpretability for IDS. With the application of SHAP, increased descriptive accuracy by which complete feature importance explanation and the impact of removing essential features on model

Received: August 03, 2025

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

performance are illustrated is achieved. Sparsity is easier to describe with SHAP so that analysts are able to easily identify features that have relatively lower contributions to model predictions. But less effectively, whereas baseline models are able to predict faster, SHAP requires more computational power to calculate Shapley values to contribute by feature, which makes the process less effective. Even with this, there is more consistency with SHAP because it gives similar feature importance scores in several different experiments and thus the model is more trustworthy. There is also more robustness with SHAP because it provides the capability to trace much better how small variations in input features affect the model's decision-making. In general, SHAP enhances model transparency, interpretability, and credibility by providing explicit understanding of feature impacts on predictions at the cost of some computational efficiency. In critical applications like cybersecurity, the benefits of using SHAP like better understanding and reliability overcome the cost of performance, and it is a valuable IDS tool.

V. Conclusion

An intrusion detection tool's primary objective is to act as a barrier or preventative measure against invaders. Automating these tools can be simplified by the incorporation of AI. This study uses a framework with the four ML models and a XAI model to classify and interpret the network traffic data to prevent the network from the hackers. The study is conducted with the ML models like XGBoost, Adaboost, SVM and RF for classification and SHAP for generating the interpretations and provides the analysis of those models. The study can be further extended to study the performance in terms of XAI performance metrics like stability, efficiency (execution time) and robustness of the model. Furthermore, the prospective avenues for IDS research could involve exploring how to incorporate Blockchain and XAI into current methodologies.

References

[1]. Sharma, Bhawana, Lokesh Sharma, Chhagan Lal, and Satyabrata Roy. "Explainable artificial intelligence for intrusion detection in IoT networks: A deep learning based approach." Expert Systems with Applications 238 (2024): 121751.

[2]. O. Arreche, T. R. Guntur, J. W. Roberts and M. Abdallah, "E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection," in *IEEE Access*, vol. 12, pp. 23954-23988, 2024, doi: 10.1109/ACCESS.2024.3365140.

[3]. A. Yayla, L. Haghnegahdar and E. Dincelli, "Explainable artificial intelligence for smart grid intrusion detection systems", *IT Prof.*, vol. 24, no. 5, pp. 18-24, Sep. 2022.

[4]. E. Roponena, J. Kampars, J. Grabis and A. Gailītis, "Towards a human-in-the-loop intelligent intrusion detection system", *Proc. CEUR Workshop*, pp. 71-81, 2022.

[5]. D. Han, Z. Wang, W. Chen, K. Wang, R. Yu, S. Wang, et al., "Anomaly detection in the open world: Normality shift detection explanation and adaptation", *Proc. Netw. Distrib. Syst. Secur. Symp.*, pp. 1-13, 2023.

[6]. A. Warnecke, D. Arp, C. Wressnegger and K. Rieck, "Evaluating explanation methods for deep learning in security", *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, pp. 158-174, Sep. 2020.

Received: August 03, 2025

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[7]. D. Gaspar, P. Silva and C. Silva, "Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron," in *IEEE Access*, vol. 12, pp. 30164-30175, 2024, doi: 10.1109/ACCESS.2024.3368377.

[8]. Mohammed, B. (2024). The Synergy of Explainable AI and Learning Analytics in Shaping Educational Insights. IAENG International Journal of Computer Science, 51(9).

[9]. Malik R., Raza H., and Saleem M., Towards A Blockchain enabled integrated library managment system using hyperledger fabric: using hyperledger fabric, International Journal of Computational and Innovative Sciences. (2022) 1, no. 3, 17–24.

[10]. Bellegdi, Sameh, Ali Selamat, Sunday O. Olatunji, Hamido Fujita, and Ondfrej Krejcar. "Explainable Machine Learning for Intrusion Detection." In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 122-134. Singapore: Springer Nature Singapore, 2024.

[11]. D. Satyanarayana and E. Saikiran, "Intrusion Detection System in Explainable Artificial Intelligence by Using Different Algorithms," 2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT*)*, Bengaluru, India, 2024, pp. 1-4, doi: 10.1109/ICDCOT61034.2024.10515463

[12]. Hooshmand, Mohammad Kazim, Manjaiah Doddaghatta Huchaiah, Ahmad Reda Alzighaibi, Hasan Hashim, El-Sayed Atlam, and Ibrahim Gad. "Robust network anomaly detection using ensemble learning approach and explainable artificial intelligence (XAI)." Alexandria Engineering Journal 94 (2024): 120-130.

[13]. Abrar, Iram, Zahrah Ayub, Faheem Masoodi, and Alwi M. Bamhdi. "A machine learning approach for intrusion detection system on NSL-KDD dataset." In 2020 international conference on smart electronics and communication (ICOSEC), pp. 919-924. IEEE, 2020.

[14]. R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," in *IEEE Access*, vol. 7, pp. 41525-41550, 2019, doi: 10.1109/ACCESS.2019.2895334

[15]. Al-Omari, M., Rawashdeh, M., Qutaishat, F. et al. An Intelligent Tree-Based Intrusion Detection Model for Cyber Security. J Netw Syst Manage 29, 20 (2021). https://doi.org/10.1007/s10922-021-09591-y

[16]. H. Satilmiş, S. Akleylek and Z. Y. Tok, "A Systematic Literature Review on Host-Based Intrusion Detection Systems," in *IEEE Access*, vol. 12, pp. 27237-27266, 2024, doi: 10.1109/ACCESS.2024.3367004

[17]. c, Osvaldo, Tanish Guntur, and Mustafa Abdallah. 2024. "XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems" *Applied Sciences* 14, no. 10: 4170. https://doi.org/10.3390/app14104170

[18]. Violet turri, 2022,What is Explainable aI? , https://insights.sei.cmu.edu/blog/what-is-explainable-ai/

[19]. Wali, Syed, and Irfan Khan. "Explainable AI and random forest based reliable intrusion detection system." Authorea Preprints (2023).

Received: August 03, 2025

Volume 38 No. 5s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[20]. Hariharan, S., Rejimol Robinson, R.R., Prasad, R.R. et al. XAI for intrusion detection system: comparing explanations based on global and local scope. J Comput Virol Hack Tech 19, 217–239 (2023). https://doi.org/10.1007/s11416-022-00441

[21]. Patil, Shruti, Vijayakumar Varadarajan, Siddiqui Mohd Mazhar, Abdulwodood Sahibzada, Nihal Ahmed, Onkar Sinha, Satish Kumar, Kailash Shaw, and Ketan Kotecha. 2022. "Explainable Artificial Intelligence for Intrusion Detection System" *Electronics* 11, no. 19: 3079. https://doi.org/10.3390/electronics11193079

[22]. Dang, Q.-V. Improving the performance of the intrusion detection systems by the machine learning explainability. Int. J. Web Inf. Syst. 2021, 17, 537–555.

[23]. C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Independently Published, 2022, [online] Available: https://christophm.github.io/interpretable-ml-book.

[24]. M. Wang, K. Zheng, Y. Yang and X. Wang, "An explainable machine learning framework for intrusion detection systems", *IEEE Access*, vol. 8, pp. 73127-73141, 2020.

[25]. Mounica, B., and K. Lavanya. "Feature selection method on twitter dataset with part-of-speech (PoS) pattern applied to traffic analysis." International Journal of System Assurance Engineering and Management 15, no. 1 (2024): 110-123.

[26]. Belavagi, M. C., & Muniyal, B. (2023). Intrusion Detection Using Rule Based Approach in RPL Networks. IAENG International Journal of Computer Science, 50(3), 988-999.

Received: August 03, 2025